GROUP ON EARTH OBSERVATIONS

QA4EO A QUALITY ASSURANCE FRAMEWORK FOR EARTH OBSERVATION

A guide to comparisons – organisation, operation and analysis to establish measurement equivalence to underpin the Quality Assurance requirements of GEO

QA4EO-QAEO-GEN-DQK-004

GROUP ON EARTH OBSERVATIONS

QA4EO A QUALITY ASSURANCE FRAMEWORK FOR EARTH OBSERVATION

QA4EO guide: QA4EO-QAEO-GEN-DQK-004 **Community Approved**

A guide to comparisons – organisation, operation and analysis to establish measurement equivalence to underpin the Quality Assurance requirements of GEO

Author:	Nigel Fox
E-mail:	Nigel.Fox@npl.co.uk
Editor:	Marie-Claire Greening
E-mail:	marie-claire@greeningconsulting.co.uk

Issued under Authority of:	QA4EO
Issue no:	Version 4.0
Date of issue:	27 March 2010

Summary of Changes since last issue:

Version 3.0 now includes edits and comments made on version 2.0 by GSICS in March 2009.

Version 4.0 incorporates edits and comments made on version 3.0 by WMO and GSICS August – November 2009.

Table of Contents

1 Abstract

This document provides guidance on how best to organise, operate and analyse the results of a comparison to evaluate the equivalence of different techniques, instruments and/or teams when used to measure or nominally process the same information. Comparisons are an essential tool within any Quality Assurance (QA) strategy as they provide a source of unequivocal information on differences and biases associated with similar activities. In general, comparisons are not a test of "right or wrong" but a means to identify and sometimes understand differences. This guide is written based on best practise guidance established by the Comité International des Poids et Mesures (CIPM), and its associated technical committees, for performing comparisons of key quantities between National Metrology Institutes (NMI) to underpin the Mutual Recognition Arrangement (MRA) of the metre convention (http://www.BIPM.org/). Comparisons need to be fair and unbiased but should also provide unequivocal evidence of performance to underpin QA statements. It is therefore essential that they are organised, and all results analysed, in a transparent and consistent manner.

2 Scope

The most effective means of demonstrating traceability and establishing an overall uncertainty budget is to perform a formal comparison against a reference standard, which, by virtue of its position in the traceability chain hierarchy, has an uncertainty significantly smaller than that being declared or that required by the process under test. Such a comparison may in general terms be called a "calibration".

This document provides guidance on the general approach that should be followed when organising comparisons to evaluate techniques or processes. It covers all "peer to peer" comparisons, both bilateral and multi-partner, but not those considered to be a "calibration". In such situations the reader is directed to QA4EO-QAEO-GEN-DQK-003.

Comparison undertaken with respect to peers are perhaps more appropriate for many situations within the Earth Observation (EO) sector. The objective of this type of comparison would be to identify and evaluate any biases between participants, not necessarily judge "right" or "wrong". Such comparisons should ideally be carried out to evaluate the performance of all aspects of a data processing chain. However, because of the relative organisational complexity and consequential cost of comparisons they are only likely to be used selectively. In this way they can provide a useful "sampling" of overall performance and, if chosen with care, they can be used to evaluate an amalgam of activities in one go. For example, if all similar satellite imagers view the same reference standard (e.g. an icefield like Dome C in Antarctica for optical imagers or the Amazon rainforest for SAR) this can test the end to end processing chain for radiometric gain. It may not test any individual aspect of the sensor (e.g. algorithm) independently, but if all proves consistent, it may be adequate to infer "suitability" of all the subcomponents and processes. If a bias is found, it can serve to highlight areas where further detailed work may be needed.

In all cases, comparisons need to be established with rigour and transparency to avoid unnecessary bias from the comparison process itself and to ensure that all participants are treated fairly and equally. Comparisons need to be carefully designed so that they can be carried out "blind" and yet still avoid issues due to simple typographical errors. The potentially large cost of organisation and participation should also be recognised.

In many cases there is no *a priori* correct answer and so a process should be adopted to establish a "comparison reference value" (CRV) for the comparison to which all results can be compared, in a fair but scientifically appropriate manner.

If such comparisons are to be used as a means of providing key evidence to underpin QA4EO, as they ideally should be, it will be important to ensure that all results are made public in a timely manner. Effort must also be made to facilitate participation by all who need to take part. This may best be carried out through a series of small linked comparisons, organised by geographical region, or alternatively through one large global exercise. In general, they will be organised under the authority of an agency or international organisation. The organising authority, which may or may not be the source of funds, will be responsible for approving the results and participants of the comparison. However, it is recommended that all results and also invitations to participate are made publicly available via appropriate communication mechanisms (e.g. the QA4EO website or other specialist portals serving the GEO community).

To ensure that such a system of comparisons is manageable and affordable but still adequate to underpin the QA4EO process, two generic types of comparison can be established:

1. **Key comparisons:** designed to test the principle aggregated techniques of the scientific field and a means of evaluating overall system bias. Such comparisons

will be organised under the auspices of an appropriate international body representing a GEO community.

2. **Comparisons:** all other comparisons organised under the auspices of any agency or organisation (including international bodies required for key comparisons) and which are in general designed to test more specific detailed aspects of a data processing chain.

This document provides a detailed structure for both administrative and technical organisation of comparisons and the analysis and reporting of results. It has been largely adopted from the best practise that has been established by National Metrology Institutes (NMIs) and is relatively formal in nature. However, this should be interpreted flexibly to serve the needs of the comparisons' stakeholders and in particular to take account of the level of maturity of understanding of the measurand or process that the comparison is evaluating.

3 Terminology

All terms within this document are based on internationally-agreed definitions that are, in many cases, derived directly from formal standardising bodies such as the International Organization for Standardization (ISO). These agreed definitions can be found on the QA4EO website (http://QA4EO.org/).

