Addressing and Presenting Quality of Satellite Data via Web-based Services

G. Leptoukh, C. Lynnes, S. Ahmad
Goddard Space Flight Center
NASA
Greenbelt, MD, USA
Gregory.Leptoukh@nasa.gov

P. Fox, S. Zednik, P. West
Tetherless World Constellation
Rensselaer Polytechnic Institute
Troy, NY, USA

***Abstract* - With the recent attention to climate change and proliferation of remote-sensing data utilization, climate model and various environmental monitoring and protection applications have begun to increasingly rely on satellite measurements. Research application users seek good quality satellite data, with uncertainties and biases provided for each data point. However, different communities address remote sensing quality issues rather inconsistently and differently.**

We describe our attempt to systematically characterize, capture, and provision quality and uncertainty information as it applies to the NASA MODIS Aerosol Optical Depth data product. In particular, we note the semantic differences in quality/bias/uncertainty at the pixel, granule, product, and record levels. We outline various factors contributing to uncertainty or error budget; errors.

Web-based science analysis and processing tools allow users to access, analyze, and generate visualizations of data while alleviating users from having directly managing complex data processing operations. These tools provide value by streamlining the data analysis process, but usually shield users from details of the data processing steps, algorithm assumptions, caveats, etc. Correct interpretation of the final analysis requires user understanding of how data has been generated and processed and what potential biases, anomalies, or errors may have been introduced.

By providing services that leverage data lineage provenance and domain-expertise, expert systems can be built to aid the user in understanding data sources, processing, and the suitability for use of products generated by the tools.

We describe our experiences developing a semantic, provenance-aware, expert-knowledge advisory system applied to NASA Giovanni web-based Earth science data analysis tool as part of the ESTO AIST-funded Multi-sensor Data Synergy Advisor project (PI: G. Leptoukh).

Index Terms – Data Quality, Aerosols, web tools, semantic web, ontology.

I. Introduction

We live in the golden era of studying atmospheric components from space by simultaneous measurements by sensor onboard several NASA and other countries' satellites . While there is a plethora of atmospheric aerosol data available at various spatial and temporal scales, the scientific community is currently vigorously debating the lack of consensus derived from these heterogeneous observations. Quantitative differences in aerosol observations "necessitate a detailed critical assessment and integrated analysis that would go far beyond simple intercomparisons of various satellite products and comparisons of satellite aerosol optical thickness results with ground-based sun–photometer data" [5].

The reported differences may result from any of the following:

- Instrumental issues: differences in calibration, instrument sensitivity, changes or drift in calibration or sensitivity over time, different wavelengths of the observations, etc.
- Retrieval algorithm issues: different assumptions about surface characteristics, aerosol properties in some regions, aerosol particle size and shape, handling of clouds, etc.
- Observational issues: rapidly varying cloud cover, viewing angles and conditions, time(s) of observation, spatial and/or temporal scale(s) of observations being compared, etc.
- Data processing issues: different handling of quality control flags, data aggregation or (re)gridding, etc.
- Real geophysical processes that vary rapidly between near–coincident observations.

The "hot" topic of current research on the global distribution and impact of aerosols is the development of a multi-sensor merged or composite dataset of aerosol properties. However, before this can be accomplished, one must first characterize the differences between the various datasets measuring the same physical parameters (assess the comparative quality), and then identify the source(s) of the differences observed by the various satellite datasets with each other, with ground–based (AERONET) data and with model data.

Many researchers and application users get the Level 3 aerosol data or perform the analysis online using Giovanni [1, 2]. The NASA Giovanni (http://giovanni.gsfc.nasa.gov) has significantly eased access to these data and provided means of performing multi-sensor intercomparisons. Giovanni, in some cases, is the first user encounter with differences between aerosol properties measured by different sensors. The Giovanni advantage is that users can easily explore various global or regional spatial and temporal slices of data. It also helps to assess sampling and missing data patterns easily, e.g., to identify quickly where MODIS is not measuring aerosols due to clouds or sun glint. On the other hand, the current version of Giovanni does not provide (yet) exhaustive filtering by quality, cloud fraction or other constraints, which prevents more comprehensive intercomparison. Giovanni does provide

limited lineage information about the processing steps performed by Giovanni.

