Customer	**:**	**ESRIN**	**Document Ref**	**:**	**IDEAS+**
Contract No	**:**		**Issue Date**	**:**	**06 May 2016**
WP No	**:**		**Issue**	**:**	**1.2**

Project Baseline for Proba-V Cloud Detection Round Robin (PV-CDRR)

Abstract : This document presents the Project Baseline for the Proba-V Cloud Detection Round Robin Exercise. The document includes the work logic, the approaches and methods, the work plan and the list of deliverables. The project baseline will be discussed within the Proba-V Quality Working Group and when agreed it will be followed along the full duration of the project.

Authors : **Fabrizio Niro, Rosario Q. Iannone, Kerstin Stelzer, Grit Kirches, Carsten Brockmann**

Distribution : **Proba-V Quality Working Group**

© 2014

Table of Contents

© 2014

AMENDMENT POLICY

The Amendment Record Sheet below records the history and issue status of this document.

AMENDMENT RECORD SHEET

ISSUE	DATE	REASON
1.0	16 Feb 2016	Initial version
1.2	6 May 2016	Final version including comments from QWG and taking into account outcome of discussion at QWG#3 Meeting (27-28 Apr 2016)

© 2014

1. REFERENCES AND ACRONYMS

1.1 References

[RD-1] Ackerman, S. A., Holz, R. E., Frey, R., Eloranta, E. W., Maddux, B. C., & McGill, M. (2008). Cloud detection with MODIS. Part II: validation. *Journal of Atmospheric and Oceanic Technology*, *25*(7), 1073-1086.

[RD-2] Ackerman, Steve, et al., Discriminating clear-sky from cloud with MODIS algorithm theoretical basis document (MOD35)." MODIS Cloud Mask Team, Cooperative Institute for Meteorological Satellite Studies, University of Wisconsin. 2010.

[RD-3] Breon, F. M., & Colzy, S. (1999). Cloud detection from the spaceborne POLDER instrument and validation against surface synoptic observations. Journal *of Applied Meteorology, 38*(6), 777-785.

[RD-4] Brockmann C., Paperin M., Danne O., Kirches, G., Bontemps, S., Stelzer, K., Ruescas, A. Cloud Screening and Pixel Characterisation: IdePix Approach and Validation Using PixBox, Sentinel-3 OLCI/SLSTR and MERIS/(A)ATSR workshop, which will be hosted in ESA-ESRIN, Frascati, Italy, from 15 to 19 October 2012.

[RD-5] Christodoulou, C., Michaelides, S. C., & Pattichis, C. S. (2003). Multifeature texture analysis for the classification of clouds in satellite imagery. *Geoscience and Remote Sensing, IEEE Transactions on*, 41(11), 2662-2668.

[RD-6] Hagolle, O., et al. Quality assessment and improvement of temporally composited products of remotely sensed imagery by combination of VEGETATION 1 and 2 images. *Remote Sensing of Environment* 94.2 (2005): 172-186.

[RD-7] Hagolle, O., Huc, M., Pascual, D. V., & Dedieu, G. (2010). A multi-temporal method for cloud detection, applied to FORMOSAT-2, VENμS, LANDSAT and SENTINEL-2 images*. Remote Sensing of Environment,* 114(8), 1747-1755.

[RD-8] Hollstein, A., Fischer, J., Carbajal Henken, C., & Preusker, R. (2014). Bayesian cloud detection for MERIS, AATSR, and their combination*. Atmospheric Measurement Techniques Discussions,* 7(11), 11045-11085.

[RD-9] Jang, J. D., Viau, A. A., Anctil, F., & Bartholomé, E. (2006). Neural network application for cloud detection in SPOT VEGETATION images. *International Journal of Remote Sensing,* 27(4), 719-736.

[RD-10] Lisens, G., P. Kempeneers, F. Fierens, and J. Van Rensbergen. Development of Cloud, Snow, and Shadow Masking Algorithms for VEGETATION Imagery. Proceedings of Geoscience and Remote Sensing Symposium, IGARSS 2000, Honolulu, HI 2: 834–836.

[RD-11] PROBA-V Products User Manual v1.3, E. Wolters, W. Dierckx, E. Swinnen, 31/8/2015

[RD-12] Proba-V QWG#1 Minutes of Meeting and Action Items, ESTEC, 28 – 29 Apr 2015

[RD-13] Proba-V QWG#2 Minutes of Meeting and Action Items, ESRIN, 28 – 29 Oct 2015

[RD-14] Proba-V Symposium Concluding Remarks, Ghent, Belgium, 26 – 28 Jan 2016

[RD-15] PV-LAC: Advanced Land, Aerosol and Coastal products for Proba-V, ESA Statement of Work, PRBV-GSEG-EOPG-SW-15-0001, 15 Feb 2015.

[RD-16] Sedano, F., Kempeneers, P., Strobl, P., Kucera, J., Vogt, P., Seebach, L., & San-Miguel-Ayanz, J. (2011). A cloud mask methodology for high resolution remote sensing data combining information from high and medium resolution optical sensors. ISPRS *Journal of Photogrammetry and Remote Sensing,* 66(5), 588-596.

[RD-17] Wolters, E.L.A., Swinnen, E., I. Benhadj, Dierckx, W., PROBA-V cloud detection evaluation and proposed modification, QWG Technical Note, 17/7/2015

[RD-18] Congalton, R.G.; Green, K. Assessing the Accuracy of Remotely Sensed Data, 2nd ed.; CRC Press: Boca Raton, FL, USA, 2009; p. 193.

[RD-19] Cohen, J. 1960. A coefficient of agreement for nominal scales. Educational and Psychological Measurement. Vol. 20, No. 1, pp. 37–40.

[RD-20] Hayes, A. F., & Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. Communication Methods and Measures, 1, 77-89.

[RD-21] Scott, W. (1955). "Reliability of content analysis: The case of nominal scale coding." Public Opinion Quarterly, 19(3), 321-325.

© 2014

[RD-22] CMS/Météo-France, 2005, Validation report for PGE01-02-03 of SAF/NWC/MSG. . Météo France / Centre de Météorologie Spatiale Report SAF/NWC/IOP/MFL/SCI/VAL/01, version 1.0.

[RD-23] Mackie, S., Embury, O., Old, C., Merchant, C. J., & Francis, P. (2010). Generalized Bayesian cloud detection for satellite imagery. Part 1: Technique and validation for night-time imagery over land and sea. International Journal of Remote Sensing, 31(10), 2573-2594.

[RD-24] Dierckx, Wouter, et al. "PROBA-V mission for global vegetation monitoring: standard products and image quality." International Journal of Remote Sensing 35.7 (2014): 2589-2614.

[RD-25] C. Vancutsem, J.-F. Pekel, P. Bogaert & P. Defourny (2007) Mean Compositing, an alternative strategy for producing temporal syntheses. Concepts and performance assessment for SPOT VEGETATION time series, International Journal of Remote Sensing, 28:22, 5123-5141

[RD-26] Curran, Paul J. "The semivariogram in remote sensing: an introduction." Remote sensing of Environment 24.3 (1988): 493-507.

[RD-27] Román, Miguel O., et al. "The MODIS (Collection V005) BRDF/albedo product: Assessment of spatial representativeness over forested landscapes." Remote Sensing of Environment 113.11 (2009): 2476-2498.

1.2 Acronyms

ATBD	Algorithm Theoretical Basis Document
BELSPO	Belgian Federal Science Policy Office
CALIOP	Cloud-Aerosol Lidar with Orthogonal Polarization
Cal/Val	Calibration/Validation
CCI	Climate Change Initiative
EO	Earth Observation
ESA	European Space Agency
ESRIN	European Space Research Institute
ESTEC	European Space Research and Technology Centre
FAR	False Alarm Rate
GeoTIFF	Geospatial Tagged Image File Format
HDF5	Hierarchical Data Format 5
IDEAS+	Instrument Data quality Evaluation and Analysis Service (extended phase)
IR	Infra-Red
MEP	Mission Exploitation Platform for Proba-V: http://proba-v-mep.esa.int
MODIS	Moderate Resolution Imaging Spectroradiometer
NetCDF	Network Common Data Form
NIR	Near-InfraRed
NN	Neural Network
OAA	OverAll Accuracy
PA	Producer's Accuracy
PDF	Portable Document Format
POD	Probability Of Detection
Proba-V	Project for on-board Autonomy-Vegetation

© 2014

PV-CDRR	Proba-V Cloud Detection Round Robin
QA	Quality Assessment
QC	Quality Control
QWG	Quality Working Group
SEVIRI	Spinning Enhanced Visible & Infrared Imager
SoW	Statement of Work
SNAP	SeNtinel's Application Platform: http://step.esa.int/main/toolboxes/snap/
SPOT-VGT	Satellite Pour l'Observation de la Terre - Vegetation
SPPA	Sensor Performance and Product Algorithm: https://earth.esa.int/web/sppa/home
SYNOP	Synoptic observation
SWIR	Short-Wave InfraRed
TIR	Thermal Infra-Red
TN	Technical Note
TOA	Top Of Atmosphere
TOC	Top Of Canopy
UA	User's Accuracy
VITO	Flemish institute for technological research
VNIR	Visible and Near-InfraRed

1.3 Definitions

The following definitions apply to the current document.