4 Background and Context

This key guideline is written as part of a set, based on the adoption of existing best practise, to form a Quality Assurance Framework for Earth Observation (QA4EO). The QA4EO was developed to meet the current and aspirational needs of the societal themes of the Group on Earth Observation (GEO)'s Global Earth Observation System of Systems (GEOSS). It was prepared as a direct response to GEO task DA-06-02 (now DA-09-01-a) to "Develop a GEO data quality assurance strategy, beginning with space-based observations and evaluating expansion to *in situ* observations, taking account of existing work in this arena".

5 Outcomes

The outcome of following this key guideline will be a set of comparisons and subsequent results that will provide evidence to identify biases and allow the evaluation of uncertainty in traceability of associated measurements or processes. This evidence will be fair and representative of the participants' capabilities at the time of performing the comparison. The results also provide the community with an assessment of the current state-of-the-art of a particular type of measurement or process.

6 Inputs

There are no specific identifiable inputs required by this procedure to enable the outcomes to be achieved.

7 Standards and Traceability

The procedure outlined in this document has no quantitatively assessable outcomes in its own right and there are no appropriate reference standards to which traceability should be demonstrated. However, this document will make reference to the use of existing documentary standards (where appropriate) within its guidance, e.g., vocabulary [1], Uncertainty analysis [2] and SI traceability [3]. As time progresses, new documentary standards and "best practises" may be developed and adopted by QA4EO that are applicable to the activities described in this document. Therefore, it is recommended that the latest version of this document be reviewed for changes prior to its use.

1. QA4EO endorsed vocabulary http://QA4EO.org/

2. http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=45315.

3. BIPM MRA http://www.bipm.org/

Any comparison carried out following these guidelines will necessarily select and make use of appropriate "reference standards" to perform the task that they describe.

8 Initiating and Organising a Comparison

8.1 Introduction

Comparisons are required to provide evidence to support claims on traceability, uncertainty and as a means of evaluating any bias from sensors or processes. Ideally (peer to peer) comparisons should be carried out to test all aspects of the data processing chain, complementing specific local calibration or characterisation activities. However, to ensure that such a system of comparisons is manageable, affordable and adequate to underpin the QA4EO process, "key comparisons" will be established. These key comparisons will serve to test the principle techniques of the scientific field, although they will be relatively few in number and will be carried out under the auspices of the appropriate international body representing the GEO community.

The procedure described in this section for selecting, conducting and evaluating comparisons, including the detailed technical protocols and periodicity of the comparisons, are designed to ensure that:

- the comparisons test all the principal techniques in the field;
- the results are clear and unequivocal;
- the results are robust;
- the results are easy to compare with those of corresponding comparisons carried out within different geographical regions;
- overall, the comparisons are sufficient in range and frequency to demonstrate and maintain equivalence between the participating organisations.

This document sets out the broad process that should be followed in conducting formal "peer to peer" comparisons, as opposed to "calibrations", and includes both bi-lateral and multi-participant versions. It is supplemented by the detailed technical protocols written for the individual comparisons (see QA4EO-WGCV-IVO-CLP-002 as an example). Although written to scope the detailed process, which is essential for "key comparisons", the principles can be adopted for all types of comparison. In principle, "informal comparisons" can, and ideally should, follow the same rigour, but where the process is simply being carried out as a "scientific exercise" and not being used for any QA based activity it may be less important to adopt some of the formality described here. In particular, the need to do "blind" measurements may not be appropriate. However, the reader may find the structure useful as a checklist to aid the process.

During comparisons it is important that up-to-date information on the progress of the comparison be readily available to all appropriate stakeholders. This implies that at least the participants, any funding organisation and the approving authority be regularly informed by the pilot institute about the status of each comparison. This can most easily be fulfilled through utilising the QA4EO website (http://QA4EO.org/) or any appropriate community specific portal, e.g., the GEO / CEOS calibration and validation (Cal/Val) portal (http://calvalportal.ceos.org/).

The following process has been adapted from that established by NMIs and so is relatively mature and formalised in terms of its structure and reporting requirements. There are a variety of good examples of comparisons undertaken within the EO sector. However, whilst in many cases these comparisons are fully consistent with the principles outlined in this document, the consistency and style of their reporting differs significantly. To ease use and to aid harmonisation, it is recommended that the formalism of the approach described in this document is implemented in all comparisons.

At this point it is worth referencing a recent example of a comparison exercise carried out on "radiative transfer models". In this case the pilot laboratory, the Joint Research Centre of the European Union, showed that the principles described here can be applied equally well to algorithms and software [1].

8.2 Types of Comparison

There are two broad types of comparison (structure as opposed to status). In the first are those comparisons for which the reference standard to be compared is assumed to have long-term stability. In the second category are those for which long-term stability cannot be assumed. The procedures for conducting the comparisons and for evaluating the results may differ between the two cases.

Comparisons of stable (long-term) reference standards are normally carried out bilaterally and on a continual basis at the convenience of the participating institutes. Typical of these comparisons might be the viewing of a community-specified reference test site, e.g., the Amazon rainforest, an ocean buoy or even the Moon. The procedure for carrying out this type of comparison should be available from the QA4EO website and results submitted to a nominated "pilot organisation" on a regular basis. Comparisons of this nature provide a means for continuous assessment of Quality Indicators (QIs) and are strongly encouraged.

Other comparisons in which the reference standards are not assumed to have long-term stability make up the majority of comparisons and are carried out under a strict time schedule. This enables all the participating institutes to make their measurements within a fixed period of time. These comparisons may require travelling standards to have good short-term stability and be robust to transportation. Much of the detailed text in what follows applies mainly to this type of comparison. In the EO sector it may be more common and appropriate for the participants to visit a single location and view a common reference standard instead. In the context of this key guideline it simply changes the "travelling aspects" to the participants' equipment rather than the reference standard. It also changes some of the timing and reporting aspects since this latter type of comparison is organisationally more difficult and, as a consequence, costly.

In all cases comparisons may have any number of participants ranging from a total of two (a bilateral) or more (multi-participant).