The Multi-sensor Data Synergy Advisor (MDSA) is a NASA ESTO-funded project (PI: G. Leptoukh) with a goal to provide Giovanni users of remotely sensed aerosol data with clear, cogent information on salient differences between data candidates for intercomparison, merging and fusion to enable scientifically and statistically valid conclusions. Tools like Giovanni make data comparison and merging too easy. This may lead to taking results at a face value without learning important details and caveats about the data. Depending on conditions, different data may have known quality issues in certain conditions: spatial location, glint/non-glint, desert, etc. MDSA will provide those cautionary notes by using a semantic web framework to characterize key properties of datasets and inference to decide what to tell the user.

II. Data Quality

A. *Why so difficult? General considerations*

There are many different qualitative and quantitative aspects of data quality. Methodologies for dealing with data qualities are just emerging. No comprehensive framework for Level 2 and Level 3 remote sensing data quality exits yet. Even the most comprehensive recent review Batini's book [3] demonstrates that there are no preferred methodologies for solving data quality issues. Little funding was allocated in the past to data quality as the priority was to build an instrument, launch a rocket, collect and process data, and publish a paper. Each Science Team handled quality differently.

What has changed? With the recent revolutionary advance in data systems, the data from multiple finally arrive to users quite easily, and all the differences in the data can easily be seen but not easily understood as the information is dispersed widely. Only now, a systematic approach to remote sensing quality is on the table. Various national and international efforts have started to address data quality issues:

Quality is perceived differently by data providers and data recipients. It is important to understand Fitness-for-purpose aspect of Data Quality. The Climate Change Modelers need gridded contiguous data with uncertainties in each grid cell. For studying long-term time series, the most important aspect is good bias assessment, especially related to sensor change and degradation, orbit and spatial sampling change. On the other hand, for the Near-Real Time monitoring of aerosol transport or high aerosol loading events, the coverage and especially the delivery timeliness is more be more important that the absolute accuracy. And educational and mass-media users (generally not well-versed in the intricacies of quality; just taking all the data as usable can impair educational lessons) need only the best "looking" products.

Working with data from multiple sources does elevate importance of harmonization. It is not sufficient just to have the data from different sensors and their provenances in one place. Before comparing or merging data, various aspects of the data need to be harmonized: Metadata (terminology, standard fields, units, scale), Data (format, grid, spatial and temporal resolution, wavelength), Provenance (source attribution, algorithm description and assumptions, processing steps), and finally Quality (bias, uncertainty, fitness-for-purpose, validation).

B. *What do the data users want and need?*

A typical question that users ask: "Which dataset is the best?" A more refined question is: "Which dataset is better for me?" This needs to be qualified by understanding that the "best" data for one user might that good for another. The challenge for the data providers and deliverers of data is to describe the data and their quality in such a way that users can make an informed decision and choose the right dataset. We need to describe the "what", the known facts about the data. We also need to provide some explanations of these facts, the "why" based on the literature. The scientists are mostly interested in "why" through "what".

The main challenge is how to characterize the quality of the data consistently across various data products, how capture this information a variety of published papers, and how to present this information to users so they can decide on which product is better for their application. It is also important to understand that most of the remote-sensing data do not provide information about uncertainty of the data. Usually, no error information is present within the data. The standard deviation values in gridded Level 3 data do not completely describe uncertainty in the data as they are the result of data variability with the grid cell convoluted with many other uncertainties in the retrieval algorithm parameters, assumptions, ancillary data uncertainties, etc.

Another important aspect is the user's trust. Only when being presented with the documented history of how the data were created, users can make their informed decision. Data Lineage or Provenance describes the source of data, including the execution history of the processes that produced them. Data by themselves without provenance are not sufficient to make accurate scientific conclusions. A perceived quality data is much lower without good provenance that is delivered together with the data in a easily understandable and clear manner.