Term	Definition
Validation Dataset	The *Validation Dataset* is the "truth" data, which we are going to use to assess the quality of the different cloud detection algorithms. This dataset will be built from visually interpreted reference images and (optionally) using ground observation from synoptic weather stations. This dataset needs to be global and representative of different environmental conditions (clouds and surface types, different seasons). The validation dataset will be kept in a "vault" and used only at the end of the Round Robin exercise to perform the final quality assessment.
Test Dataset	The *Test Dataset* is a representative sample of the validation dataset, which is provided to the Round Robin participants as an indicator of our pixel classification criteria and definition. This dataset is a subset of the validation dataset, in which all the relevant pixel classes are adequately represented.
Training Dataset	The *Training Dataset* is a statistically significant ensemble of pixels, which is used by the algorithm's providers to train and "calibrate" their methods. Ideally the training dataset must be much bigger than the validation dataset in order to include a sufficient ensemble of cases for training its prediction capability. The selection of a suitable training dataset will be responsibility of the algorithm's providers.

© 2014

2. INTRODUCTION

2.1 Background

Land remote sensing observations in the visible and infrared domain are limited by the presence of clouds. Cloud detection and removal is therefore the first and one of the most critical pre-processing steps in any algorithm for the retrieval of surface properties from satellite observations. Despite the vast literature on the subject, uncertainties still remain on the used cloud detection approaches that can have significant impact for a wide range of land applications. The proposed algorithms in fact lack for generality, being targeted for the considered sensor and application; moreover algorithm validation is challenging due to the scarcity of fully reliable and coincident truth data to compare against. Finally, cloud detection in some environmental conditions and specifically over land can be extremely difficult due to the potential similarity of radiometric and spectral response of the clouds with the underlying surface as well as to the spatial heterogeneity of the landscape.

As for any other optical sensor, also for Proba-V, clouds screening represents a critical pre-processing step and undetected clouds are a major concern for a number of land and coastal water remote sensing applications. The current cloud detection algorithm for Proba-V is inherited from SPOT-VGT [RD-10] and it is based on adjusted thresholds on the reflectances in the Blue and SWIR channels. The limitation of such thresholds approaches is that the cloud detection performances depends largely on the amount of contrast in radiometry and in spectral response between the clouds and the underlying surfaces as well as on the accuracy of the radiometric calibration. This limitation is even more stringent for Proba-V, giving the limited number of spectral bands and the lack of TIR channels or dedicated band for cirrus cloud detection. The determination of a threshold, which is globally valid, is therefore extremely difficult. The problem is hence reduced to a trade-off between over-detection, to the detriment of clear pixels' availability, or under-detection with potential residual noise in the land products. The drivers of this trade-off are ultimately the users, since different retrieval algorithms and applications may be more or less tolerant to residual clouds contamination.

On the other hand, alternative approaches were proposed in the past years both for SPOT-VGT [RD-6] and Proba-V [RD-17], addressing the issue of defining a global threshold for cloud detection. The rationale of these new approaches is to add a-priori seasonal-dependent information on the surface reflectance, extracted from land cover database or from albedo climatology. This ends up in adopting a "dynamic threshold", being higher over bright surface and lower over dark surface, allowing, for instance, to avoid false detection over desert areas, while being more sensitive to thin clouds over dark surfaces. Similarly, different methodologies, such as statistical, multi-temporal or spatial coherence approaches may offer advantages over the thresholds methods in discriminating different pixel classes when radiometric and spectral properties overlap.

These considerations were the basis for proposing and setting up a Round Robin exercise on cloud detection for Proba-V, the baseline for this inter-comparison exercise is described in details in the present document.

2.2 Motivations

The need for an improved cloud detection method for Proba-V products was raised already during the first Quality Working Group Meeting [RD-12], held in ESTEC on March 2015. This need was additionally underlined during the last Proba-V Symposium [RD-14], held in Ghent on January 2016, as being one of the major issues to be addressed for improving data quality. Several presenters at the Symposium reported under-detection of clouds with the current algorithm with semi-transparent clouds representing a major concern for land cover applications as well as for surface properties retrieval. This need is particularly pressing since three data exploitation activities will be started during 2016 [RD-15] to investigate the potential of Proba-V 100 m for coastal and land cover applications as well as to study advanced methods for joint aerosol-surface properties retrieval. All these three advanced studies, in particular the land cover and the surface-aerosol retrieval ones, are extremely sensitive to residual clouds contamination and the quality of the current cloud flag is not adequate to reach the objectives set out in the Statement of Work [RD-15].

© 2014

In order to answer to this need, while considering the time constraints of the three scientific studies, it was agreed during the last Proba-V QWG Meeting in October 2015 [RD-13] to proceed with two parallel activities with different time frame:

- On a short-term, the new method proposed by VITO, based on the usage of GlobAlbedo climatology [RD-17], after a validation phase, will be implemented in the operational chain, with the target to start the reprocessing of the full mission within first half of 2016, in time for providing an improved set of data to the three exploitation studies.
- On a long term, a Round Robin exercise will be initiated in order to test and inter-compare the performances of various cloud screening algorithms on Proba-V data. This inter-comparison will be open to a wide scientific community with the intent of testing different approaches and assess their performances and suitability for various applications. This project, involving interaction with a wider community, extending beyond the QWG team, and including the definition of a large and globally representative validation dataset, will ultimately have a longer duration, with a target end date on January 2017.

A detailed description of the work rationale and approaches that we intend to adopt for running this long-term Round Robin exercise is the subject of this technical note.

2.3 Objectives

The objectives of the Round Robin exercise are:

- To inter-compare different cloud screening methodologies for Proba-V and learn on advantage and drawbacks of the various techniques for various clouds and surface conditions
- To provide final recommendations to ESA on potential best candidate for implementation in the operational processing chain
- To review and consolidate users' requirements within the Proba-V community on clouds clearing and decide on trade-off between under-detection of clouds and clear pixels' availability
- To collect lessons learnt on cloud detection in the VNIR and SWIR domain for land and coastal water remote sensing and reuse them in the frame of Sentinel-2 and Sentinel-3 cloud detection
- To increase awareness on Proba-V mission by inviting new scientific teams in the Round Robin exercise and organising a final workshop on project results

© 2014

3. APPROACHES AND METHODS

3.1 Rationale and Requirements

3.1.1 Rationale

Cloud screening methodologies for SPOT-VGT and Proba-V suffer from the lack of TIR channels or dedicated cirrus band, thresholds methods are particularly limited by the difficulty in determining a threshold, which is globally valid for the different surface characteristics. In order to compensate from this lack of information, additional constraints can be included in a cloud detection scheme in order to improve its performances, such as the temporal stability of the surface reflectances with respect to the clouds (multi-temporal methods, e.g., [RD-7]), the contextual information on the clouds textural features (spatial coherence methods, e.g., [RD-5]), any a-priori knowledge on the surface characteristics (dynamic thresholds methods, e.g., [RD-6]). Alternatively, all the spectral and radiometric signature of the clouds and their physical properties can be optimally used in a statistically sound approach, using various techniques, in particular Neural Network (e.g., [RD-9]) or Bayesian methods (e.g., [RD-8] and [RD-23]).

The idea of testing various algorithms for Proba-V cloud detection and learn how they can be useful in improving performances of the current method, especially in discriminating pixel classes in case the radiometric properties overlap, is the foundation of this algorithm intercomparison exercise.

