Minutes: Product Quality Metrics Telecon, Held August 17, 2010

Greg Hunolt, updated, September 3, 2010

The minutes begins with a list of attendees at the telecon, followed by a summary, then telecon notes, and finally a comment on the draft minutes received by an MPARWG member.

A. Attendees:

There were 32 persons attending Product Quality Metrics Telecon:

Name	Project or Affiliation
Mark Anderson (for David Robinson)	Northern Hem. Snow and Ice Climate Data Records
Brian Beckley (for Richard Ray)	Multi-Mission Ocean Alt. Data for Climate Research
Wes Berg	Long-Term Precip Dataset w/Uncertainty Information
Corey Bettenhausen (for Christina Hsu)	Long-Term Aerosol Data Records
Jennifer Bohlander	Ice Velocity Mapping of the Great Ice Sheets: Antarctica
Clyde Brown – MPARWG Co-Chair	Database for Validation of Global Models of Atmos. Compos.
Saurabh Channan	ESDRs of Global Forest Cover Change
Gao Chen	Database for Validation of Global Models of Atmos. Compos.
Toshio Mike Chin	Multi-Sensor Ultra-High Resolution Global SST Field
Peter Cornillon	GAC and HRPT AVHRR Reprocessing to GHRSST
Kamel Didan	Vegetation Phenology & Enhanced Veg Index Products
Tom Farr	Definitive Merged Global Digital Topographic Data Set
Eric Fetzer	Multi-Sensor Water Vapor Climate Data Record
John Forsyth	Improvement of the NVAP Global Water Vapor Data Set
David Hancock	IceSat-2 SIPS
Gina Henderson (for David Robinson)	Northern Hem. Snow and Ice Climate Data Records
Greg Hunolt	SGT, support to Rama
Calli Jenkerson	Vegetation Phenology & Enhanced Veg Index Products
Steve Kempler	GSFC / GES DISC (a.k.a. GSFC DAAC)
John Kimball	ESDR for Land Surface Freeze-Thaw State
Mike Kobrick	Definitive Merged Global Digital Topographic Data Set
Gordon Labow (for Jay Herman)	Earth Surface & Atmos. Reflectivity from Multiple Satellites
Stephane Maritorena	Beyond Cholorophyll: Innovative Ocean Color ESDRs
Kevin Murphy	GSFC / EMS
"Rama" Ramapryian: Co-Chair MPAR	GSFC / ESDIS
Bill Rossow	Global Cloud Process Studies: ISSCP Data Analysis
David Roy	Web-Enabled Landsat Data & Surface Characterizations
Bernd Scheuchl	UC-Irvine
Deborah Smith	DISCOVER - Climate / Ocean Products & Visualizations
Ron Weaver	NSIDC
Frank Wentz	DISCOVER - Climate / Ocean Products & Visualizations
Victor Zlotnicki	ESDR of Changes in Earth Masses from GRACE, etc.

Rama thanked all of the attendees for taking time out from their busy schedules to participate in the telecon.

B. Summary:

This summary is based on the detailed telecon notes that follow below.

The telecon began with general agreement with Rama's goals for the telecon: "Agree on approach to providing Program Level metric(s) on usability of MEaSUREs products by the user community" and "Identify details to be worked out at follow-on discussions in October 2010 meeting in New Orleans".

The discussion that followed was focused on several topics:

1) Approach

There was general agreement that the approach to product quality metrics should involve development of information about specific product quality criteria, such as those suggested by Dr. Frouin, from which a small number of condensed programmatic level metrics could be produced and tracked over time.

There was general agreement that, using Dr. Frouin's input as a starting point, a set of criteria or categories pertaining to product quality could be arrived at. Then, for each criterion, a set of questions could be developed, in the form of 'has the project done this' where 'this' would be a specific item or task (e.g. is each product validated, is each well documented). This would amount to a checklist that the project would gradually complete. The degree to which the questions for a criterion are answered 'yes' at any point in time would measure the degree to which the project meets that criterion, and a 'high', 'medium', or 'low' or similar completeness scale could be developed. As the work of the project proceeds, more of the questions for the criteria would be answered in the affirmative, providing a measure of the project's progress toward fully meeting the criteria. Condensed or summary program level metrics could be developed based on the project's progress on the criteria (i.e. their answers to the questions), and these program level metrics could be tracked over time and also aggregated as a measure of the MEaSUREs program's overall progress.