C. *Facets of Quality: control vs. assessment*

There are many dimensions or aspects of quality. Some of them are quite objective and can be quantified. The other depend on how those objective aspects are perceived by specific communities depending on their objectives [15–18].

The Quality Control (QC) facet: the reported Pixel-level QC flags in the data are assigned during data processing according to algorithmic guess at usability of data point (some say it reflects the algorithm "happiness"). The Granule-level Quality facet reflects the statistical roll-up of Pixel-level Quality.

Quality assessment, on the other hand, is done by analyzing the data "after the fact" through validation, intercomparison with other measurements, self-consistency,

etc. The Product-level Quality facet is perceived based on how closely the data represent the actual geophysical state. And then there is the Record-level Quality: how consistent and reliable the data record is across generations of measurements

Different quality types are often erroneously assumed having the same meaning. However, there is no one-to-one correspondence between the above quality facets. Filtering out bad QC flags does not mean necessarily ending up with good data. For example, if the certain algorithm assumptions about the measurement conditions are not right, the corresponding data are not good while the QC flags might still indicate data being of the best quality.

D. Quality of Aerosol Data

Perceived quality of aerosol data depends on when and how the data are going to be used.

For example, if one is interested in a quality of satellite observations to monitor and evaluate heavy pollution events, then one should use the measurements performing well in heavy aerosol loading conditions. This also means that some factors affecting the aerosol measurement might be ignored in this case, e.g., surface properties are not important as under heavy aerosol loading the surface is not visible at all. However, an ability to distinguish thick aerosol layers from clouds becomes very important as heavy aerosols are more easily misclassified as clouds.

On the other hand, if interested in distinguishing between fine mode and course mode, it is important to account for different aerosol modes and also to know the absorptive properties of aerosols.

Another important aspect is a sensor's ability to measure aerosols close to the surface. If the sensor sensitivity is poor near the surface (it cannot see inside the Planetary Boundary Layer), then only some climatological data or data from models are used for that portion of the atmosphere. In this case, if the aerosol layer height cannot be determined, and even if the total aerosol depth is measured rather accurately, the uncertainty in air pollution measurement is rather high, i.e., the quality of these aerosol data for this particular purpose is poor [4].

E. Facets of aerosol data quality

Usually, uncertainty is the first thing scientists look for assessing quality of the data. Unfortunately, not a lot has been done in addressing uncertainty of remote sensing data. On the other hand, there are aspects of Level 2 and Level 3 data that can be captured and characterized to enable users of the data to evaluate these data and get some perception of quality of these data as applicable to their needs.

In case of aerosol properties measured by spaceborne and ground-based sensors, the following aspects can be identified (note: the semantic aspect of data quality-related terminology): besides the usual expected information on uncertainty, bias, error budget, etc.:

The simplest to identify and present are Completeness, Consistency, and Representativeness. They can be applied to various aspects of the data, for example:

Completeness:

- Spatial: MODIS covers more than MISR
- Temporal: Terra mission has been longer in space than Aqua
- Observing Condition: MODIS cannot measure over sun glint while MISR can

Consistency:

- Spatial (e.g., not changing over sea-land boundary);
- Temporal (e.g., trends, discontinuities and anomalies)
- Observing Condition (e.g., exhibit variations in retrieved measurements due to the viewing conditions, such as viewing geometry or cloud fraction)

Representativeness:

Neither pixel count nor standard deviation fully express representativeness of the grid cell value

It is much more difficult is to characterize spatial and temporal sampling. Fig. 1 illustrates examples of different data quality aspects:

- Completeness: MODIS dark target algorithm does not work for deserts
- Representativeness: monthly aggregation is not enough for MISR and even MODIS
- Spatial sampling patterns are different for MODIS Aqua and MISR Terra: "pulsating" areas over ocean are oriented differently due to different direction of orbiting during day-time measurements.