We define this exercise a “Round Robin”, namely, benefits and drawbacks of the different methods will be evaluated for the different conditions and recommendations to ESA and BELPSO will be drawn. These recommendations will be discussed within the Proba-V users’ community and the QWG to agree on the most suitable cloud detection approach for the future operational baseline.

3.1.2 Users Requirements

The definition of what is a cloudy pixel and what is a clear pixel is not generally accepted, but it depends on the targeted users’ application, being different algorithms more or less sensitive to undetected clouds. The users requirements will be collected within the Proba-V QWG and as part of the advanced Proba-V scientific studies. These requirements will be essential at the end of the inter-comparison exercise to draw the final conclusions and provide recommendations for implementation. At this final stage (after the Round Robin), a representative number of Proba-V key users will be involved in the discussion and in the preparation of the final report.

3.1.3 Algorithms requirements

The Round Robin exercise will be open to any interested algorithm’s provider, who wants to test and inter-compare its own algorithm for cloud detection. The adopted approach for cloud detection can be of any type, e.g., including for instance multi-temporal approaches.

On the other hand, a performance criterion must be taken into account; therefore the participant should provide documented information on the algorithm computing resources. The overall exercise is in fact, aimed to provide recommendations for an operational processor, namely the algorithms should be accurate, but fast enough to be considered as potential candidate for implementation in the Ground Segment.

3.1.4 Clouds Flags requirements

We will focus on a binary flag (clouds/clear) to ease inter-comparison of the different algorithm, although we will accept also algorithms providing “ambiguous” cases, where ambiguous corresponds to semi-transparent clouds in the visible and NIR spectrum. In this context, semi-transparent means that clouds optical thickness in the visible is such that surface properties and spatial characteristics are still “visible” in the satellite images. This definition is in line with the actual pixel classification approach being implemented for this Round Robin exercise and the test dataset will provide indication of what we consider clear sky, semi-transparent or thick cloud. Examples of semi-transparent clouds in Proba-V RGB images over water and over land are shown in Figure 1.

© 2014

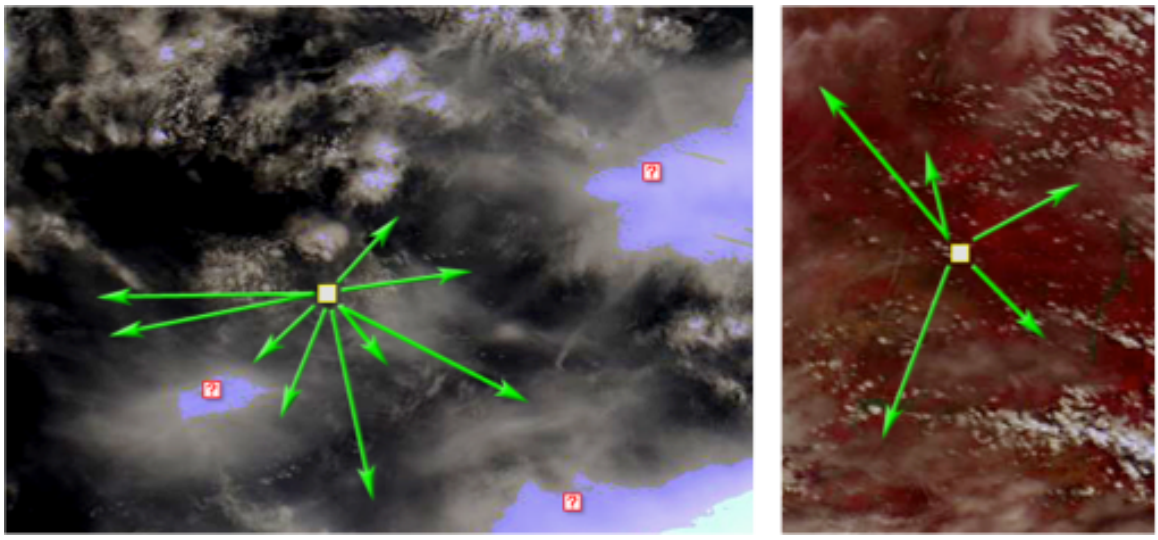


Figure 1 – The arrows indicate example of semi-transparent clouds in Proba-V RGB imagery: on the left over water, on the right over land. In the RGB images thick clouds appear as white, i.e. the surface signal is completely absorbed and only the spectrally flat reflectance of the clouds is visible, for semi-transparent clouds the surface reflectance signal is mixed with the clouds signature.

3.1.5 Round Robin requirements

The Round Robin is by definition an impartial method for discriminating the performance of various and independent algorithms over a common set of input data. The definition of the input data, the test and validation dataset and the associated quality assessment criteria should be made such as not to bias the results of the exercise and to provide fair conditions for all participants.

The input reference scenes should be global and covering all the environmental conditions, providing to the algorithm's providers a sufficiently wide set of input data from which they can extract their own training dataset. We will not provide any training dataset within the Round Robin, in order not to interfere with the algorithm tuning and calibration. Similarly, we will not provide any auxiliary data, which is deemed necessary to constraint a specific algorithm (e.g., land cover maps, surface albedo). For multi-temporal approaches, historical reference scenes from the past days (e.g. 10 days) will be provided on request in order to exploit the multi-temporal information.

Guidelines for the participation to the Round-Robin exercise (we call them *Protocols*) will be provided to the algorithm's providers. The dataset on which the final quality assessment will be made (we call it *Validation Dataset*) should be hidden to the participants, while a small, but statistically representative, subset of this validation dataset (we call it *Test Dataset*) will be made available to the participants in order to provide example quality criteria on how the algorithms' results will be judged and an indication of our pixel classification criteria.

The detailed set-up of the Round Robin exercise is described in the following paragraph.

3.2 Round Robin Set-up

3.2.1 Input reference scenes

The input products that will be the subject of the Round Robin exercise (i.e., the ones to be processed with the various algorithms) are Proba-V Top-Of-Atmosphere (TOA) products. More specifically, *Level 2a* products will be provided to the users, consisting of TOA reflectances in the 4 Proba-V bands, radiometrically and geometrically corrected, projected and resampled to the chosen spatial resolution.

© 2014

We prefer the Level 2a to the standard daily synthesis (S1) users products [RD-11], because the temporal compositing, based on Maximum Value Composite, is a processing step beyond the clouds flagging. The daily temporal compositing is in particular detrimental for those algorithms using the contextual information on clouds spatial structure, since this latter is altered by the mosaicking process. An example of such mixing of pixels coming from different time in the synthesis daily products is provided in Figure 2. The usage of Level 2a is also the preferred option for the validation approach that we are going to adopt, which is based on the PixBox tool [RD-4], as we will further explain in the next paragraphs.

As regards to the resolution, we restraint our Round Robin only to one resolution: 333m.

We require that the algorithms should be able to work at all spatial resolutions, and it would be eventually interesting to study performances at different resolutions in order to investigate how they change in relation to clouds fraction. On the other hand, we will consider this as an optional activity to be performed at a later stage, requiring additional work and time (e.g., the generation and processing of three separate test and validation datasets), while we will focus in this first Round Robin exercise only on 333m products.

The input data to be used for the intercomparison will consist of four days (for the four seasons) global Level 2a products for a to-be-defined year. Note that since these products are not available in the operational chain, we will ask ad-hoc processing at VITO. Note also that during this ad-hoc processing the clouds, the shadows and the snow/ice flagging processing will be disabled in order to provide a blind input dataset to the participants, only information on radiometric quality will be available in the status map.

