
# Data Quality and Earth Science Data: A Community Discussion

Gregory Leptoukh & Christopher Lynnes

NASA GSFC

# Background

- Users of the satellite data are asking for better information about data quality…
- But this turns out to be more surprisingly complicated.

# Objective

- Assess various aspects of quality and uncertainties in satellite data from the data user perspective
- Expose some known issues
- But mostly … solicit inputs for the community

# What Is Data Quality?

- "The state of completeness, validity, consistency, timeliness and accuracy that makes data appropriate for a specific use*"

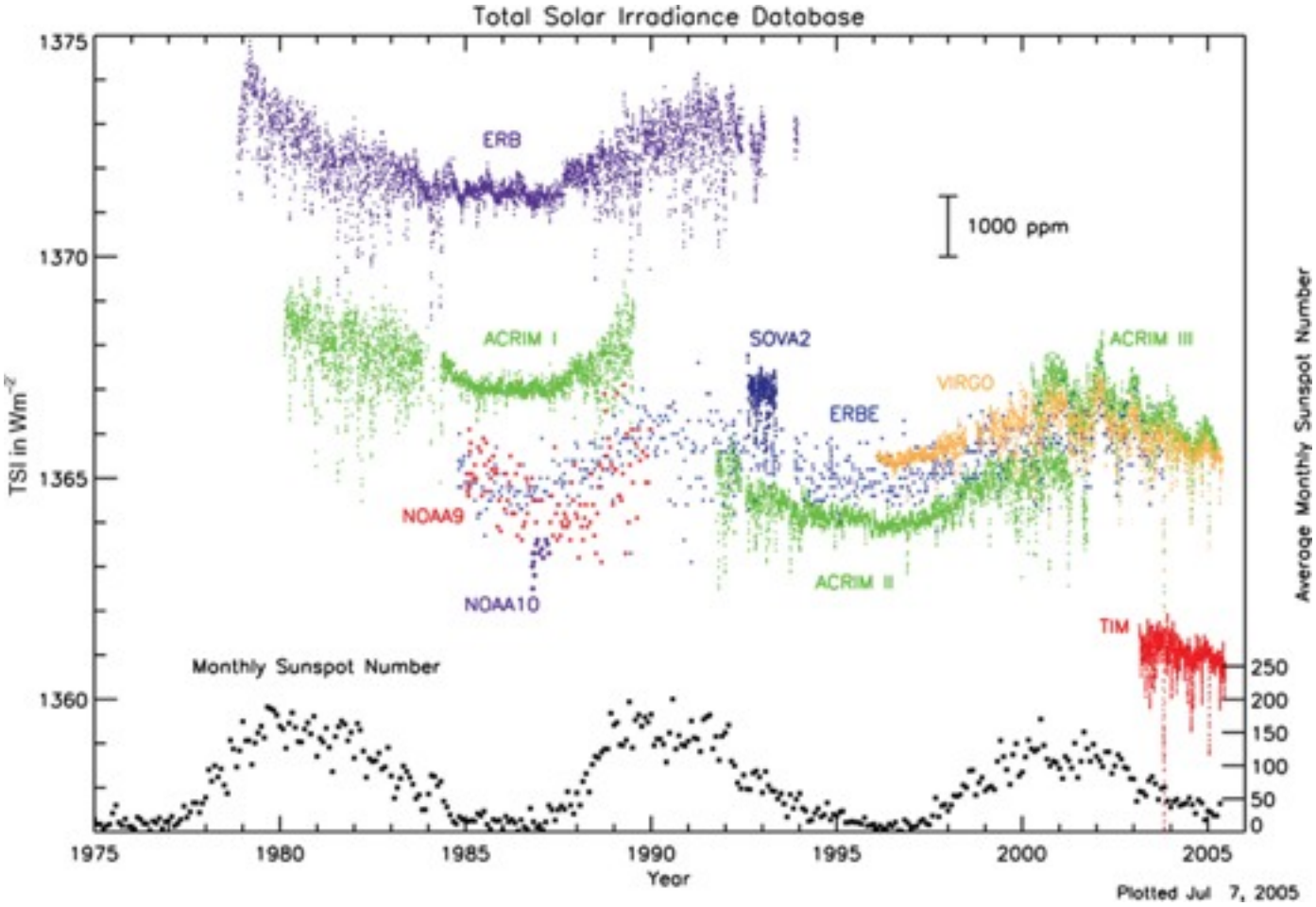
*Wikipedia, from Provincial Govt of British Columbia