The discussion on approach reached two conclusions for the telecon:

1) The MPARWG will take the criteria table and rework it at the October meeting.

2) The MPARWG will work on how to produce a condensed set of program level metrics.

2. Concerns with "Science Quality"

"Science quality" has a subjective element: the level of science quality depends on what the product is being used for – a given product may be of sufficient quality for some uses but not for others. Accuracy can be quantified objectively (some examples were cited), e.g. through validation, i.e. confirmation of accuracy by comparison with independent reference measurements that leads to error bars or other appropriate measures. It was noted that projects need to address biases and structural or situation dependent errors, and that overall error bars are

too crude. Simple error bars or single error/uncertainty numbers are for the most part meaningless when dealing with spatial & temporal variant data (global or regional). Error and uncertainty need to be put in their proper spatial and temporal context for them to be useful and meaningful.

Quality and accuracy are not the same thing, but are two distinct topics that should be handled separately. Quality assessment is not the same thing as accuracy assessment, so there should be separate accuracy and quality metrics, so you should consider accuracy, quality, and usability separately.

It was suggested that scientific value be used as a criterion instead of quality or accuracy. It was also noted that science value / quality would not be static but would vary with time. A new product's quality might be seen as very high because it provides a first look at a parameter(s), e.g. the very first SST product. Subsequently that first SST product might be seen as lower in comparison to later versions or entirely new SST products.

3. Concerns Regarding the Criteria

It was suggested that from Robert Frouin's table that only the criteria of accuracy, consistency, and completeness are appropriate for the projects to address as data producers (perhaps with one more 'other issues' category) and that that other criteria were not appropriate for the P.I. to address. Instead criteria like uniqueness, interpretability, relevance and accessibility were best answered through some form of independent review or community feedback, perhaps with involvement of the program scientists, or perhaps by having the projects review each other (i.e. play the role of users of other projects' products). If left to the projects alone, the result would likely be a meaningless self evaluation, with "yes" responses to all of the questions.

Reproducibility was suggested as a criterion, but there was a general reaction that reproducibility was very difficult, especially when you have a succession of different versions of a project (which may correct errors present in previous versions or be based on improved algorithms). It would be difficult to preserve all of the input data as it existed at the time each version of a product was produced, as well as the algorithm software used for each version.

4. Concerns Regarding Documentation

It was noted that quality metrics cannot replace full documentation, and that the program should require and fund the projects to produce detailed documentation, not withstanding that users may ignore a product's documentation and then claim that the product was not usable. [This begs the question of a standard for documentation – at a minimum, a checklist of items that should be included in the documentation that the project could complete. Such a checklist could be vetted by user groups. Completion of all of the items on a documentation checklist would not guarantee that users would read the documentation, but would certify that the project had made the information available, which would meet the program's requirement for the project.]

It was noted that usability of a product package was distinct from usability of the delivery mechanism used by the user to obtain the product package – he noted that the background paper identified accessibility as a separate consideration.

5. User Feedback Problems

While quality feedback from users is needed, agencies do not fund user evaluation of data sets, so projects have to get people to do it voluntarily, which is very difficult. It was suggested that projects should provide easy and simple ways for users to provide feedback.

Projects noted the difficulty of dealing with user feedback. It was observed that if a project was to get 500 feedbacks, the project would have no-one with the time to go through and analyze them unless the project were funded to do that. Even so, some attendees reported that user feedback has been very valuable to their projects.

C. Telecon Notes:

These are "raw" notes tracking the discussion as it happened.

Rama began by going over the four introductory charts. There was general agreement with the telecon goal given on the first chart: "Agree on approach to providing Program Level metric(s) on usability of MEaSUREs products by the user community" and "Identify details to be worked out at follow-on discussions in October 2010 meeting in New Orleans".

David Roy began the discussion by observing that the background material for the telecon used accuracy and quality if they were interchangeable. He said that instead they are distinct topics that should be handled separately.

Peter Cornillon noted that Rama assumed in his introductory statement that science quality was assured, but Peter pointed out that the level of science quality depends on what the product is being used for – a given product may be of sufficient quality for some applications but not for others.