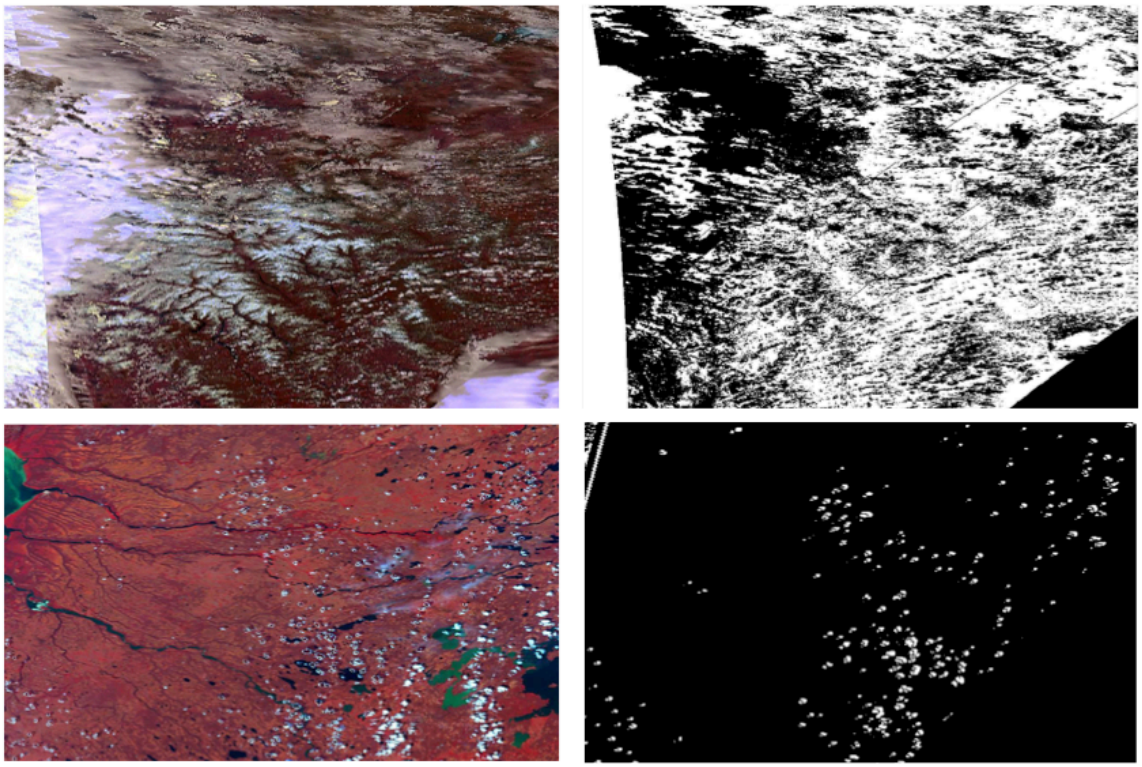


Figure 2 – Examples of the effects of temporal compositing in Proba-V S1, daily synthesis images. On the left two RGB images (NIR-Red-Blue) are shown with artefacts visible in the clouds spatial texture. Artefacts are due to mixture of pixels coming from different overpass time, temporally composed in the daily synthesis. The mixture is evident in the figures on the right showing the two different time of measurement (white and black colours). In the bottom right figure the effect of compositing is confined in the inner part of the clouds, where the MVC method preferably select pixels from different acquisition time.

© 2014

3.2.2 Validation dataset

For validation dataset we intend the ensemble of pixels on which the quality assessment of the various algorithms will be made. The validation dataset will consist on a relatively high number (some ten thousands) of carefully chose pixels, extracted from the input reference scenes (the four days of Level 2a products).

This dataset needs to be statistically representative of the different pixel classes, which are relevant for our scope. The following main pixel classes will be statistically represented in the validation dataset: clear sky, thick clouds, and semi-transparent clouds over different surfaces: land, snow/ice and coastal water. The statistical distribution of the classes should be representative of mean global clouds cover condition, i.e.: 60% of cloudy pixels (half of which semi-transparent clouds), 40% of clear pixels. The distribution of surface type is in-line with the typical observation scenario of the Proba-V sensor, acquiring largely over land (70%), with the remaining 30% equally distributed over coastal, inland water and snow/ice.

In addition of being statistically significant, the validation dataset needs to be global and representative of different seasons, meaning that the selected pixels need to be equally spread over the globe for the four seasons, allowing to correctly represent the different climatological conditions and land cover types. Finally different geometry of observation should be included (sun and viewing geometries) in order to represent different illumination condition and atmospheric path radiance.

The validation dataset will be populated using a dedicated tool, called PixBox [RD-4], which allows the pixel classification by visual inspection of trained experts. The pixel classification is made using a set of visualization tools (e.g., SNAP), to help the visual discrimination of classes. A-posteriori, a Neural Network is applied to identify outliers within the database and raise a flag for potentially misclassified expert pixels. The PixBox tool has been already successfully used in the frame of several ESA CCI projects for the quality assessment of pixel classification algorithms (including cloud screening).

Note that the validation dataset will be hidden to the Round Robin participants, in order to avoid that algorithms are adjusted to the pixels, where the actual quality assessment will be made. In other words, the participant will not know which pixel, among the input reference scenes, will be part of the validation dataset.

3.2.3 Test dataset

The test dataset represents a small subset of the validation dataset. This subset needs to be representative of the different pixel classes and environmental conditions. The test dataset will include metadata information on pixel clouds coverage derived by visual inspection using the PixBox tool, e.g., cloudy, clear sky, semi-transparent clouds.

The purpose of the test dataset is to provide to the participants example of our pixel classification criteria and nomenclature. Since the goal is only to give identification on how the different algorithms will be quality assessed, while we want to avoid the algorithms to be tuned on this dataset, the extension of the test dataset will be limited, but still providing a representative example of all the pixel classes.

An explicative example of our definition of the reference scenes, the validation dataset and the test dataset is provided in Figure 3.

© 2014

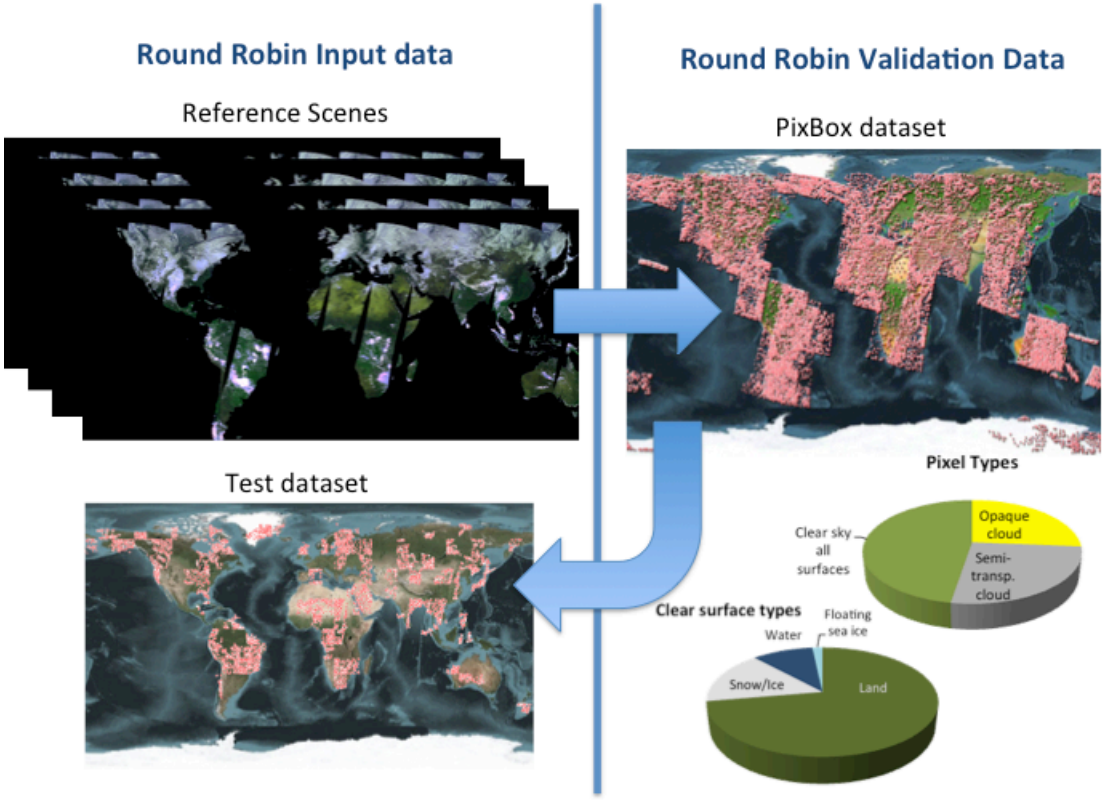


Figure 3 – Example of the different dataset, which will be used for the Round Robin exercise. From top left in the clockwise direction: input reference scenes, validation dataset and associated statistical distribution of clouds/clear and surface type pixels, and test dataset. The figures are only for the sake of example, showing how the datasets will be globally spread and visualizing their relative size and statistical distribution of classes.

