Data Quality Challenges for Analysis Ready Data (ARD)

Paper #: IN13B-0575
2023 AGU Fall Meeting

Zhong Liu[1], Robert Downs[2], Ge Peng[3], David F. Moroni[4], Hampapuram Ramapriyan[5], Yaxing Wei[6], and Chung-lin Shie[7]

[1] NASA GES DISC/GMU; [2] NASA SEDAC, ESCO, Astromat & CIESIN; [3] University of Alabama in Huntsville and NASA MSFC IMPACT; [4] Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA; [5] SSAI and NASA Goddard Space Flight Center, [6] NASA ORNL DAAC; [7] UMBC (ret.)

Abstract

Data quality plays a critical role in research and applications. The Earth Science Information Partners (ESIP) Information Quality Cluster (IQC) defines four aspects of information quality: **Science, Product, Stewardship, and Services**. The ESIP IQC has become internationally recognized as an authoritative and responsive resource of information and guidance to data producers and distributors on how to implement data quality standards and best practices for their science data systems, datasets, and data/metadata dissemination services.

In recent years, cloud computing environments have provided scale-up capabilities such as data archives and services, enabling interdisciplinary science and applications. More value-added products are expected from data service providers, including Analysis Ready Data (ARD). ARD refers to data that has been preprocessed into a form that allows immediate analysis by the end user, processed to a minimum set of requirements and provides interoperability over time and across multiple datasets. Once a dataset has been developed from its original form to produce ARD, what quality characteristics should the derived dataset or ARD possess? Also, is it safe to assume that the quality of the ARD is consistent with the quality of the source data, or are there special attributes to an ARD that would warrant a secondary, independent quality assessment? What provenance (also called "data lineage") information needs to be included in ARD? It is important to answer these questions, especially given the ease of use of ARD, and the consequent temptation by users to trust ARD without understanding the limitations or possible variations in quality compared to the source data. In this presentation, we will discuss data quality challenges for ARD products and services and introduce IQC for participation.

Analysis Ready Data (ARD) - Key Questions

- ARD refers to data that have been preprocessed into a form that allows immediate analysis by the end user (Lynnes and Hua 2019), processed to a minimum set of requirements and provide interoperability over time and across multiple datasets (https://ceos.org/ard/).
- Once a dataset has been developed from its original form to produce ARD, what characteristics would be needed to identify and represent the quality of the ARD?
- Is ARD consistent with the quality of the source data, especially data from multiple sources and different measurements or algorithms, or are there special attributes to an ARD that would warrant a secondary, independent quality assessment?
- It is important to answer these questions, especially given the ease of use of ARD, and the consequent temptation by users to trust ARD without understanding the limitations or possible variations in quality compared to the source data.

Why is ARD so Important?

ARD plays a key role in accelerating research and applications.

- Accelerate research and applications: more time on research, less on data processing (e.g., format, grid, aggregation) – the 80/20 rule in data science
- Facilitate compliance with FAIR guiding principles (Findable, Accessible, Interoperable, and Reusable)
- Promote open science (workflow transparency)
- Enable interdisciplinary science

ARD Data Quality Challenges

Data quality plays a critical role in research and applications (e.g., climate change, extreme weather events, model prediction, AI/ML). However, there are many challenges to generate quality information in the following four aspects from the IQC: science, product, stewardship, and services that ARD depends on. These four aspects are summarized in Table 1 (Ramapriyan et al. 2017) below:

Table 1: Different information quality aspects, associated data product life cycle stages and responsible groups.

Information Quality Aspect	Life Cycle Stage	Responsible Group
Scientific	1. Define, develop, and validate	Science Team
Product	2. Produce, assess and deliver (to an archive or data distributor)	Science Team
Stewardship	3. Maintain, preserve and disseminate	Archive/Distributor
Service	4. Enable use, provide support and service	Archive/Distributor

More details can be found in Ramapriyan et al. 2017. In addition to these aspects, user needs are equally important. User involvement plays a key role to ensure data products and services meet their needs such as research, applications, and education. In order to tackle data quality challenges, participation from all related parties is needed in an integrated, coordinated, open and networked way (Hills et al. 2022).

Information

IQC information:
IQC Wiki: http://wiki.esipfed.org/index.php/Information_Quality
Subscribe: https://lists.esipfed.org/mailman/listinfo/Esip-infoquality

The ESIP Information Quality Cluster: Providing authoritative information and guidance to data providers on how best to implement data quality standards.