III. Experiences Developing a Semantic Representation of Product Quality, Bias, and Uncertainty for a Satellite Data Product

A. Easy web data access

Web-based science analysis and processing tools allow users to access, analyze, and generate visualizations of data while alleviating users from having directly managing complex data processing operations. These tools provide value by streamlining the data analysis process, but usually shield users from details of the data processing steps, algorithm assumptions, caveats, etc. To interpret the final analysis results correctly, user of these tools need understand how the data have been generated and processed; and what potential biases, anomalies, or errors may have been introduced during the processing. An easy data access without proper assessment of the joint data usage might be quite dangerous as it becomes easy to misuse the data.

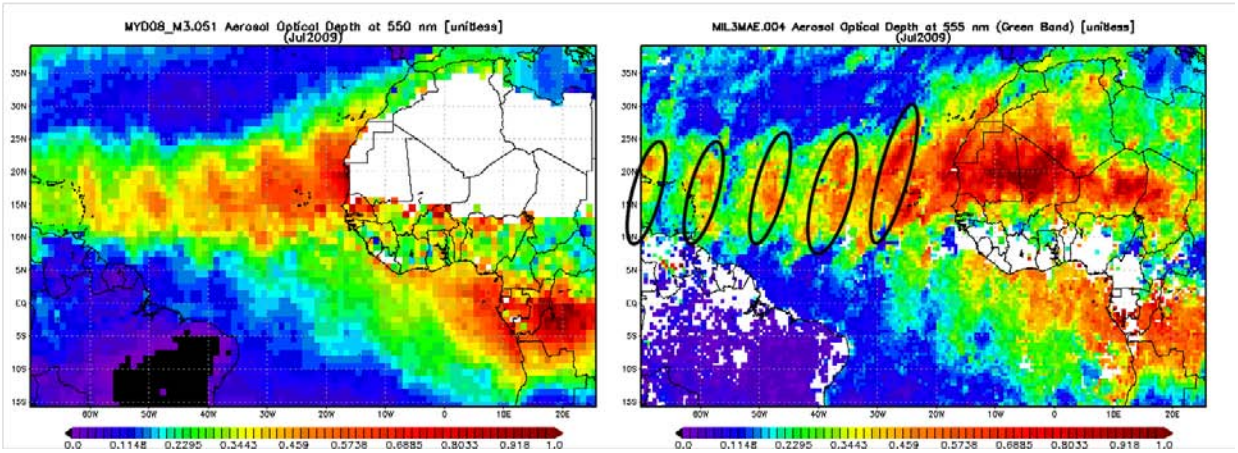

Fig. 1. Quality aspects illustrated by looking and comparing maps of monthly Aerosol Optical Thickness measured by MODIS Aqua (left) and MISR sensors.

B. The MDSA approach

The NASA ESTO-funded MDSA project (PI: G. Leptoukh) focuses on the value-added capabilities of the NASA Giovanni online tool for data access, visualization and analysis to improve usage of NASA's remote-sensing data, and specifically, addressing aerosol data.

The MDSA approach is to augment Giovanni with semantic web technologies and ontologies to support data inter-comparisons from different sensors or models, encode dataset variable characteristics and related quality to derive inter-comparison rules, and add data provenance (essential parameter details, quality and production caveats), and provide users of remotely sensed aerosol data with clear, cogent information on salient differences between data candidates for intercomparison, merging and fusion to enable scientifically and statistically valid conclusions. This could greatly enhance scientists' ability to perform valid comparisons, draw quantitative conclusions, and then merge/fuse data from multiple sensors.

Dealing with data quality is paramount to this project to address: Recently, the project has concentrated on trying to discover in literature and standardize assessment of MODIS AOT biases [5-8]. Biases can be related to spatial, temporal, vertical, pixel quality, clear sky and surface type issues.

IV. Technology for Capturing Quality Information

To characterize aerosol data quality, capture quality aspects from the literature and relations between these aspects, and develop rules behind those relations, the MDSA project has employed several technologies described below.