8.3 Responsibilities for Choosing Key Comparisons

It is the responsibility of international bodies representing specific GEO communities to determine which, if any, comparisons in their technical discipline should be established and designated as "key". Such "key comparisons" should be chosen to enable the principle measurement techniques in the field to be tested. On the basis of the results of the key comparisons, statements of equivalence or knowledge of bias can be made. This will cover a wide range of measurements that use similar techniques, not just the measurements directly tested by a key comparison. The periodicity of the comparisons should be chosen to ensure continuity of confidence of the equivalence statement without overloading the participating organisations. In some cases this may be very regular (e.g., monthly), whilst in others less frequency may be acceptable (e.g. annually or greater).

QA4EO, under guidance from the GEO members, is responsible for maintaining a list of key comparisons. Of course, any organisation and/or region can decide to organise a comparison on any topic at any time, and this is encouraged, but these will not be designated "key comparisons". A key comparison would be an important community event and, as such, it should be regarded as a major focal exercise with dedicated effort from all relevant parties. Key comparison events would therefore not be very frequent, although when they did occur they would be a priority exercise for the community being targeted.

8.4 Initiating a Key Comparison

Key comparisons are initiated by the appropriate international body and notification should be given to the QA4EO secretariat. Within each broad GEO community there will be sub-groupings of specialists who will ultimately be responsible for identifying and proposing comparisons, both general and the more specific candidate key comparisons.

However, to ensure that this selection process is carried out in a controlled manner, the decision on which should be considered "key" should be taken by an organisation designated by GEO.

For each comparison, a pilot institute should be identified to take the main responsibility for running the key comparison. The pilot will be responsible for the technical protocol, analysis and reporting of the comparison; this may or may not be the local organiser. The pilot needs to be considered as providing an independent view on all aspects of the comparison process, but this does not preclude them from also participating.

- In drawing up the provisional list of participants and an approximate timetable, the organising authority should ensure that an adequate number of participants from each of the main geographical regions take part so that any corresponding regional comparisons are properly linked to the key comparison.
- In some key comparisons the number of participants may need to be limited for technical reasons.
- Two or three institutes from the provisional list should be nominated by the organising authority to help the pilot institute in drawing up the technical protocol and timetable for the comparison.
- The timetable of this and any other comparisons decided by the organising authority should be discussed to ensure that the work load of the whole set is not too great for the participating and pilot institutes and that the results will be available in a timely manner.

8.5 Initiating a Comparison: General

Comparisons can be organised by any agency or organisation and may be proposed to a GEO-approved authority as a "key comparison" (see Section 8.4). The initiating agency or organisation takes responsibility for authorising the results of the comparison (although they can seek advice through QA4EO or others) to ensure that it meets the requirements of this and other appropriate guidelines. Such comparisons should be registered with the QA4EO secretariat and receive a unique comparison identifier. It is the responsibility of the pilot to notify the QA4EO secretariat of the intent to carry out any comparison.

8.6 Organisation of a Comparison

The organisation of a comparison is the responsibility of the pilot institute, assisted by two or three nominated participants. The first task of this small group would be to invite participants, to draw up the detailed technical protocol for the comparison (see Section 8.7) and to dispatch the protocol. The draft protocol must be sent to the chairman of the relevant authorising organisation, or their nominee as appropriate, for review. The invitation to participate should be sent directly to the delegates of all member institutes of the organising authority in the case of a key comparison. Copies of the invitation and the draft protocol should also be sent to the QA4EO secretariat and posted on the QA4EO website. This latter process should also be carried out for non-key comparisons.

The main points to be decided by the small group headed by the pilot institute are the following:

- The list of participants with full details of mailing and electronic addresses.
- The "reference standard" or "reference standards" to be used in the comparison.
- The need or otherwise for any preliminary work to be carried out amongst a restricted number of participants to verify the performance of the reference standard or methodology.
- The pattern of the full-scale comparison. This can be achieved through:
    - the simple circulation of a single travelling standard around all the participants,
    - the sending of an individual travelling standard directly to each participant from the pilot institute or from each participant to the pilot institute or
    - for all participants to visit a common reference standard or some combination of these.
- The starting date, detailed timetable, means of transport and itinerary to be followed by each travelling standard (if appropriate). This starting date is subsequently referred to as the starting date for the comparison.
- The procedure in the case of failure of a travelling standard or a measurement campaign due to external factors, e.g., weather.
- The procedure in the case of an unexpected delay at a participant institute.
- The customs documents or logistics that may need to be in place to facilitate the comparison.
- Plans for "bilaterals" or "follow-ons" in the event of a serious unexpected bias.

8.7 The Technical Protocol for a Comparison

The pilot institute together with two or three nominated participants should draw up the detailed technical protocol. The technical protocol is an important part of the comparison and specifies in detail the procedure to be followed for the comparison. It is important to remember, however, that the purpose of a comparison is to compare the measurement or process as performed by the participating institutes, not to require each participant to adopt precisely the same methodology and procedure for doing the task. The protocol should, therefore, specify the procedures necessary for the comparison, but not the procedures used for the measurement of the reference standard being used for the comparison unless this is a *de facto* process.

Among the points that should be treated in the protocol are the following:

- Detailed description of the reference standard being used: make, type, size, weight, characteristics, location, etc., and any technical data needed for its use.
- Advice on using or viewing the reference standard, including handling if appropriate.
- Any tests to be carried out before measurement.
- The conditions of use of the reference standard during measurement.
- Instructions for reporting the results.
- A list of the principal components of the uncertainty budget to be evaluated by each participant and any necessary advice on how uncertainties are estimated (this is based on the principles laid out in the Guide to the Expression of Uncertainty in Measurement, published by ISO – see QA4EO-QAEO-GEN-DQK-006). In addition to the principal components of the uncertainty, common to all of the participants, individual institutes may add any others they consider appropriate. Uncertainties are evaluated at a level of one standard uncertainty and ideally information should be given on the number of effective degrees of freedom required for a proper estimate of the level of confidence.
- The traceability to SI (or other agreed community reference) of each reference standard or instrument participating in the comparison.
- A timetable and procedure for the communication of the results to the pilot institute. Early communication helps to reveal any problems with the reference standard during the comparison. For example, when comparisons are carried out as part of a campaign, i.e., where all attend the same location, it may be

appropriate for the results to be delivered to the pilot in near real time to allow analysis, etc., to be undertaken whilst participants are still present. In such a situation it may be possible for a participant who has any "issues" to take part in a "follow-on" bilateral within days of a first comparison.