References:

Hills, D., J. Damerow, B. Ahmmed, et al. N. Catolico, S. Chakraborty, T. Y. Chen, C. Coward, R. Crystal-Ornelas, W. Duncan, L. Goparaju, C. Lin, Z. Liu, M. Mudunuru, Y. Rao, R. Rovetto, Z. Sun, B. Whitehead, L. Wyborn, and T. Yao. 2022. Earth and Space Science Informatics Perspectives on Integrated, Coordinated, Open, Networked (ICON) Science Earth and Space Science [10.1029/2021EA002108]

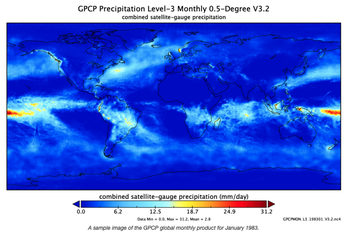
Huffman, G.J., A. Behrangi, D.T. Bolvin, E.J. Nelkin (2022), GPCP Version 3.2 Satellite-Gauge (SG) Combined Precipitation Data Set, Edited by Huffman, G.J., A. Behrangi, D.T. Bolvin, E.J. Nelkin, Greenbelt, Maryland, USA, Goddard Earth Sciences Data and Information Services Center (GES DISC), Accessed: [*Data Access Date*], 10.5067/MEASURES/GPCP/DATA304

Lynnes and Hua 2019, Analysis Ready Data in Analytics Optimized Data Stores for Analysis of big Earth Data in the Cloud. Available: https://ntrs.nasa.gov/api/citations/20190033486/downloads/20190033486.pdf

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

Acknowledgements: These activities were carried out across multiple United States government-funded institutions (noted above) under contracts with the National and Space Administration (NASA). Government sponsorship is acknowledged. However, it does not constitute or imply the endorsement by NASA and the affiliated institutions.

A Case Study – Satellite-based Precipitation Products from Global Precipitation Climatology Project (GPCP)



A sample image of the GPCP global monthly product for January 1983.

Table 1. Key data fields in the MEaSUREs GPCP V3.2 daily dataset

Data Field Name	Description	Units
precip	Precipitation estimate	mm/day
probability_liquid_phase	Probability of liquid phase	percent

Table 2. Key data fields in the MEaSUREs GPCP V3.2 monthly dataset

Data Field Name	Description	Units
sat_gauge_precip	Combined satellite-gauge precipitation estimate	mm/day
sat_gauge_error	Combined satellite-gauge precipitation random error estimate	mm/day
satellite_precip	Satellite-only precipitation estimate	mm/day
satellite_source	Source of the contributing satellite estimate	index values
gauge_precip	Wind-loss adjusted gauge precipitation	mm/day
probability_liquid_phase	Probability of liquid phase	percent
gauge_relative_weight	Gauge relative weighting	percent
quality_index	Quality index	index values

- GPCP generates (Huffman et al. 2022) multi-satellite, multi-sensor global precipitation products that are widely used in the climate community.
- Random error estimate only exists in the GPCP monthly product but is missing in the daily product.
- Products include single sensor and multi-sensor merged products (e.g., IMERG – Integrated Multi-satellitE Retrievals for GPM).
- Products are developed with multiple sources (e.g., in-situ, satellites, models).
- It is difficult (not enough knowledge) to generate data quality information, especially on a global scale. Additional research is needed as more derived products are developed in clouds as Analysis Ready Data.
- Satellite sensor conditions and operations (e.g., quality flags) are relatively easy to generate and many already exist (e.g., in Level-2 products)
- Uncertainties can be large due to a lack of in-situ observations (gauges, ground radars) for ground validation, bias correction, algorithm refinement, especially over oceans and in polar regions.

What can we do?

- Encourage participation to support data (e.g., in-situ, satellite) generation, collection, integration, standardization, and the FAIR principle on a global scale.
- Ensure provision of quality information for open science (e.g., satellite instrument conditions, anomalies).
- Promote DQ standards and best practices (examples, tools).
- Support data quality research (e.g., AI/ML, derived products (more urgent)).
- Facilitate provision of information on service and stewardship quality (in addition to data quality).
- Capture and collect user needs and identify gaps and solutions.

Your involvement maters.

Earth Science Information Partners (ESIP) Information Quality Cluster (IQC)

Vision:
Become internationally recognized as an authoritative and responsive resource of information and guidance to data providers on how best to implement data quality standards and best practices for their science data systems, datasets, and data/metadata dissemination services.

Goals:
- Bring together people from various disciplines to assess aspects of quality of Earth science data and information
- Establish and publish baseline of standards and best practices for dataset quality for adoption by inter-agency and international data providers
- Build framework for consistent capture, harmonization, and presentation of dataset quality for the purposes of climate change studies, Earth science and applications

Collaboration Area Objectives:
- Actively evaluate community data quality best practices and standards
- Improve capture, description, discovery, and usability of information about data quality in Earth science data products.
- Ensure producers of data products are aware of standards and best practices for conveying data quality, and data providers/distributors/intermediaries establish, improve and evolve mechanisms to assist users in discovering and understanding data quality information.
- Consistently provide guidance to data managers and stewards on how best to implement data quality standards and best practices to ensure and improve maturity of their datasets

Things our collaboration area needs to accomplish our objectives
- Active participation by cluster members from multiple agencies (e.g., NASA, NOAA, USGS)
- Connections with international partners
- Continued support from student fellow(s)

Our recommendations: https://www.earthdata.nasa.gov/esdis/esco/standards-and-practices/recommendations-from-the-data-quality-working-group

Our joint session with the ESIP Data Readiness cluster at the 2024 ESIP January Meeting (Jan 22-26, virtual): https://www.esipfed.org/meetings

Summary

- There are many challenges in providing data quality information for analysis ready data (ARD)
- Community efforts, participation, and collaboration (users, product developers, data service providers) are important (e.g., to guide ARD development and services)
- More data quality research activities are needed to better understand and provide data quality information in ARD