Peter also felt that the question of the usability of a product package was distinct from usability of the delivery mechanism used by the user to obtain the product package – he noted that the background paper identified accessibility as a separate consideration.

David Roy noted that the MODIS Land Team was responsible for developing the product maturity levels for the EOS program. He observed that:

1) Validation is confirmation of accuracy by comparison with independent reference measurements that leads to error bars or other appropriate error measures.
2) Maturity levels distinguish between small scale or single point vs. overall systematic or global comparisons with reference data.
3) Quality assessment is not the same thing as accuracy assessment, so there should be separate accuracy and quality metrics; accuracy, quality, and

usability should be considered separately.

Gao Chen suggested that we consider scientific value as a criterion instead of quality or accuracy.

Bill Rossow noted that evaluation of science value / quality was not necessarily static but would vary with time. A new product's quality might be seen as very high because it provides a first look at a parameter(s), e.g. the very first SST product.

David Roy observed that the quality of that first SST product might be seen as lower in comparison to subsequent versions or entirely new SST products.

Bill Rossow agreed with the need to treat accuracy and quality separately.

David Roy emphasized that while accuracy can be quantified objectively, quality is inherently subjective because if depends on what the product is used for.

Gao Chen noted that his project used aircraft data - asked how the accuracy of the data can be assessed.

Gao Chen suggested that quality of the products might have more to do with the accuracy of the input data from which they are produced.

Rama noted that projects are producing particular products in long term time series. It is the quality of the outputs from the MEaSUREs products rather than the input quality that we would be concerned with for purposes of this discussion.

Bill Rossow, referring to the lists in the background document suggested an approach to arriving at metrics. For each category, a question or questions could be written along the lines of 'was this [particular thing] done by the project?' This sets out a series of tasks for the project. The compilation of answers to those questions could then become the basis for a metric measuring the project's progress in completing those tasks. This progress could be tracked over time.

Deborah Smith agreed with Bill's suggested approach – noting that you can see progress as the tasks are completed – but she observed that you may only get a self-evaluation.

Rama asked can you get the community to answer some of the questions.

Bill Rossow noted that agencies do not fund user evaluation of data sets, so projects have to get people to do it voluntarily which is very difficult so you are likely to only get self-evaluation.

Bill also suggested use of a validation checklist.

Rama suggested that the program scientists could participate in the evaluation.

David Roy said that you really need to get users to provide quality feedback.

David also agreed with the idea of a validation checklist, but noted that validation won't happen without funding.

Peter Cornillon said feedback is difficult to get because we haven't provided users with an easy and simple way to provide feedback; we should do this.

Bill Rossow said trying to get user feedback won't work because people don't have time to do it voluntarily, so they won't unless they are funded. Not only that, if a project was to get 500 feedbacks, the project would have no-one with the time to go through and analyze them unless the project were funded to do that.

An attendee observed that what you would get would be a variety of biased inputs of uncertain value anyhow.

David Roy agreed that dealing with user feedback is a difficult task.

Peter Cornillon reported that user feedback has been very valuable and useful to his projects.

An attendee also reported that user feedback has been very valuable to the PO.DAAC at JPL.

After some additional back and forth on user feedback, Frank Wentz suggested getting back to the goal of the telecon.

Rama re-iterated the telecon goals.

Frank Wentz noted that in getting to the programmatic level metrics the meanings of "high", "medium", and "low" would have to be defined.

David Roy said that using Bill Rossow's question and answer approach, getting answers would then drive the "high", "medium" and "low" gradation.

Deborah Smith agreed with the question and answer approach.

An attendee observed that user input would be needed on accessibility and support levels.

David Roy suggested that from Robert Frouin's table that the categories of accuracy, consistency, and completeness be used, with one more 'other issues' category. He observed that items other than accuracy, consistency, and completeness were not appropriate for the P.I. to address.

A representative from the ESDSWG Software Re-Use group suggested that the software re-use levels could serve as a model. This prompted a general reaction that assessing suitability for re-use of software was not comparable to assessing product usability.

Bill Rossow noted that users may ignore documentation and then complain that the product was not usable. The metrics should measure if products are validated and documented, which you would be able to determine from the answers to the task questions.

An attendee noted that quality metrics should help the user use the data, i.e. provide guidance to the users.

.

Bill Rossow responded that quality metrics cannot replace detailed documentation.