- Financial aspects of the comparison, noting that in general each participating institute is responsible for its own costs for the measurements, transportation and any customs charges, as well as any damage that may occur to its own property or any comparison property whilst in its possession. Overall costs of the organisation of the comparison including the supply and access to any reference standard or transfer devices are normally born or arranged by the pilot institute.

An example technical protocol is provided in QA4EO-WGCV-IVO-CLP-002, which relates to the CEOS comparison of instrumentation and methods used to measure surface emitted IR radiance / brightness temperature. This is relevant to the calibration and validation of satellites where spectral irradiance at the primary realisation level through the use of circulated incandescent lamps as transfer devices is required. Other EO-specific protocols will be available via the QA4EO website as they become available.

8.8 Reporting the Results of a Comparison

The participating institutes must report the results of a comparison to the pilot institute as soon as possible, but at least within six weeks after the measurements are completed unless the protocol specifies differently. The measurement results, together with the uncertainties and any additional information required, should be reported in the format given in the instructions as part of the protocol, usually by completing the standard forms annexed to the instructions. Where "real time" reporting and analysis is anticipated, participants must send details on their intended measurement procedure, traceability and uncertainty, including estimates of anticipated repeatability, prior to the start of the comparison to allow the necessary peer review process (pre-draft A; see section 8.9) to take place.

8.9 Preparation of the Report on a Comparison

The pilot institute is responsible for the preparation of a report on the comparison. The report passes through a number of stages before publication and these are referred to here as drafts A and B. The first draft (draft A) is prepared as soon as all the results have been received from the participants. It includes the results transmitted by the participants, identified by name, and is confidential to the participants. The bulk of draft A can, in principle, be completed in advance of the comparison for situations where the comparison will be analysed in near-real time.

The second draft (draft B) is subsequently prepared for the final review body. In the case of a key comparison this would be the authorising body, in others it might be an individual agency or other nominated approving body. Draft B includes an Appendix containing proposals for a comparison reference value and degrees of equivalence.

Before the publication of Draft A, it is essential that all participants have an opportunity to review and edit their submitted results and supporting information prior to detailed analysis and publication (between participants only). This is because from this point onwards results and/or uncertainties cannot be changed unless the error is caused by the pilot organisation. The pre-draft A phase is structured into two parts and provides an opportunity for all participants to carry out a peer review of other participants' methodology and associated uncertainty budget prior to being influenced by the visibility of any results on publication of Draft A. In some circumstances it will be advantageous to carry out a peer review phase for the pre-draft A prior to the comparison. This would facilitate a rapid analysis of results in near-real time so that the results can be made available to the participants whilst still at the comparison site. In such circumstances, once published, Draft A will remain final and no results will be allowed to change. However, in the event that a participant has a result that they consider unrepresentative of their expected measurement capability, they may be able to resolve any error and arrange and take part in a subsequent bilateral without leaving the comparison site. In this latter case the first comparison will stand and be published, the second bilateral considered a separate follow-on comparison with a separate analysis and report. Any participant would be expected to explain fully any changes made to their method or instrumentation.

The review body is responsible for ensuring that the comparison meets all the requirements set out in this key guideline, particularly if it is to be used as evidence to support uncertainty claims or traceability.

During the comparison, as the results are received by the pilot institute, they are kept confidential by the pilot institute until all the participants have completed their measurements and all the results have been received, or until the date limit for receipt of results has passed.

The following sections provide more detail in the procedure to be followed.

8.9.1 Pre-Draft A – Process 1: Distribution of Uncertainty Budgets

The uncertainty budget (table of uncertainty components and uncertainty contributions, as well as description of measurement technique and facility) must be submitted by participants to the pilot, together with their results. The overall uncertainty values alone will not be sufficient. If the uncertainty budget received is not complete, the pilot organisation contacts the participating organisation to provide a complete uncertainty budget.

- Specific instructions on reporting the uncertainty budget must have been given in the protocol of the comparison that was agreed before the start of comparison.
- If a participant fails to provide their uncertainty budget in the required detail within a given deadline, the pilot organisation may request that the approval authority remove the participant from the comparison or from the calculation of the comparisons' reference value, as appropriate (in this case, the fact will be stated in the final report).
- The technical description and uncertainty budget (minus the components due to the comparison itself) can in certain circumstances be requested prior to the actual comparison exercise.

After all the descriptions with uncertainty budgets from all participants have been submitted they are collated and distributed as a set to all participants by the pilot. This allows for a peer review prior to any knowledge of results and is done within an agreed period from receipt of all information. In some cases (see above) this process may occur prior to the comparison and the submission of actual final results.

Any participant, including the pilot, can send questions or comments on other participants' uncertainty budgets or methodologies and ask for further information. For example, a participant's uncertainty may be considered unusually small (or large) or some important uncertainty components may be missing.

Comments or questions from any participant are accepted within six weeks from distribution of the uncertainty budgets.

- Comments or questions should be sent to the pilot who will then forward them anonymously to the participant being asked with all other participants in copy. The pilot laboratory is responsible for maintaining records of all communication.

Participants who receive comments must respond promptly and, if necessary, can revise their uncertainty budget, although revision is in no way obligatory. At this stage, any participant can submit a correction of their uncertainty budget, even without receiving comments. However, revision of uncertainty components is only allowed in the direction to increase the overall uncertainty.

Responses to comments and revisions of uncertainty budgets (if any), are accepted within an agreed period from distribution of the uncertainty budgets.

Replies to comments should go to the pilot laboratory and be forwarded by the pilot to all.

If any correction or changes of the uncertainty budget are submitted at this stage, the changes of values and the reason will be reported in the appendix of the comparison report.