# Different kinds of reported data quality

- Product-level Quality: how closely the data represent the actual geophysical state
- Pixel-level Quality: algorithmic guess at usability of data point
- Granule-level Quality: statistical roll-up of Pixel-level Quality

These types are often erroneously assumed having the same meaning

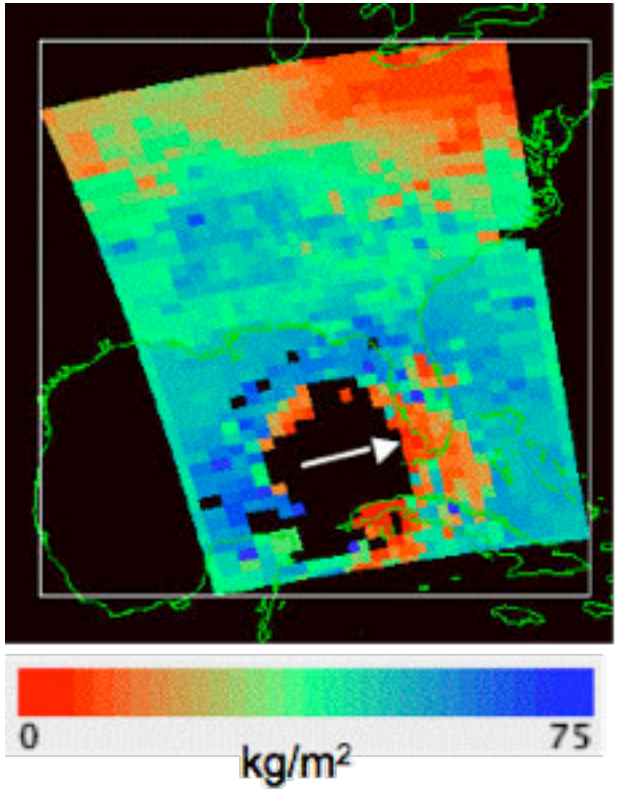
# Product-level: "Which of products to use?"



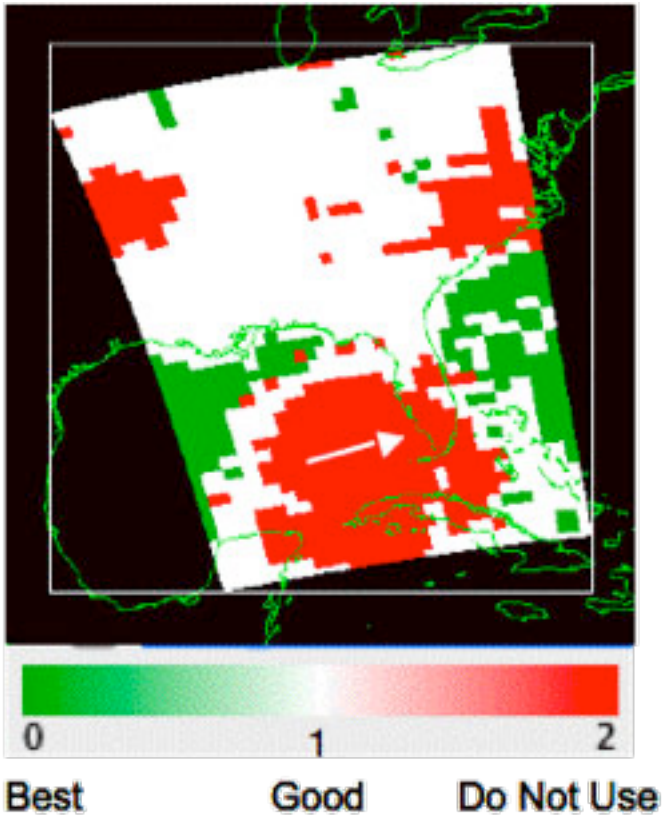

Top-of-atmosphere Total Solar Irradiance (TSI) data measured by various satellites from 1975 to 2005 (from Datla et al., 2010, Int. J. of Rem. Sensing, 31, 867–880)

# Pixel-level: "Use only pixels with quality "Good" or better."

Total Column Precipitable Water




Quality




Using bad quality data is in general *not* negligible: use bad pixels and hurricanes may look dry in the AIRS image above

# Granule-level: "Fetch only granules with >90% of pixels Good or better"

- Employed via (some) search and order tools
- Can be deceiving if the user area constitutes just a small part of the whole granule coverage