A. Mind Maps

Capturing quality information in tables has proved to be very inefficient as there are too many facets of quality information. Not only the number of facets is high, it also depends on a sensor measurement capabilities and specific of the retrieval algorithm. A breakthrough has happened when we switched to FreeMind and started mapping results of the latest papers on MODIS Aerosol Collection 5 validation. FreeMind allowed creating multi-dimensional snapshots of the various quality and condition aspects. We collected the most recent papers that address AOT biases over different regions and seasons [6, 7, 8] and extracted plots of correlation between MODIS and the ground-based Aeronet measurements and the corresponding statistical fitting results. The results were systematized by sensor, region, season, surface type, etc. Also, the available explanation of discrepancies between MODIS and Aeronet were analyzed and recorded.

B. Knowledge presentation

Ontologies describe and organize knowledge in a form that is machine-readable and provide a foundation for rulesets. The MDSA Ontologies focus on the genesis of science products, in particular, aerosol data. They include information about sensors, retrieval algorithms, averaging algorithms, data-day differences, etc. Ontologies for various levels of aerosol data are developed based on the information collected in FreeMind maps. Ontologies are scoped using the use cases and by assessing the current state of capability that end-user scientists were ready to use. The project advisor infers how "similar" datasets (or processes) are by rigorously comparing their ontological descriptions.

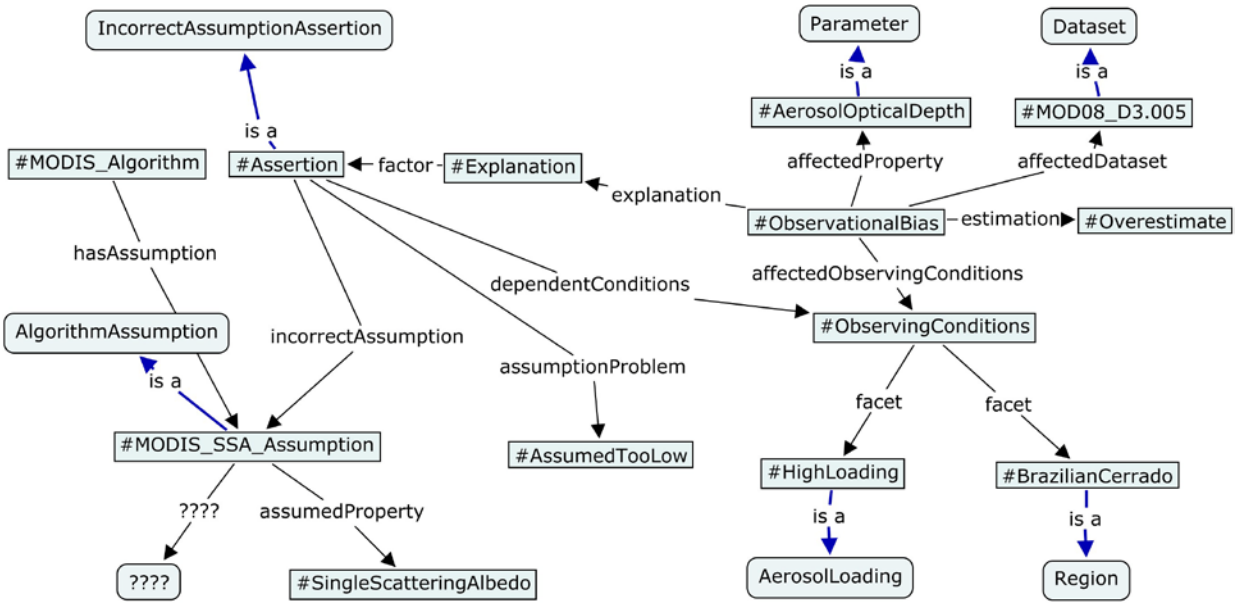

Fig. 2. MODIS Aerosol Bias Estimation ontology view. .MODIS overestimates AOD in fine-dominated cases where the observed SSA is greater than that assumed. For example, in the Brazilian Cerrado, the assumed SSA = 0.86 whereas more recent AERONET data suggests SSA = 0.91 (SSA = Single Scattering Albedo)

The following ontologies have been developed: Processing; Aerosol data; and Data Quality. To both validate and verify the developed ontologies and clearly identify suitable provenance encodings and presentation, a tabular form of ontology registration was developed.