8.9.2 Pre-Draft A – Process 2: Review of Relative Data

After the results have been submitted from all the participants and any measurements or checks of the pilot have been completed (within an agreed period, in some cases this may be in near-real time), the pilot sends to each participant:

1. their reported values, as received by the pilot, for verification, and
2. their results reduced in such a way that only the internal consistency of any measurements made by the participant are visible (such data is called Relative Data).
    - Relative Data can be obtained by calculating the ratios of values of all transfer standards measured by the participant and by the pilot laboratory (or other stable reference) and normalising the ratios to their mean. This normalisation removes any relationship of the participant's absolute scale to the pilot laboratory and leaves only internal consistency information. See Annex A for an example.
    - The pilot laboratory sends, to each participant, their Relative Data only, i.e. not other laboratories' data. Relative Data are kept confidential between participants.

The participants review the Relative Data as well as their reported values (as retyped by the pilot) and examine if there are any errors. If any errors are found, the participant can correct their results at this stage. The participants can also examine the stability of any reference standards that may have been used. If significant changes or drifts in any of the reference standards are identified that can be attributed to uncontrollable external effects, e.g., weather, the participant can discuss the removal of the data of the particular reference standards or measurements (or even the re-measurement of the reference standard if necessary).

Each participant must respond to the pilot laboratory within an agreed timescale to confirm that there is no problem with their data or to request any corrections (this may only be hours in some circumstances). All participants should respond but, if no response is received by the deadline, the original data will stand.

Data of particular transfer standards that exhibited problems can be removed with agreement between the participant and the pilot laboratory.

Re-measurement can be done only when it is absolutely necessary and when it will not delay the schedule of the comparison significantly (the pilot makes the decision, although a participant can request it).

It is the participant's responsibility to identify any anomalous feature of their Relative Data that imply errors. If the pilot laboratory finds obvious anomalous results for any participant that cannot be identified from Relative Data, a warning should be sent to ALL participants (without specific information) so as not to specifically alert or influence any one participant.

If any corrections of data are submitted from participants in this stage, the changes of values and the reason will be reported in the appendix of the report of comparison.

If data of any transfer standards are removed, the fact will be stated in the report of the comparison.

Note: Process 1 and Process 2 can proceed simultaneously.

8.9.3 Identification of Outliers

After the Pre-Draft A process, if obvious outlier(s) are observed in the comparison results and where the CRV value would be significantly skewed, the pilot should discuss with all the participants the removal of such data from the calculation of the CRV before Draft A. This should be done so without disclosing the absolute results. For example, the ratios of deviation from CRV and stated uncertainty (k=2) might be distributed for discussion without the identification of the specific participants whose data is skewing the results.

8.9.4 Preparation and Distribution of Draft A

After the Pre-Draft A processes are complete, the pilot prepares and distributes Draft A to all the participants. Draft A discloses the absolute results of the comparison with identification of all the participating organisations. Draft A should tabulate all the results as well as present them in graphical form as necessary. It is recommended that the pilot also distribute the data of the analyses in a spreadsheet file. Draft A should be distributed within an agreed period after completion of all the measurements of the comparison. In some situations this may be in near real time.

Draft A should be given a unique identifier based on the authorising body and the year and the type of comparison.

The default method for calculating CRV is the weighted mean with "cut-off". Use of other methods can be discussed only when the pilot finds serious problems in using the default method and should be discussed before distribution of Draft A. Other methods may be used with consensus of all the participants and subsequent approval of the approval authority. The use of a "cut-off" is to remove any undue bias to any one participant in establishing a CRV.

The cut-off value for the uncertainty, as a default, is determined as the average of the uncertainty values of those participants that reported uncertainties smaller than or equal to the median of all the participants. For example, if there are 10 participants, the cut-off value will be the average of the 5 smallest values of uncertainty. The use of a cut-off value other than the default, if necessary, should be discussed and agreed by all participants before Draft A is distributed:

- The determination of the weights is based on the participants' reported uncertainties, adjusted by the cut-off and combined with the transfer uncertainty of the comparison (i.e. the reproducibility of measurements at the pilot and also other components associated with differences in measurement conditions between the pilot, its participants, etc.).

When discussing use of other methods, the pilot must be careful not to disclose the results of the comparison, while still providing some data for discussion. For example, it is acceptable to disclose the standard deviation of the results, the average of the stated uncertainties, the Birge ratio, etc. Plots of absolute results with uncertainty bars, even with anonymous participant identification, must not be distributed for discussion because identification might be inferred from such data. The ratios of deviation from CRV and stated uncertainty (k=2) of each laboratory (without identification) can be plotted.

The data analysis should be as simple as possible and the calculation process should be made transparent so that the final results can be reproduced by others, without difficulty, from participants' reported measurement results included in Draft A. The data analysis program and intermediate results should be made available for all participants. The approach used should be agreed by all participants before the publication of Draft A. An example of a commonly used data analysis for an intercomparison is provided in Annex B. The calculation process may be elaborated as necessary for each comparison. Alternative calculation techniques based on the least-square model approach (e.g., [2]) may be used as appropriate.

8.9.5 Review of Draft A by Participants

Each participant carefully reviews all the data presented in Draft A and reports to the pilot if they find any clerical errors made by the pilot or send any other comments.

Comments should be sent within an agreed period from distribution of Draft A. In some circumstances this will take place in near real time.

After Draft A has been distributed, correction of the results (reported values and uncertainty values) due to errors by participating labs, for any reason, cannot be accepted.

Once Draft A has been distributed, the whole or any part of a participant's results cannot be withdrawn even if they are found in error. Under special circumstances it might be allowed if the approving authority accepts the justification.

If a participant has found an error(s) that they made in their measurements or in data analysis that affected the reported results, the fact should be reported to the pilot. The corrections are documented in the appendix of the report. In this case, under the pilot's decision, the participants' results (or part of the results) may be excluded from the CRV calculation, with the fact stated in the report.

A change of the method for calculating CRV, if necessary, should be discussed and agreed in Pre-Draft A stage. However, if such a discussion did not take place and Draft A shows serious problems, it can be changed with consensus of all the participants and subsequent agreement by the approval authority.

Removal of partial results should be discussed in the Pre-Draft A stage and it is not allowed at this (Draft A) stage except when the problem in the reference standard(s) or another external factor was not clearly shown in Pre-Draft A stage. In this case, consensus between all the participants must be sought.