# Use cases and data quality needs

- Climate Change:
  - Model validation - gridded contiguous data with uncertainties in each grid cell
  - Long-term time series – bias assessment is the must
- Studying phenomena using multi-sensor data:
  - Consistently processed and presented data with quality information
- Applications:
  - Near-Real Time for transport and event monitoring - in some cases, coverage might be more important that quality
  - Monitoring (e.g., air quality exceedance levels) – uncertainty
- Educational (users generally not well-versed in the intricacies of quality; just taking all the data as usable can impair educational lessons) – only the best products

Quality should include assessment of uncertainty and bias

Other terms used: accuracy and precision

# Open questions

- What do users want?
- What do users need?
- What do providers want users to pay attention to?

# General Product-Level Issues

- How can we determine biases from product-level quality?
- How can we extrapolate validation knowledge about Level 2 product quality to the corresponding Level 3 gridded product quality?
- How can we harmonize quality across products – which one has better quality over certain areas?

# General Pixel-Level Issues

- How well we extrapolate validation knowledge about selected Level 2 pixels to the Level 2 (swath) product?
- How can we harmonize terms and methods for pixel-level quality, e.g. AIRS “good” vs. MODIS “3”?
- What part should these different qualities play in provenance – quality provenance?
- When is granule-level quality useful?

# Level 3 (Gridded) data quality issues

- Modelers need gridded “non-gappy” data with error bars in each grid cell
- Many differences between Level 3 data from different sensors and little uncertainty information
- Standard deviation within a grid cell reflects spatial variability at low-mid latitudes but mostly temporal variability at high latitudes
- What is validation of Level 3 product?

# Bias-related Issues

- How does bias relate to product-level quality?
- How does sampling bias affect product quality?
  - –Spatial: sampling polar area more than equatorial
  - –Temporal: sampling one time of a day only
  - –Vertical: not sensitive to a certain part of the atmosphere thus emphasizing other parts
  - – Pixel Quality : filtering by quality may mask out areas with specific features
  - – Clear sky: e.g., measuring humidity only where there are clouds may lead to dry bias
  - –Surface type related

# Current initiatives

- NASA 2010 ESDSWG / MPARWG initiative expands 2008 MEaSUREs and ACCESS programs emphasis on data quality. Legacy of the NASA Guidelines for Ensuring Quality of Information, 2001
- ESA is currently implementing contractual requirements for providing quality information within the Climate Change Initiative
- New (May'10) Guideline for the Generation of Datasets and Products Meeting GCOS Requirements
- CEOS QA4EO provides recommendations for capturing uncertainties but basically stops at Level
- ISO 19115 provides rich metadata structure for QA

Any other known initiatives?

# How it is done now?

Different disciplines have different approaches to quality handling:

- Sea Surface Temperature – plenty of measurements, good assessment of biases
- Precipitation – multiple rain gauges, appreciation of sampling bias
- Ocean Color – good Cal/Val program
- Land
- Atmospheric – not very consistent

Opinions?

# What do we propose to do?

**A framework for consistent assessment, capture and presentation of data quality information**

- Extend QA4EO effort to Level 2 and 3 data
- Address various biases
- Consistently aggregate to Level 3 products to ensure compatibility between data from different instruments
- Deliver quality information to users of data in a way that users can understand and use it

# Announcement:
# Session on Data Quality Vocabulary

- To discuss various dimensions of data quality, e.g., algorithm accuracy, application dependency, etc.
- To come up with common terminology for the future ESIP Federation workshops