3.2.4 Protocols

In addition to the input data and the test dataset, the participants will be provided with a guidelines document that we call *Round Robin Protocol*. Within this document the overall project baseline will be briefly outlined and the requirements for participation to the inter-comparison exercise will be detailed, this includes in particular the description of the input data and the requirements for the output data format and relevant documentation.

Although the Protocols will be issued at a later stage, together with the disclosure of the test dataset, we can already anticipate that we will give freedom to the users to choose among a set of commonly used formats (NetCDF, HDF5, GeoTIFF). Note also that we will require the participant to provide a relevant documentation with the output products, namely an ATBD (but it can be also a peer-reviewed paper), describing the algorithm processing flow and input/output data description.

3.3 Quality Assessment

3.3.1 Potential QA methods

In order to perform a quality assessment (“validation”) of the various algorithms we need first to define what is the truth against which we are going to compare the Round Robin results. It is well known that the “validation” of a cloud detection algorithm is a very challenging task due to the scarcity of generally accepted and fully reliable ground truth data to compare against. Moreover, the capability of an observing system to “see” a cloud depends on the measurement principle, the wavelength and on the observing conditions, in particular on the viewing geometry. Inter-comparison with other remote sensing cloud data is also extremely challenging due to stringent

© 2014

constraints in time and space collocation, unless the considered sensor fly in formation with other sensor, e.g., MODIS on Aqua and CALIOP on Calipso [RD-2].

A generally accepted practice in evaluating the performance of a cloud detection algorithm is to use an expert analyst to interpret the satellite imagery in terms of clear and cloudy regions (e.g., [RD-16]). The human eye is in fact able to recognize on a satellite image cloud structure much better than any automatic algorithm. On the other hand, the human recognition is not an independent "validation" and it may induce mislabeling in case the visual interpretation is more problematic (thin clouds, broken and patchy clouds structure).

Alternative approaches consist in using independent surface observations of cloud cover, such as from synoptic network of meteorological weather stations [RD-3] or from active ground-based instruments [RD-1]. We should note, however, that these ground measurements have their own limitation, mainly because the up scaling of point-measurements to the sensor pixel data will introduce systematic bias. In addition, ground cameras allow only verification of clouds fraction over the total sky area imaged by the camera, and the size of this area depends on the altitude of the clouds, meaning that a coincident active measurement of cloud height should be made to improve the accuracy of the validation. Validation with ground based active sensors (e.g., lidar or radar) is more accurate, but it lacks for global coverage.

3.3.2 Adopted QA approach

Taking into account the challenges in cloud mask validation, as well as the limitation of the potential ground based measurements, we decided to base our QA approach primarily on the PixBox tool and associated validation dataset, which was illustrated in previous paragraph.

The pixel classification performed within PixBox at the time of the database collection will represent our truth data to be used for comparing the various algorithms. PixBox has the advantage of being based on visual inspection of satellite images, which is the most effective approach for clouds identification, and implements, in addition, a set of statistical tools for verifying the representativeness and the correctness of the performed visual classification, an a-posteriori QC based on NN allows in particular to identify potential human errors in the classification. This ensures a statistically robust and globally consistent approach for validation.

Furthermore, visual inspection of single images will be used as complementary information to understand the performances of the various algorithms for different environmental conditions. Visual inspection in fact will always provide contextual information on the clouds structure, which is lost in the PixBox pixel-wise comparison.

Alternative methods for clouds validation are deemed at the moment not sufficiently reliable to be part of the baseline for this project. However, as an optional activity, we will explore the possibility of using collocated surface global observations from weather stations, such as the SYNOP dataset [RD-22], which were successfully used here [RD-23] for cloud detection validation. An example of the geographical distribution of SYNOP surface observations used for validating SEVIRI cloud masking over Europe and North Africa is shown in Figure 5.

Another optional QA method consists in the derivation of Level 3 monthly composite images. This method is particularly useful in order to evaluate the trade-off between cloud detection and clear pixels' availability, in addition, it allows to detect misclassification and under-detection of clouds, which appear as artifacts in the resulting composites. An example of such artifacts in Level 3 temporal compositing for MERIS 300m products is shown in Figure 4. In order to use this QA method, however a full month of Level2a data needs to be processed ad-hoc at VITO, this input data needs to be analyzed by the different algorithm providers and the output masks needs to be post-processed to generate the composites (Mean Compositing technique [RD-25] will be the preferred option for compositing). This entails an additional significant burden in the effort allocated for this project; therefore we consider this method as optional, despite its undeniable interest for the sake of cloud algorithm verification.

© 2014

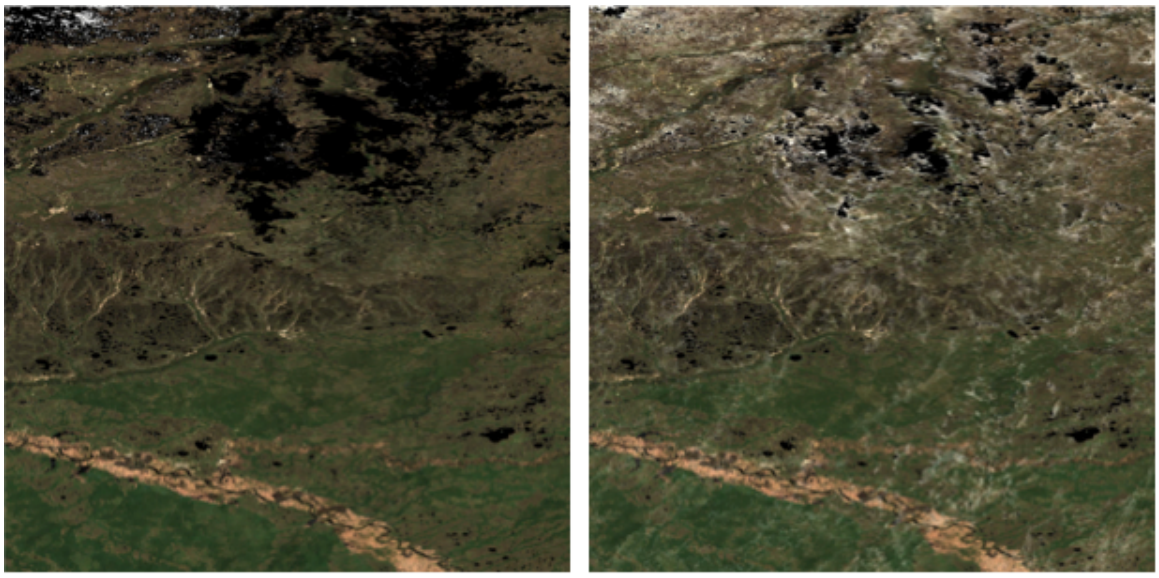


Figure 4 – Effect of different cloud masks on Level 3 products. The two RGB images show 7-days synthesis products from MERIS 300m data obtained using two different algorithms for cloud screening. Residuals undetected semi-transparent clouds are clearly visible in the composite image on the right, this is the result of a less conservative cloud screening algorithm.

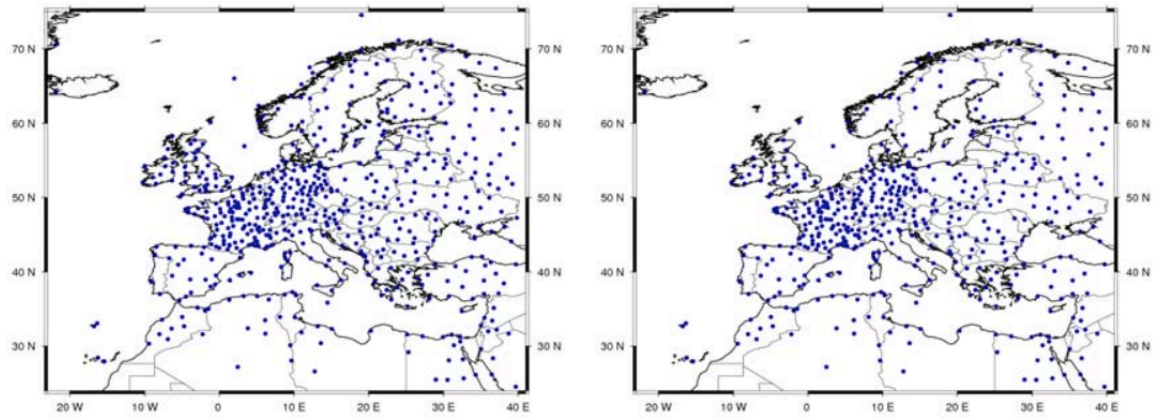


Figure 5 – Geographical distribution of SYNOP weather stations used for validation of clouds masking of MSG-2/SEVIRI geostationary satellite, figure extracted from [RD-22]. On the left the total number of stations is shown, on the right only the ones, which were actually used for the statistics. The SYNOP data used are routine weather observations, coded by the observers into WMO synoptic code they contain information on cloud cover for low, medium and high clouds.