If one or more participants make comments, these comments should be circulated to all participants. If the comments are significant, the pilot laboratory can discuss with the participants whether and how changes are to be made for the next Draft A version. If necessary, further data can be distributed as Supplement to Draft A. When changes are made to address comments, the revised draft will be called Draft A-2 and should be distributed again to all the participants for approval. In this case, the revised draft should be distributed within an agreed period from closing comments. If further comments are made to the revised draft, the process can be repeated (Draft A-3, Draft A-4, etc.). The pilot can consult the approving authority in case of dispute. When all the participants approve the Draft A-x, it will become Draft B.

Draft A is considered as confidential, i.e., for the participants only. The data in Draft A should not be distributed or presented to the general public.

A participating organisation that considers its result unrepresentative of its capability may request a subsequent separate bilateral comparison with the pilot institute or one of the participants. This should take place as soon as possible after the completion of the comparison in progress. The subsequent bilateral comparison is considered as a new and distinct comparison. In some circumstances it may be possible to organise such a comparison directly following the first one and prior to approval of its Draft B. However, this must be treated as independent and any differences obtained by participants in this second comparison must be fully explained in the final reports.

8.9.6 Preparation of Draft B

When all participants have agreed the final version of Draft A, it becomes Draft B. The pilot submits Draft B for approval by the appropriate authority within an agreed period from distribution of Draft A (if no further version of Draft A need to be prepared).

- Draft B includes an Appendix containing proposals for a reference value and degrees of equivalence (unilateral and bilateral) presented in tables.
- Draft B will be reviewed by the approval authority (and no longer by participants). The review may request changes to Draft B from the pilot. If a revision is produced, it is called Draft B-2 (and subsequesntly Draft B-3, Draft B-4, etc., as required) and this would be reviewed again by the approval authority. Participants do not participate in this process unless some major revision is proposed. When Draft B-x is approved it becomes the "Final Report".
- Any versions of Draft B are not considered confidential and may be the subject of a publication, with the exception of the Appendix containing proposals for the reference value and degrees of equivalence.

8.9.7 Publication of Final Report

The final reports of all comparisons should ideally be published (or accessible) through the QA4EO website. All key comparisons and/or those used to support claims of traceability must be made available. If the pilot chooses to do so, the reports can also be published in a formal journal.

8.9.8 Recommended Time Line

The following time line is the maximum that should be needed for a comparison. In some circumstances this can be reduced significantly and of course, as discussed above, may be modified to allow the Pre-Draft A phase to take place prior to the comparison.

Month 0: Pilot receives all the results with uncertainty budgets and finishes all the measurements.

Month 2: Pilot distributes the uncertainty budgets of all the participants to all participants (comments due within 6 weeks).

Month 2: Pilot sends out Relative Data to each participant and their reported values as recorded by the pilot for checking (response due within one month).

Month 3: Responses to Relative Data from all participants due. N.B. if no response is received by pilot by this deadline the original reported values stand.

Month 3.5: Comments on the uncertainty budgets closed.

Month 4: Responses to comments on uncertainty budgets and revision of uncertainty closed.

Month 6: Draft A distributed (approval / comments due within two months).

Month 8: Comments on Draft A due.

Month 10: Draft B submitted to approval authority (approval due within 6 weeks) or Draft A-2 distributed to participants (comments due within one month).

Month 11.5: Draft B approved by approval authority or comments on Draft A-2 sent to pilot (any further revisions requested within one month.)

Month 12: Final Report published.

N.B. * Due date for comments after revision of Draft A or Draft B may be adjusted depending on the degree of changes.

* The progress of each comparison will be monitored by the approval authority and reminders will be sent to pilot if the schedule is significantly delayed from the recommended time line.

8.9.9 References

1. Widlowski, J-L., M. Taberner, B. Pinty, V. Bruniquel-Pinel, M. Disney, R. Fernandes, J.-P. Gastellu-Etchegorry, N. Gobron, A. Kuusk, T. Lavergne, S. Leblanc, P. Lewis, E. Martin, M. Mottus, P. J. R. North, W. Qin, M.Robustelli, N. Rochdi, R.Ruiloba, C.Soler, R.Thompson, W. Verhoef, M. M.Verstraete, and D. Xie, "*The third RAdiation transfer Model Intercomparison (RAMI) exercise: Documenting progress in canopy reflectance models*" Journal of Geophysical Research **112** (2007).

2. Woolliams E., Fox N.P., Cox M & Harris P. http://kcdb.bipm.org/AppendixB/appbresults/ccpr-k1.a/ccpr-k1.a_final_report.pdf.

9 Conclusion

This document has outlined the process that should be followed when initiating, organising and reporting the results of comparisons to underpin QA4EO. This key guideline proposes a detailed set of instructions that **can** be followed for all types of comparison but **should** be followed for formal key comparisons. In addition to this guideline, the reader is directed to the QA4EO website (http://QA4EO.org) for a list of active and past comparisons and example protocols and reports.

Annex A

An example is given here for an intercomparison of spectral responsivity where three detectors (NIST04, 08, 10) were used as transfer standards. The detectors were measured by the pilot laboratory, then by a participant, then by the pilot laboratory again. Figure A1 shows the plots of the absolute ratios of the responsivity values of the three detectors as measured by a participant (Lab-1) and the pilot laboratory (Pilot Lab) before and after transportation. So, there are six points at each wavelength. From this, the Pilot Lab sees an obvious anomaly for detector NIST10 at 900 nm. However, the Pilot Lab does not know yet if it is a numerical error by Lab-1 or some problem caused by the detector. These absolute results, of course, must not be sent to participants before Draft A. Instead, Relative Data is sent to the participant to let them identify the problem.

Figure A2 shows the plots of the Relative Data (as described in 8.9.2) for this example. The six values at each wavelength are normalised in such a way that the average of the six values at each wavelength is always 1. Therefore, the relationship of the scales between Lab-1 and the Pilot Lab is removed at each wavelength. Only the internal consistency of measurements of three transfer standards is presented.