Thursday, 4:30 pm

Room 403

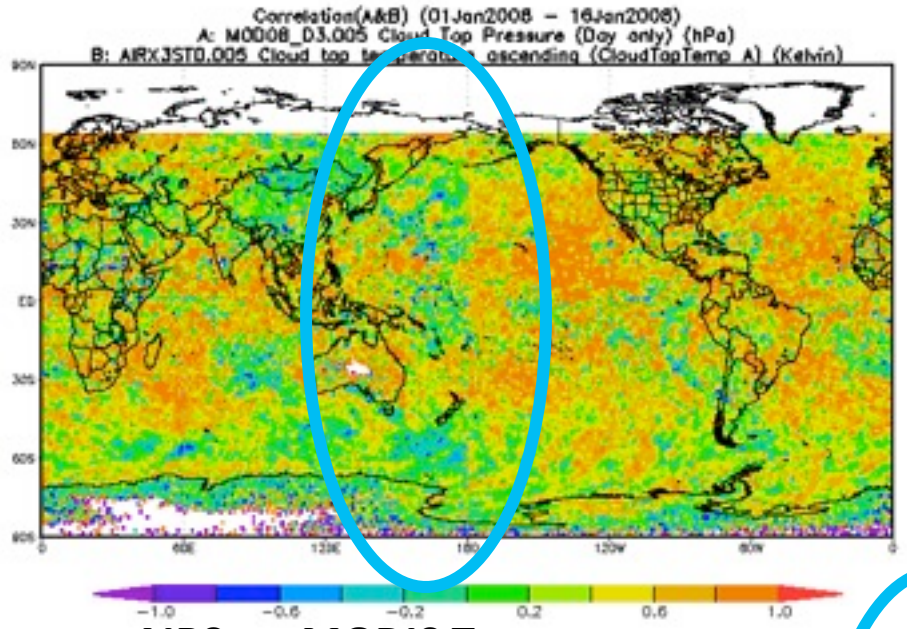
# More background material

1. Differences in quality assignment for similar pixels within the same product
2. Peculiarities and differences between Level 3 data from different sensors
3. Effects of different aggregations from Level 2 to Level 3
4. Data intercomparison methods

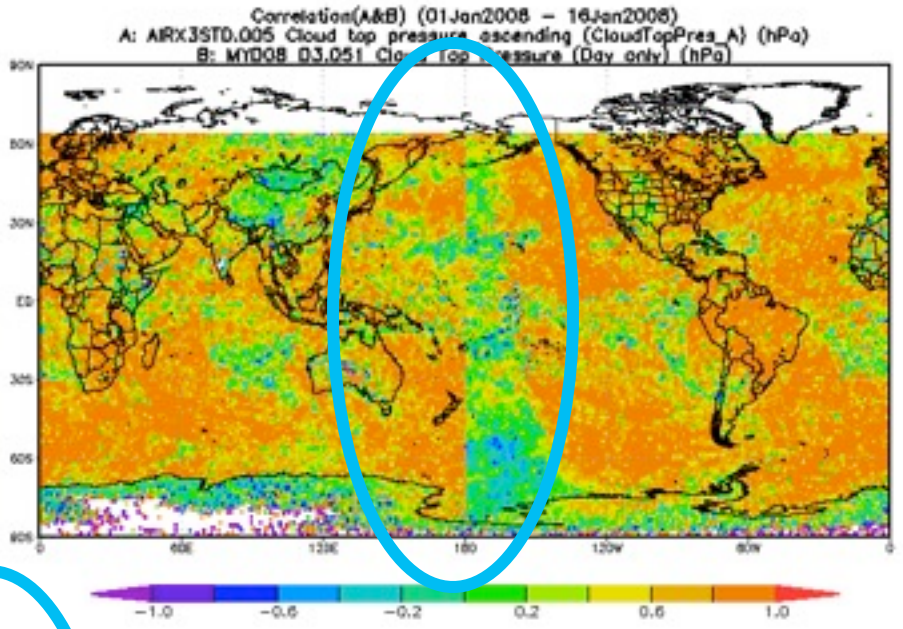
# 1. Differences in quality assignment within the same product

- MODIS Aerosol Optical Depth over ocean and land – different decision trees and meaning for QA=3 over ocean and land
- AIRS – quality threshold may differ with latitude to ensure similar coverage