3.3.3 Adopted Quality Metrics

As explained in the previous paragraph, our QA method will be largely based on the PixBox tool and its associated performance metrics. Once the PixBox database is populated, in fact, this tool ensures a comprehensive set of statistical quality metrics in order to inter-compare the performances of the various algorithms. An example of a typical output of the PixBox tool is provided in Figure 6.

The quality metrics provided as output of PixBox tool include the following statistical indicators:

- *Confusion Matrix* [RD-18] – The confusion matrix is calculated by comparing the location and class of each ground truth pixel (in our case the PixBox classification) with the corresponding location and class in the classification image (in our case the considered

© 2014

cloud detection algorithm). The ground truth will consist of cloudy and clear classes, with an additional semi-transparent clouds class, for those algorithms providing also "ambiguous" cases. For binary classification algorithms, the impact of semi-transparent clouds will be investigated by studying the variation of QA metrics when considering only cloudy/clear classes and when including semi-transparent clouds in the cloudy class.

- *Producer's Accuracy, PA* [RD-18] – The producer's accuracy (%) is a measure indicating the probability that the classifier has labeled an image pixel into Class A given that the ground truth is Class A. In our case e.g., the percentage ratio between how many pixels were correctly flagged as cloud by the considered algorithm with respect to how many pixels are classified as cloud in the PixBox reference database.
- *Omission Error* – Errors of omission represent pixels that belong to the ground truth class but the classification technique has failed to classify them into the proper class, i.e., *OE (%) = 100 - PA*
- *User's Accuracy, UA* [RD-18] – The user's accuracy (%) is a measure indicating the probability that a pixel is Class A given that the classifier has labeled the pixel into Class A. In our case e.g., the percentage ratio between how many pixels were flagged correctly as cloud by the algorithm with respect to the total number of pixels which were identified as clouds by the algorithm.
- *Commission Error* – Errors of commission represent pixels that belong to another class that are labeled as belonging to the class of interest, i.e., *CE (%) = 100 – UA*
- *Overall Accuracy, OAA* [RD-18] – The overall accuracy (%) is the percentage ratio of the number of pixels classified correctly with respect to the total number of pixels.
- *Cohen's Kappa coefficient* [RD-19] - The kappa (κ) coefficient measures the agreement between classification and ground truth pixels. A kappa value of 1 represents perfect agreement while a value of 0 represents no agreement. The Cohen's kappa coefficient is generally considered a more robust measure than simple percent agreement calculation, since it takes into account the agreement occurring by chance.
- *Scott's pi* [RD-21] – Scott's pi is similar to Cohen's kappa coefficient in that it considers also the agreement that might be expected by chance. However, the expected agreement is calculated slightly differently.
- *Krippendorf Alpha* [RD-20] - Krippendorff's alpha is more general than the other coefficients listed in here. It adjusts to varying sample sizes and affords comparisons across a wide variety of reliability data.

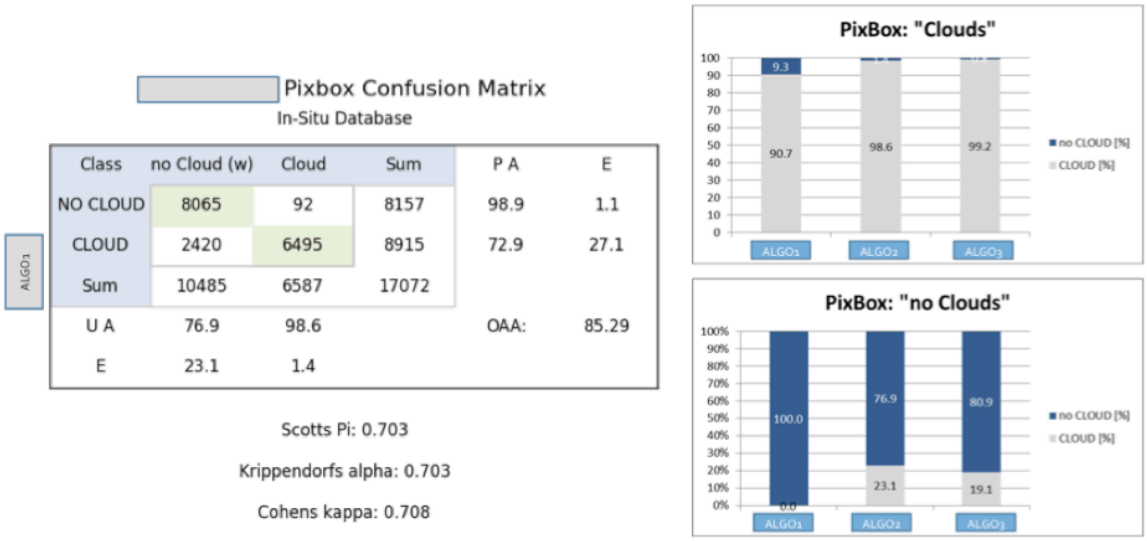
In case of the (optional) validation using a dataset of synoptic observations, the QA will be performed in line with [RD-22], using standard quality indices such as:

- *Probability of Detection, POD* - is the rate of correctly detected cloud observations, i.e. targets classified as cloudy and observed cloudy.
- *False Alarm Rate, FAR* – Is the rate of missed clear observations or false flagging of clouds, i.e. the targets classified as cloudy but observed clear (it expresses cloud over-detection errors)

In case of the (optional) validation using Level 3 monthly composite, the QA will be based on:

- Visual inspection of the derived composites for detecting artifacts due to undetected clouds
- Quantitative assessment of spatial coherence in the generated composites, using as measure the semivariogram [RD-26] and the Relative Coefficient of Variation [RD-27].
- Estimation of cloud coverage to study trade-off between cloud detection performances and clear pixels availability

© 2014



Pixbox Confusion Matrix

In-Situ Database

Class	no Cloud (w)	Cloud	Sum	P A	E
NO CLOUD	8065	92	8157	98.9	1.1
CLOUD	2420	6495	8915	72.9	27.1
Sum	10485	6587	17072		
U A	76.9	98.6		OAA:	85.29
E	23.1	1.4			

ALGO1

Scotts Pi: 0.703

Krippendorfs alpha: 0.703

Cohens kappa: 0.708

Figure 6 - Example of PixBox quality assessment metrics. On the left panel a Confusion Matrix, where PixBox clouds/no clouds classes (the reference "truth") are compared against the results of one algorithm under validation. The associated quality metrics are also shown, including User's Accuracy (UA), Producer's Accuracy (PA), Overall Accuracy (OAA) and relevant statistical indices (Scott Pi, Krippendorf's Alpha, Cohen's Kappa). On the right, the cloud detection performances of three algorithms are inter-compared for PixBox clouds and no-clouds classes.

3.3.4 Known Critical issues

Cloud detection and removal from Proba-V data is very challenging as already discussed within this document, in particular over bright surfaces (desert, snow/ice) and more generally when the radiometric and spectral response of the clouds and the underlying surfaces overlap, such as for high altitude ice clouds over snow/ice surfaces. These are the basic challenges to which all algorithms will face during the Round Robin exercise and which are common to all optical sensors working in these wavelength regions. There are however, other specific issues, which are more specific to the Proba-V instrument and which may be relevant at the time of the quality assessment of the different algorithms. These quality issues will be briefly recalled within this paragraph, they will be relevant in order to further assess the quality of the various approaches and their robustness to critical cases.

3.3.4.1 Channels Saturation

Proba-V dynamic range was carefully adjusted in order to cover the range of natural variability of radiances for different land cover types; integration time is further optimized, being dependent on illumination conditions (latitude and season), in order to dynamically change compression settings [RD-24]. On the other hand, channel saturation may happen over high altitude ice clouds and deep convective clouds, for which the integration settings are not optimally designed. This channel saturation is particularly visible in the Blue, but also in the Red channel, as illustrated in an example case in Figure 7. The PixBox validation dataset will be specifically selected in order to avoid such saturated pixels. Therefore these saturated pixels will not be part of the ensemble of points where the QA will be made.