By examining the Relative Data, Lab-1 finds the anomaly at 900 nm, but confirms that all other data are fairly consistent. It can also be seen that all detectors reproduced well before and after transportation. Lab-1 checks their results at 900 nm. If they find any error (numerical or technical) on this point, they can correct this value. If not, they might suspect some problem of detector NIST10 at this wavelength and can request the removal of the data. Or, if the Pilot Lab sees some common problems, it can propose, to all participants, the removal of some detectors at particular wavelengths. The

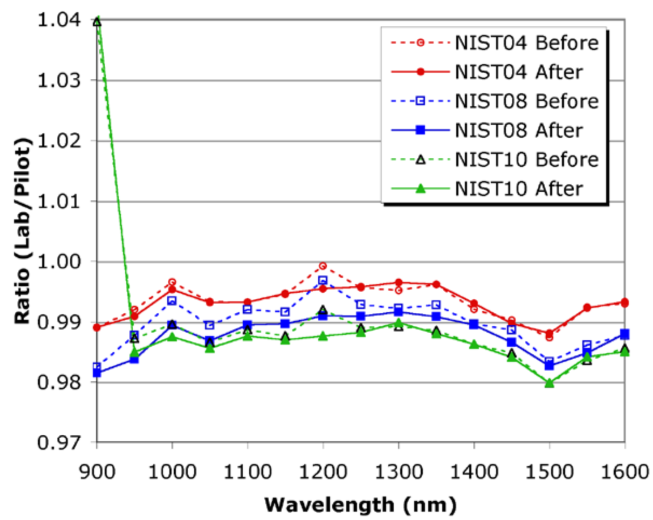

Figure A1: Plots of absolute results in the ratio (Lab-1 / Pilot Lab) participants can then look at their Relative Data to see if it is reasonable or how it may affect their results. Such a request and/or discussion can be done in a fair manner using the Relative Data but without disclosing any absolute results.

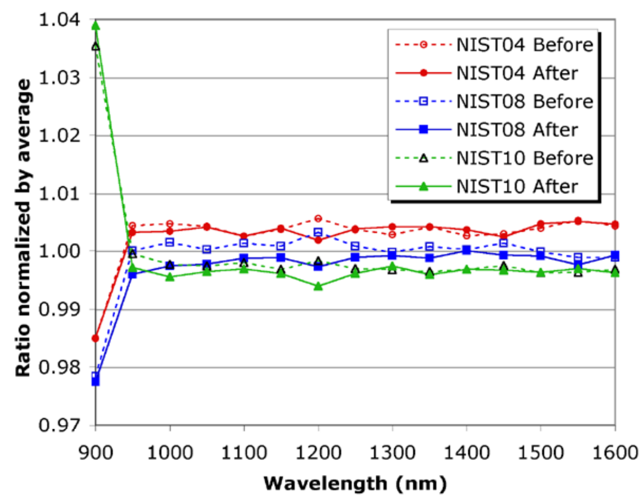

Figure A2: Plots of Relative Data of Lab-1

Annex B

Below is an example of commonly used data analysis for an intercomparison of a spectral quantity. Measurements at each wavelength are taken for each separate comparison. The same analysis will apply to the results at all wavelengths. This example employs three lamps of the same type prepared by each NMI and measured by the NMI, then measured at the pilot laboratory (Pilot Lab), then measured at the NMI again. The two measurements at each NMI (before and after those undertaken by the Pilot Lab) are referred to as round 1 and round 2. The total uncertainty of measurement for each lamp at each round is reported. The total uncertainty and reproducibility of the Pilot Lab measurements for each lamp are reported. In this method, simple arithmetic means are taken in all the intermediate steps for the results from the three lamps and the two rounds within each NMI, then weighted mean with cut-off is applied at the last step as agreed by the CCPR.

The following notations are used:

N	Number of participant NMIs, not counting the Pilot Lab.
$E_{i,j,r}$	Spectral irradiance of lamp j (=1 to 3) of NMI i, measured by the NMI in round r (=1 to 2).
$u_{\mathrm{rel}}(E_{i,j,r})$	Total relative uncertainty of $E_{i,j,r}$ reported by the NMI.
$E_{i,j}^{\mathrm{P}}$	Spectral irradiance of lamp j (=1 to 3) of NMI i, measured by the Pilot.
$u_{\mathrm{rel}}(E_{i,j}^{\mathrm{P}})$	Total relative uncertainty of $E_{i,j}^{\mathrm{P}}$.
$u_{\mathrm{rel}}(E_{i,j}^{\mathrm{PR}})$	Reproducibility of Pilot measurements of lamp j of NMI i, including the stability of the comparison scale during the period of comparison and repeatability of the transfer lamp.

1. For each NMI i for each lamp j, the NMI measurements of two rounds are averaged:

$$\overline{E}_{i,j} = \frac{1}{2}\sum_{r=1}^{2} E_{i,j,r} \tag{1}$$

and its uncertainty by

$$u_{\text{rel}}(\overline{E}_{i,j}) = \frac{1}{2}\sum_{r=1}^{2} u_{\text{rel}}(E_{i,j,r}). \tag{2}$$

Note1: This uncertainty calculation is an approximation, assuming that the results from the two rounds of the same lamp measured by the same NMI are nearly fully correlated. This is normally the case when the uncertainty of transfer measurements (random components) is much smaller than the uncertainty of the scale.

Note2: If the uncertainty of measurements are reported separately for the uncertainty of the scale $u_{\text{rel}}(E_i^{\text{S}})$ of the NMI and the transfer uncertainty $u_{\text{rel}}(E_{i,j,r}^{\text{T}})$ for the particular measurement, the uncertainty of the average of *M* rounds (*M*=2 in example above) is given with correlation taken into account:

$$u_{\text{rel}}(\overline{E}_{i,j}) = \sqrt{u^2{}_{\text{rel}}(\overline{E}_i^{\text{S}}) + \frac{1}{M^2}\sum_{r=1}^{M} u^2{}_{\text{rel}}(E_{i,j,r}^{\text{T}})} \tag{2a}$$

2. For each NMI i for each lamp j, the relative difference $\Delta_{i,j}$ between NMI measurement (as an average of two rounds) and Pilot measurement is given by,

$$\Delta_{i,j} = \frac{\overline{E}_{i,j}}{E_{i,j}^{\text{P}}} - 1 \tag{3}$$

and its uncertainty by

$$u(\Delta_{i,j}) = \sqrt{u_{\text{rel}}{}^2(\overline{E}_{i,j}) + u_{\text{rel}}{}^2(E_{i,j}^{\text{PR}}) + u_{\text{rel,add}}{}^2(E_{i,j})}\,. \tag{4}$$

where $u_{\text{rel,add}}(E_{i,j})$ is an additional uncertainty in the comparison of lamp j of NMI i, arising from those components such as changes of the artifact due to transportation (if identified) and different measurement conditions between Pilot and participants that affected comparison results (if applicable) – often related to characteristics of the artifacts.