# 2. Some Level 3 peculiarities

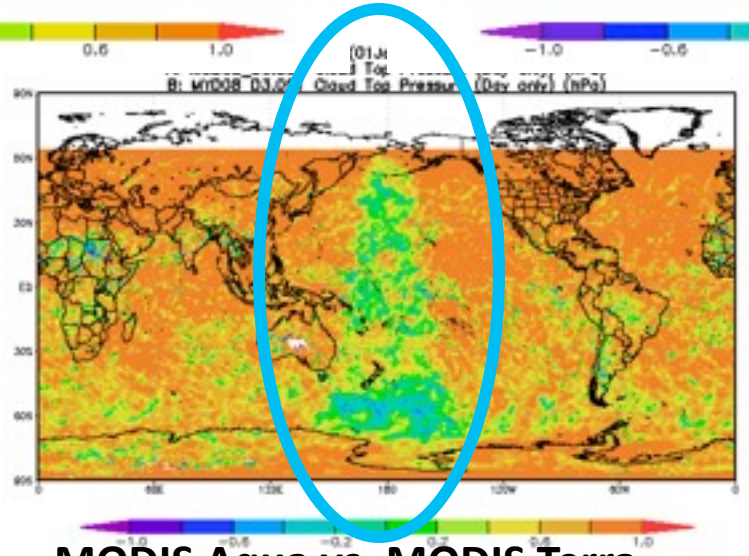



**AIRS vs. MODIS Terra**




**AIRS vs. MODIS Aqua**

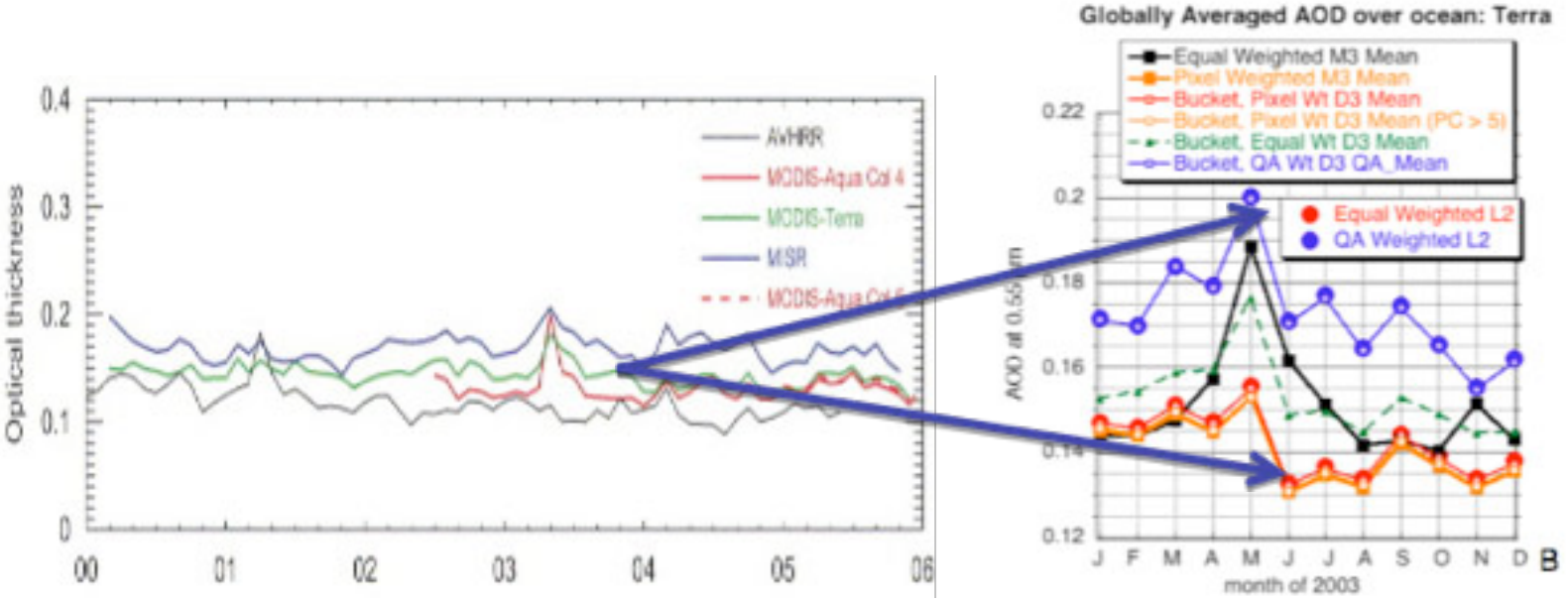
Cloud Top Pressure Correlation maps for Jan 1 – 16, 2008



**MODIS Aqua vs. MODIS Terra**

**Difference in Level 3 data day definition leads to artifacts in intercomparison**

# 3. Effects of aggregation




Aerosol Optical Depth (AOD): difference between various sensors

Mishchenko et al., 2007

MODIS -Terra AOD: difference between different aggregations

Levy, Leptoukh, et al., 2009

For MODIS-Terra alone, *AOD differences can be up to 40% depending on the aggregation method and order used to go from* L2 to L3 monthly