3.3.4.2 Time difference in Proba-V bands

A Proba-V specific feature, due to instrument's optical design, is that the different bands have slightly different time of observation. The temporal separation between bands is up to 12 s (between NIR and SWIR), and the location of clouds in the SWIR image can be different from the location of clouds in the NIR and Blue bands, because of cloud movement [RD-11].

We will not require the participants to specifically address this particular feature of Proba-V instrument, e.g. by providing different cloud masks for VNIR and SWIR cameras. However this issue should be taken into account while performing our quality assessment, specifically at the

© 2014

edge of clouds, where temporal shift may cause some discrepancies between algorithms exploiting information from different spectral bands.

3.3.4.3 Stripes in SWIR detector

Bright stripes in SWIR camera are sometimes observed in Proba-V RGB (SWIR-NIR-Red) images. The stripes are caused by saturated SWIR pixels, for instance in case of sudden increase of dark current in the relevant pixel. These bright stripes can cause cloud detection algorithm to erroneously identify clouds.

The problem is known and clearly discernible by visual interpretation of the images, therefore these regions will be avoided in the PixBox selection of the validation dataset, so that they will not impact our QA metrics.

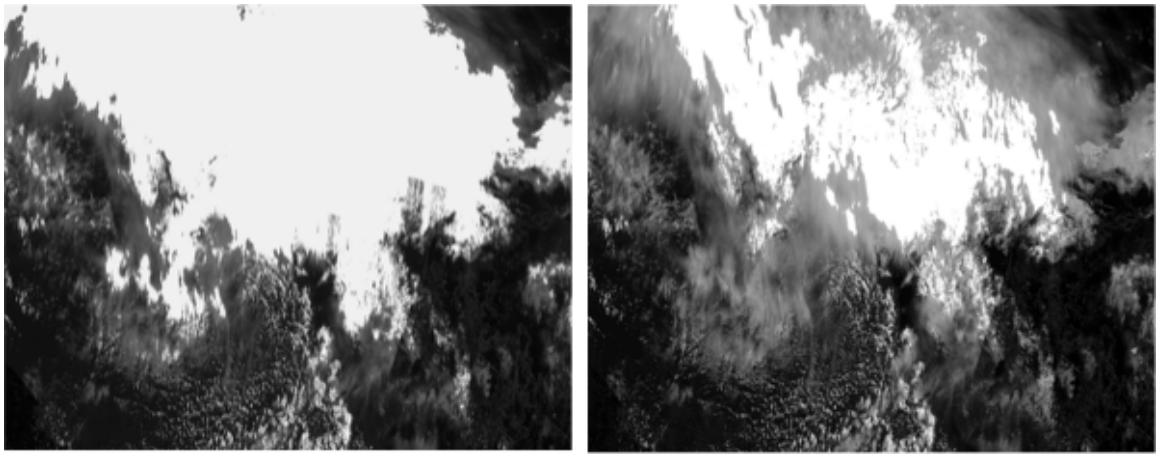


Figure 7 – Reflectance in Blue and Red Proba-V channels, showing in white the saturation of the signal over bright thick clouds.

© 2014

3.4 Summary

A summary table providing the main requirements and settings for the Round Robin exercise is provided here after for the sake of clearness.

	Requirement and Settings
General Requirements	• Provision of input reference scenes and test dataset • Validation dataset hidden to the participants • No provision of training dataset • No provision of auxiliary data
Algorithm Requirements	• Any type of algorithm and methods is accepted • Accepted flags: cloudy, clear, semi-transparent
Input Reference Scenes	• 4 days (4 seasons) global dataset • Level 2a Proba-V products • Spatial resolution: 333m
Validation Dataset	• Statistically significant (more than 20000 pixels) database • Visually classified and cross-checked with PixBox tool • 30% thick clouds, 30% semi-transparent clouds, 40% clear • From clear pixels: 70% land, 15% snow/ice, 15% water • Globally spread covering the four seasons
Test Dataset	• Randomly extracted from Validation Dataset • Representative of all classes and conditions • Pixel classification visible to the participants
Quality Assessment and QA Metrics	• Use PixBox classification as reference ground truth with associated quality metrics: confusion matrices and QA indices (UA, PA, Scott's pi, Cohen's Kappa and Krippendorf alpha coefficients) • Use visual comparison of selected images to investigate performances in clouds structure delineation and for assessing critical cases (thin, patchy, or cirrus clouds) • (Optional). Generate Level 3 monthly composites to investigate impact of undetected clouds on synthesis images (e.g., spatial coherence, residuals clouds contamination) • (Optional). Explore the use of clouds cover information extracted from synoptic meteorological data using standard quality metrics (POD, FAR)

© 2014

4. ORGANIZATION AND PLANNING

4.1 Involved teams

The Round Robin project will be carried out in the frame of IDEAS+ ESA contract. Within this contract, Serco and Brockmann Consult are providing support to the ESA/ESRIN SPPA section in a number of tasks, ranging from Data Quality to Cal/Val and algorithm evolution.

Serco will coordinate the project with the scientific support of Brockmann Consult for the validation and test dataset definition as well as for the quality assessment. Serco will also manage the external scientific partners via an open call, providing technical support for the data processing and output delivery and it will liaise with ESA/BELSPO and the QWG for providing visibility on the project's plan, schedule and deliverables. The QWG will be in particular the representative of the Users' requirements, providing recommendations for the project baseline definition and for the clouds screening tolerance level for the various applications. Serco and Brockmann Consult will collaborate for the project reporting toward ESA and the QWG and for the outreach activities (e.g. presentation of project outcomes to meeting and Conferences).

VITO is expected to provide support to the projects, in particular for the generation of the reference scenes up to the required Level 2a products, since these products are currently not systematically available to the users.

The ESA RSS team will provide the Cloud Toolbox facility on which the Round Robin exercise will be made. This consists of a remote virtual machine including all input data, a set of pre-installed tools and a customized SW environment where to run the algorithms.

External teams will be invited with the open call to participate to the Round Robin, the idea is to have a wide range of methodologies to inter-compare, and therefore participants from outside the Proba-V community will be warmly encouraged. Serco will set up a web site, as part of the ESA SPPA web pages in order to provide visibility to the overall project.

Note that a dedicated fixed budget is available for each external team to participate to the inter-comparison exercise.

The agreed project organization is depicted here below.

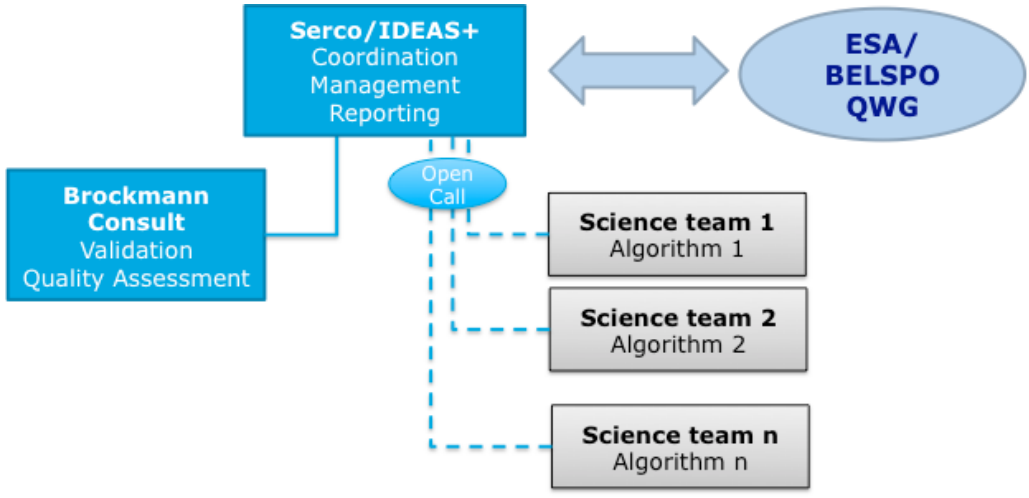


Figure 8 – Working team for the Proba-V Cloud Detection Round Robin exercise and interaction within the different partners.