Note: The term $u_{\mathrm{rel}}^2(E_{i,j}^{\mathrm{PR}})$ rather than $u_{\mathrm{rel}}^2(E_{i,j}^{\mathrm{P}})$ is used for Pilot lab uncertainty because Pilot measurements $E_{i,j}^{\mathrm{P}}$ are strongly correlated with each other, and only uncorrelated components in Pilot measurements contribute when $\Delta_{i,j}$ are further reduced to calculate DoE.

3. For each NMI *i*, the relative differences Δ_i (average of the three lamps) is obtained by

$$\Delta_i = \frac{1}{3}\sum_{j=1}^{3}\Delta_{i,j} \tag{5}$$

and its uncertainty by

$$u(\Delta_i) = \frac{1}{3}\sum_{j=1}^{3}u(\Delta_{i,j}). \tag{6}$$

Note: This uncertainty calculation is an approximation, assuming that the results from the three lamps measured by the same NMI are nearly fully correlated.

For Pilot lab ($i = 0$ is used hereinafter),

$$\Delta_0 = 0 \quad \text{and} \quad u(\Delta_0) = u_{\mathrm{rel}}(\overline{E}^{\mathrm{P}}) \tag{7}$$

where $u_{\mathrm{rel}}(\overline{E}^{\mathrm{P}})$ is the average total uncertainty of all measurements at Pilot lab:

$$u_{\mathrm{rel}}(\overline{E}^{\mathrm{P}}) = \frac{1}{3N}\sum_{i=1}^{N}\sum_{j=1}^{3}u_{\mathrm{rel}}(E_{i,j}^{\mathrm{P}}) \tag{8}$$

4. The relative uncertainty of measurements of NMI i, averaged for all lamps, is determined by

$$u_{\text{rel}}(\overline{E}_i) = \frac{1}{3}\sum_{l=1}^{3} u_{\text{rel}}(\overline{E}_{i,j}) \tag{9}$$

For convenience of calculation hereinafter,

$$u_{\text{rel}}(\overline{E}_0) = u_{\text{rel}}(\overline{E}^{\text{P}}) \tag{10}$$

5. The KCRV is calculated using weighted mean with cut-off. The cut-off value $u_{\text{cut-off}}$ is calculated by

$$u_{\text{cut-off}} = \text{average}\{u_{\text{rel}}(\overline{E}_i)\} \quad \text{for} \quad u_{\text{rel}}(\overline{E}_i) \le \text{median}\{u_{\text{rel}}(\overline{E}_i)\} \quad ; i = 0 \text{ to } N \tag{11}$$

The reported uncertainty $u_{\text{rel}}(\overline{E}_i)$ of each NMI i is adjusted by the cut-off,

$$\begin{aligned} u_{\text{rel, adj}}(\overline{E}_i) &= u_{\text{rel}}(\overline{E}_i) \quad \text{for} \quad u_{\text{rel}}(\overline{E}_i) \ge u_{\text{cut-off}} \\ u_{\text{rel, adj}}(\overline{E}_i) &= u_{\text{cut-off}} \quad \text{for} \quad u_{\text{rel}}(\overline{E}_i) < u_{\text{cut-off}} \end{aligned} \quad i = 0 \text{ to N} \tag{12}$$

The transfer uncertainty component in $u(\varDelta_i)$ is separated by

$$u_{\text{T}}(\varDelta_i) = \sqrt{u^2(\varDelta_i) - u^2_{\text{rel}}(\overline{E}_i)} \tag{13}$$

The uncertainty of $\varDelta_i$ after cut-off is given by

$$u_{\text{adj}}(\varDelta_i) = \sqrt{u^2_{\text{rel, adj}}(\overline{E}_i) + u^2_{\text{T}}(\varDelta_i)} \tag{14}$$

The weights w_i for NMI i is determined by

$$w_i = u_{\text{adj}}^{-2}(\Delta_i) / \sum_{i=0}^{N} u_{\text{adj}}^{-2}(\Delta_i) \tag{15}$$

The KCRV, Δ_{KCRV}, is determined by

$$\Delta_{\text{KCRV}} = \sum_{i=0}^{N} w_i \Delta_i \tag{16}$$

The uncertainty of the KCRV (weighted mean with cut-off) is given by

$$u(\Delta_{\text{KCRV}}) = \sqrt{\sum_{i=0}^{N} \frac{u^2(\Delta_i)}{u_{\text{adj}}^{4}(\Delta_i)}} \Bigg/ \sum_{i=0}^{N} u_{\text{adj}}^{-2}(\Delta_i) \tag{17}$$

6. The unilateral DoE of NMI i is given by

$$D_i = \Delta_i - \Delta_{\text{KCRV}} \tag{18}$$

$$U_i = k\sqrt{u^2(\Delta_i) + u^2(\Delta_{\text{KCRV}}) - 2\left(\frac{u^2(\Delta_i)}{u_{\text{adj}}^{2}(\Delta_i)} \Bigg/ \sum_{j=0}^{N} u_{\text{adj}}^{-2}(\Delta_j)\right)} \quad ;\ k=2 \tag{19}$$

Note: Eq.(19) takes into account the effect of correlation between Δ_i and Δ_{KCRV}. For any labs that are excluded from KCRV calculation, a simpler form applies:

$$U_i = k\sqrt{u^2(\Delta_i) + u^2(\Delta_{\text{KCRV}})} \quad (20)$$