In some cases, ontology alone is not enough to express the relationships between concepts. For particularly rich and/or complex relationships, inference rules may be needed. At this point, the MDSA project is populating the ontology and fleshing out the use cases with an eye toward seeing if the ontology and SPARQL queries alone will be enough to express how bias is related to observing conditions, location, aerosol loading, etc. If not, rulesets will be added in a manner similar to the initial MDSA findings on the effect of data-day and orbital parameters on MODIS Aqua and Terra daily aerosol comparisons [9, 10].

V. Presenting Quality

The MDSA project is a technology exploration project. Therefore, we do not aim to provide a comprehensive framework to capture and present all the possible aspects of data quality and provenance. Instead, we concentrate on representative use cases.

Based on the data quality ontology developed by the MDSA project, and the bias explanation ontology in particular (Fig 2), we have moved to the next stage of presenting the results of the literature review as a structured web page (Fig 3). The page (currently a mockup) contains necessary elements needed by computer to infer the necessary relations (based on user inputs) from the ontology-based knowledge base. Also (we hope) the elements displayed on the page provide sufficient information to users assess behavior (and quality) of MODIS Aerosol measurements over their selected area and time.

VI. Conclusions

It is very hard to characterize remote-sensing data quality. Different communities perceive and assess different and inconsistent measures of quality. Products with known Quality (whether good or bad quality) are more valuable than products with unknown Quality. Harmonization of data quality is even more difficult that characterizing quality of a single data product. Well-presented aspects of quality help users of the data to assess fitness-for-use correctly. Presenting various data quality aspects via web services is a very challenging task. The presentation needs to be translucent (showing only the important aspects of quality and provenance) and needs to be understood by scientists. The MDSA project is making progress towards assessing and presenting quality and provenance of multi-sensor remote sensing data by employing semantic web technologies.

The remote-sensing community as a whole needs to move forwards a common framework for data quality. We hope that the MDSA approach and methodologies employed for the prototypes can serve as a starting point for such a framework for web-based services.