- Consistent aggregation from Level 2 to Level 3 is needed

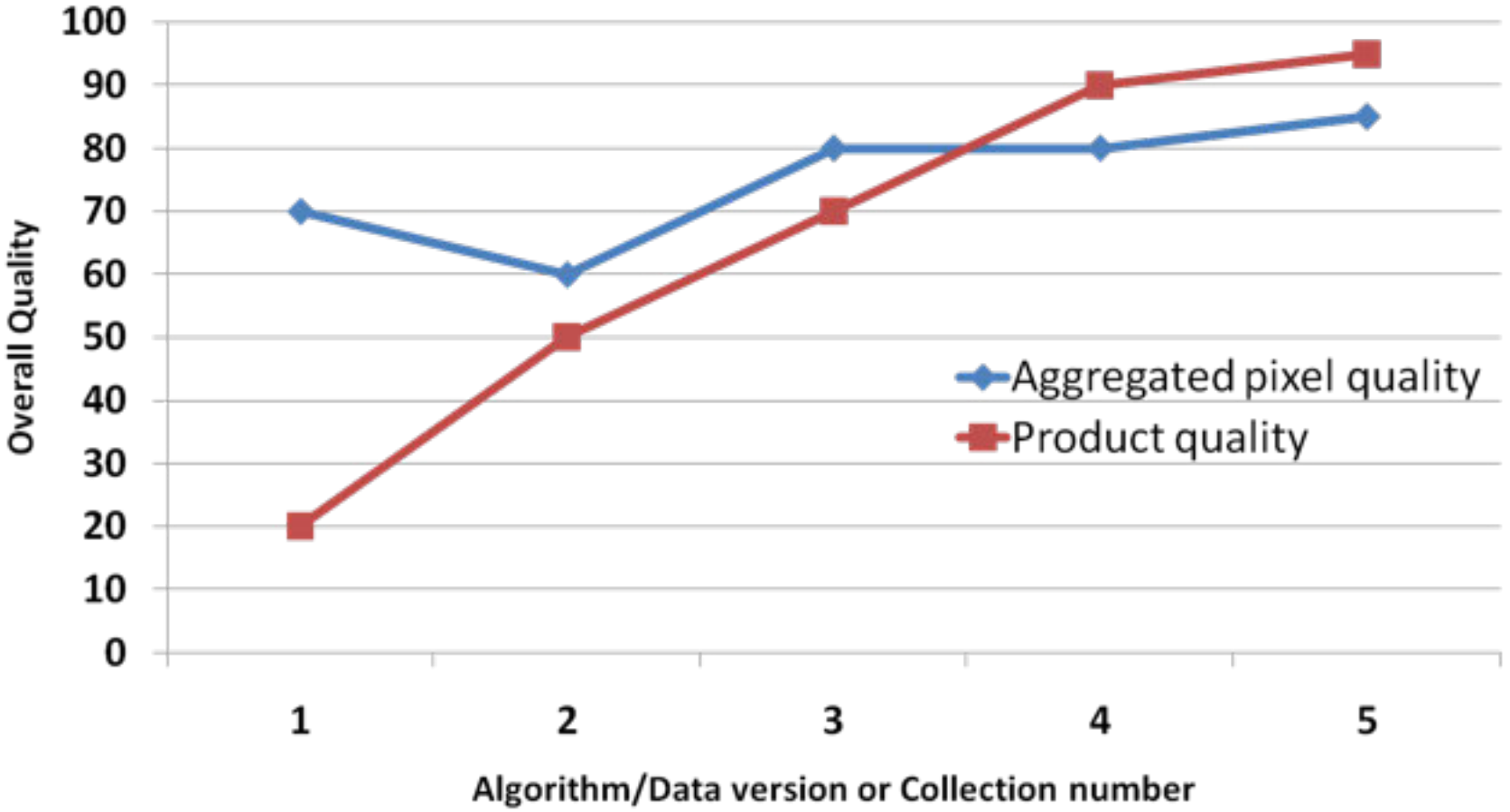
# 4. Data intercomparison methods

- Coincident data – the most straightforward
- Comparing against ground-based measurements
- Comparing via mediator (e.g., model)
- Self-consistency checks: zonal means, time-series, difference maps, …
- Using PDF
- Assimilation

# *Danger:* machines cannot do science yet

- When the data delivery protocols, metadata, authentication and other interoperability issues are resolved, there could be false impression that everything has been resolved
- However… even when different data are brought together after some harmonization, so they can easily be compared… there are many other issues to be aware of: sensor and retrieval caveats, quality, biases…

# Product Quality (based on validation) and Aggregated Pixel Quality (notional graph)




Product quality ≠ aggregated pixel quality but they are getting closer as the product matures