4.2 Facilities

The following facilities will be used in the frame of this project:

© 2014

- The PixBox tool developed by Brockmann Consult in the frame of past ESA CCI project will be the principal tool for the validation dataset and test dataset generation as well as for the final quality assessment.
- The ESA RSS Cloud Toolbox facility to host the Round Robin development and processing environment
- The ESA SPPA web site will be used for outreach and promotional activities and for presenting the projects' outcomes and plan.
- The VITO Ground Segment will be used for generating the input reference scenes up to Level 2a products level.
- Internal ESRIN room and facilities will be used for organising the final Workshop.

4.3 Task Description

The main tasks identified for carrying out this project are briefly described here.

4.3.1 Task 1: Project Baseline

The first task is to propose and agree the baseline for the Round Robin exercise; the main deliverable of this task consists in the present document. The baseline will be iterated within ESA, BELSPO and the Proba-V QWG. This task is expected to last 2 months since the start of the project (1st Jan 2016) and it is a crucial step, since when the baseline will be consolidated it will be followed upon during the whole duration of the project (13 months). The project baseline should be the guideline for the setting up the Round Robin exercise, for the Quality Assessment approach as well as for the planning and schedule. The project baseline will be prepared as a joint effort between Serco and Brockmann Consult.

4.3.2 Task 2: Community Involvement

The reaction of the users community to the open call is expected to be good enough to start a "meaningful" inter-comparison exercise, i.e., different and independent approaches (at least 3-4) need to be inter-compared in order to draw final conclusion and learn advantages and drawbacks of various methodologies. In order for the project to be successful, therefore, it is crucial to ensure that the optical community is sufficiently aware of this opportunity and that the interested participants are progressively involved within the project. This task will include promotional activity, such as e-mails, dedicated announcement in ESA SPPA web site, as well as direct contact and solicitation to a number of potential interested participants. Some of them are already expected to contribute to the project, although we need to enlarge participation to cover a broad range of methodologies.

4.3.3 Task 3: Test Dataset and Input data

While the validation dataset is being populated, a statistically representative subset of this dataset will be made available to the participants at the time of starting the Round Robin. The other input for starting the Round Robin will be also provided, this includes the input reference scenes (4 days of L2a products generated at VITO) and the Round Robin Protocols provided by Serco and Brockmann Consult.

4.3.4 Task 3: Validation Dataset

As soon as the Project Baseline is consolidated, the generation of the validation dataset will be started at Brockmann Consult. VITO support will be required to process the reference scenes. As final output of this task the validation dataset will be made available together with a short Technical Note on the content of the dataset (e.g., statistical and geographical distribution of pixel classes). This activity is expected to last 3 months.

4.3.5 Task 5: Round Robin

The Round Robin exercise will be open and coordinated by Serco, this includes the support to data processing and the preparation of a dedicated web site, to be hosted on the ESA SPPA web pages, where all the information on the project will be stored, including documentation, planning and announcement of the open call. The RSS Cloud Toolbox facility will be used for the delivery

© 2014

of the reference scenes and test dataset together with additional tools (e.g., SNAP). In particular, a virtual machine with associated input data and processing resources will be distributed to the participants in order to develop, train and test their algorithms. ESA RSS team will ensure the availability of the Cloud Toolbox for this purpose. The participants are expected to deliver the output data following the rules detailed in the protocols. An ATBD should be provided for each algorithm, describing high level algorithm processing model and the input and output data. The Round Robin exercise is expected to last 4 months.

4.3.6 Task 6: QA and Recommendations

The final task will consist in collecting all the Round Robin output and performing the Quality Assessment using the PixBox statistical tools, the visual inspection, as well as the use of the temporal compositing for testing the impact of undetected clouds on daily composite. The task will be concluded with the preparation of a report to be iterated and agreed within the Round Robin participants and the QWG. Outreach activities are also part of this task, this includes in particular the organization of a Workshop (in ESRIN), where all the results will be presented and discussed. This final task is expected to last 3 months with a target date for the workshop at the end of January 2017. As a final deliverable from this task a document of recommendations will be provided to ESA/BELSPO in order to advice on the potential best algorithm for cloud detection of Proba-V data. Key Proba-V Users will be involved at this stage to provide suitability of different methods for various applications.

4.4 Schedule

A provisional schedule of the project is provided here below, internal progress meetings are not shown. Note that the dates for Proba-V QWG Meetings will be synchronized to this schedule whenever possible, for instance, QWG Meeting #4 is scheduled at the end of the Round Robin phase: 24 – 25 November 2016.

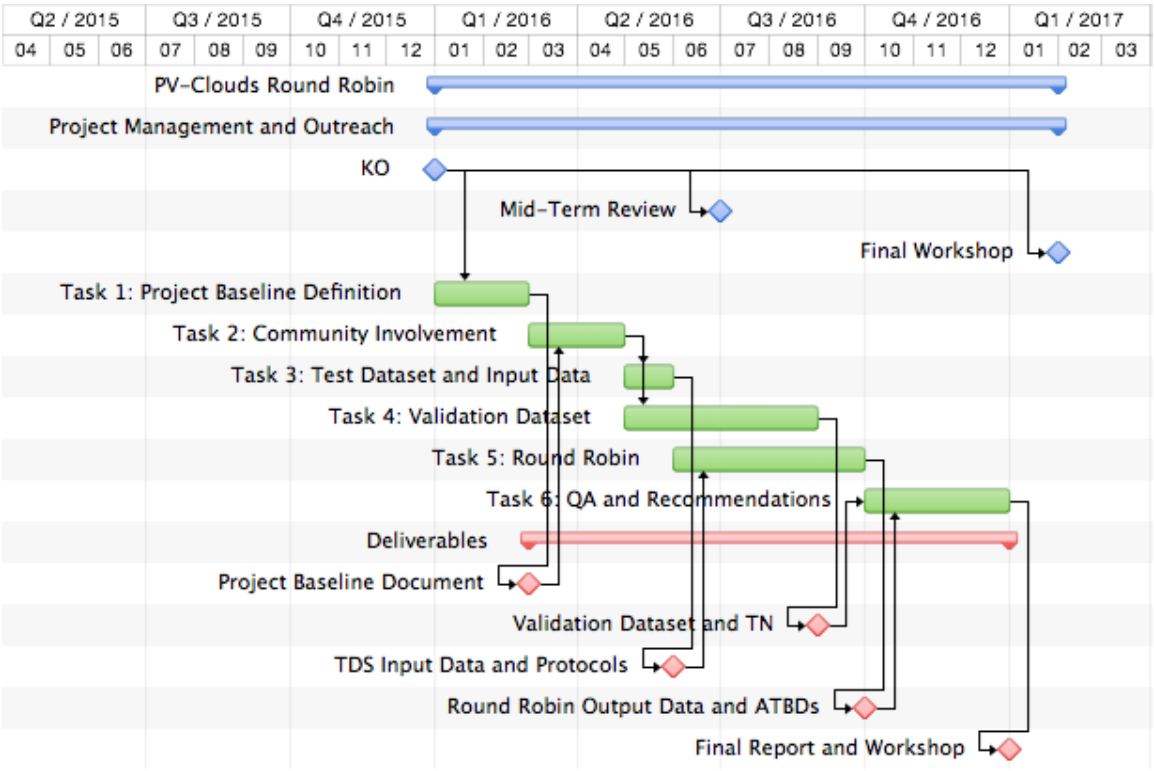


Figure 9 – Schedule and deliverable plan for the Proba-V Cloud Detection Round Robin exercise.

© 2014

4.5 Deliverables

The list of deliverables to be provided is presented in the following table; the current document is the actual 1st deliverable of the project.

ID	Title	Task	Delivery date
D-1	Project Baseline	Task 1	01/03/2016
D-2	List of Round Robin participants	Task 2	13/05/2016
D-5	Round Robin Protocols	Task 4	15/05/2016
D-3a	Test Dataset	Task 4	01/06/2016
D-3b	TN on Test Dataset	Task 4	01/06/2016
D-4a	Validation Dataset	Task 3	01/09/2016
D-4b	TN on Validation Dataset	Task 3	01/09/2016
D-6	Round Robin Output data	Task 5	01/11/2016
D-7	ATBDs	Task 5	01/11/2016
D-8	Final Report and Recommendations	Task 6	01/02/2017
D-9	Workshop	Task 6	01/02/2017

© 2014

End of Document

© 2014