Acknowledgment

G. L. thanks the NASA ESTO for providing support under the NASA ROSES 2008 program.

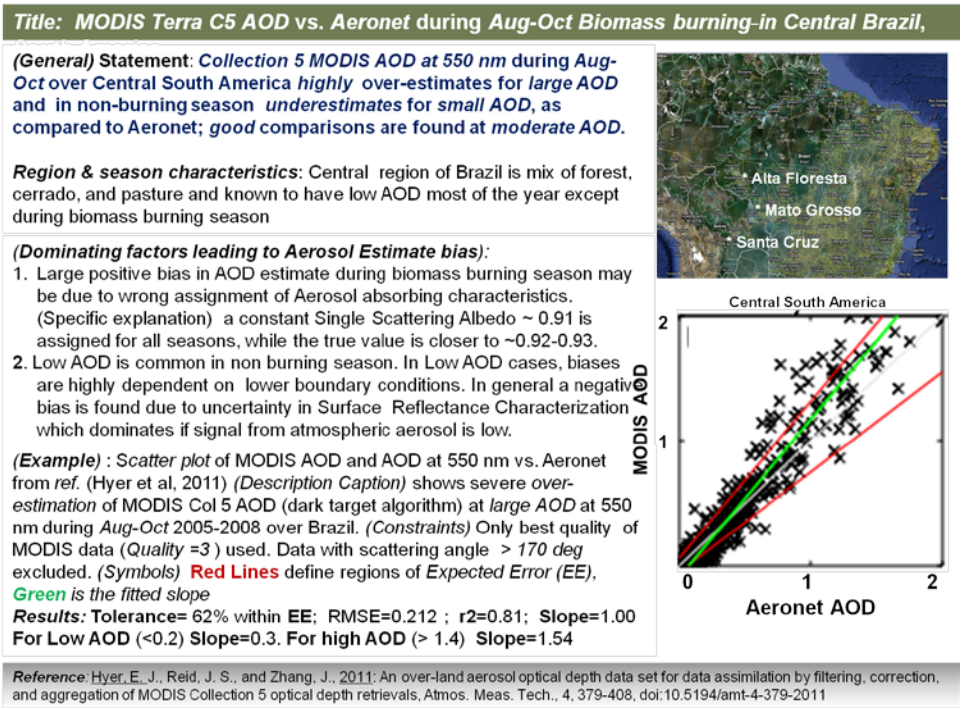

Fig. 3. A prototype of a structured MDSA presentation of Aerosol biases. The example is for MODIS AOT over Brazil.

REFERENCES

[1] J. Acker and G. Leptoukh, "Online analysis enhances use of NASA earth science data," *Eos, Trans. AGU*, vol. 88, 2007, p. 14–17.

[2] S.W. Berrick, G. Leptoukh, J.D. Farley, and H. Rui, "Giovanni: A Web Service Workflow-Based Data Visualization and Analysis System," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 47, Jan. 2009, pp. 106-113.

[3] C. Batini and M. Scannapieca, *Data Quality: Concepts, Methodologies, and Techniques*, Springer, 2010

[4] A. Prados, et al, "Access, Visualization, and Interoperability of Air Quality Remote Sensing Data Sets via the Giovanni Online Tool," *Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal of*, vol. 3, 2010, p. 359–370.

[5] M.I. Mishchenko and I.V. Geogdzhayev, "Satellite remote sensing reveals regional tropospheric aerosol trends.," *Optics express*, vol. 15, Jun. 2007, pp. 7423-38.

[6] E.J. Hyer, J.S. Reid, and J. Zhang, "An over-land aerosol optical depth data set for data assimilation by filtering, correction, and aggregation of MODIS Collection 5 optical depth retrievals," *Atmospheric Measurement Techniques*, vol. 4, Mar. 2011, pp. 379-408.

[7] R.C. Levy, et al, "Global evaluation of the Collection 5 MODIS dark-target aerosol products over land," *Atmospheric Chemistry and Physics*, vol. 10, Nov. 2010, pp. 10399-10420.

[8] T. Mielonen, et al, "Evaluating the assumptions of surface reflectance and aerosol type selection within the MODIS aerosol retrieval over land: the problem of dust type selection," *Atmospheric Measurement Techniques*, vol. 4, Feb. 2011, pp. 201-214.

[9] P. Fox and J. Hendler, "Changing the Equation on Scientific Data Visualization," *Science*, 2011, vol. 331, pp. 705-708.

[10] G. Leptoukh et al, "Sensitivity of Aerosol Multi-sensor Data Intercomparison to the Dataday Definition," in progress.

[11] G. Leptoukh, D. Lary, S. Shen, and C. Lynnes, "Sensitivity of Aerosol Multi-Sensor Daily Data Intercomparison to Level 3 Dataday Definition," EGU 2010, Vienna, May 2010

[12] R. C. Levy, et al, "A critical look at deriving monthly aerosol optical depth from satellite data," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, pp 2942-2956, 2009.

[13] G. Leptoukh and J.-P. Muller, "Data Quality, Error budget and Level 3 cases," WGISS/WGCV, Montreal, September 2010

[14] S. Zednik, et al, "A Semantic Provenance-aware Expert Advisory System in a Web-based Science Data Analysis Tool," Fall AGU, Dec 2010

[15] G. Leptoukh, "Towards Consistent Characterization of Quality and Uncertainty in Multi-sensor Aerosol Level 3 Satellite Data," Fall AGU, Dec 2010

[16] C. Lynnes and G. Leptoukh, "Ambiguity of Data Quality in Remote Sensing Data," Fall AGU, Dec 2010

[17] P. West, et al, "Experiences Developing a Semantic Representation of Product Quality, Bias, and Uncertainty for a Satellite Data Product." EGU, April 2011

[18] G. Leptoukh, Quality Issues in Multi-sensor Aerosol Level 3 Satellite Data, EGU, April 2011