David Roy suggested that documentation include a one-page summary that would point to other more detailed documents. The project can prepare the one page summary, user guides, and detailed documentation, but can't make the user read them.

Bill Rossow said that the program should require and fund the projects to produce detailed documentation, not withstanding that users may ignore full documentation. He said we can develop a checklist of what the projects can do.

Clyde Brown returned to David Roy's points about the items in Robert Frouin's table, and asked why a project could not address the category of interpretability.

David Roy responded that projects will just say "yes", which would not be meaningful or useful. He noted that a way to approach the categories of uniqueness, interpretability, relevance and accessibility would be to get an independent review, either by community representatives or have projects review each other.

Rama asked would a good agenda topic for the October meeting be to go over the criteria and questions for each?

Kamel Didan suggested adding context to the description of the criteria when dealing with quantitative metrics like error and uncertainty. Simple error bars or single error/uncertainty numbers are for the most part meaningless when dealing with spatial & temporal variant data (global or regional). Error and uncertainty need to be put in their proper spatial and temporal context for them to be useful and meaningful.

Bill Rossow noted that we needed a deeper treatment of "error" than provided in our simple start.

Peter Cornillon suggested that projects focus not on subjective quality but objective accuracy measures, e.g. point comparisons of satellite derived SST measurements to buoy measurements, and address biases in the data.

Bill Rossow responded that projects need to address biases and structural or situation dependent errors, and that overall error bars are too crude.

Saurabh Channan suggested that reproducibility would be another good quality metric.

Rama agreed.

There was a general reaction that reproducibility was very difficult, especially when you have a succession of different versions of a project (which may correct errors present in previous versions or be based on improved algorithms).It would be difficult to preserve all of the input data as it existed at the time each version of a product was produced, as well as the algorithm software used for each version.

Rama offered the following conclusions based on the telecon discussion:

1) We will take the criteria table and rework it at the October meeting.

2) We will work on how to produce a condensed set of program level metrics.

Peter suggested that we see what the SST and other are doing to work on assessing accuracy. Rama asked Peter to pursue that for the group and Peter agreed to contact the appropriate group leads.

D. Post-Telecon Comment:

Chung-Lin Shie provided the following comments on the draft telecon minutes:

(1) I fully endorse the argument that "Quality and accuracy are not the same thing .... Quality assessment is not the same thing as accuracy assessment, so there should be separate accuracy and quality metrics". Actually, I personally believe that we (the PIs or scientists) by all means are all looking for an ultimate goal of producing an ideally "accurate" product, yet more practically or realistically we might at most produce better and better "quality" products year-in year-out. Nonetheless, accuracy assessment is still needed and demanded to ensure that our products are heading the "accurate" direction. I also think Gao Chen's "scientific value" suggestion was a useful and very practical criterion that may be considered as a secondary criterion, along with scientific and accuracy assessment.

(2) As for the user evaluation of data sets, it certainly would be one of the most difficult issues to resolve, especially based on my own current experiences with my recently completed GSSTF2b product. I have had quite a few "pilot" users of GSSTF2b, but only received a couple of feedback so far. However, indeed, I consider each individual user feedback useful and constructive regardless how important it could be. I generally agree with Bill Rossow's frank comment and suggestion of "self-evaluation" because of the potential "difficulty of getting user evaluation of data sets".

(3) I would also like to echo on the point brought up by Gao Chen that the product quality might have more to do with the accuracy of the input data. I had actually advocated/discussed this point/issue during both my talks at our two previous sessions in AGU meetings (2008, 2009). I invented and used a term "Rice Cooker Theory" as an analog/metaphor. In short, my point was that the quality (tasteful or not) of rice soup (our products) you were to produce/cook would highly depend on the quality (good or bad) of the raw rice you put in (input) regardless how good

(even perfect) the rice cooker (our models or algorithms) would be. E.g., my produced latent heat flux of GSSTF2b highly relies on my retrieved surface air humidity, which was found significantly dependent of the SSM/I brightness temperature (produced/retrieved by Frank Wentz' RSS) that I used as input. The quality (let alone accuracy) of brightness temperature would certainly affect my produced latent heat flux. I totally agree with your point that the output products were our (MEaSUREs) major concern, I however suggested that the PI's may briefly address the input issue (if they do have any like my case) in their metrics that should help users better using and analyzing the products.