SME-LET Announcement of Opportunities 2009: Cal/Val and User Services -Calvados

Requirements Baseline
Version 1.2.1
16. July 2010

N. Fomferra, C. Brockmann

Index

1 General

1.1 Changelog

Version	Date	Change	Authors
0.5	30.04.2010	The initial draft version of this document.	NF, CB
1.0	12.05.2010	Changes after review 1 by E. Kwiatkowska 1. Further clarified **R-SOW-2.4** in that 2 years of MERIS RR data is actually not sufficient 2. Removed doubled and incomplete **R-FN-3.8, R-FN-3.9, FN-2.7** 3. Extended **R-FN-3.8** (binning) by flag masking requirements 4. Changed priority of **R-FN-3.13** to 'expected' 5. Fixed wrong references [RD-7] and [RD-4] in **R-FN-4.2**	NF
1.1	03.06.2010	Changes after RB review during progress meeting 1: 1. Further detailed the L2 processors to be realised: **R-FN-3.3** 2. Revised section 5.2.4 "Analysis requirements". Detailed the tests to be implemented. Created new and updated requirements **R-FN-4.1** to **R-FN-4.7** 3. Added new requirements section 5.3.3 "Demonstration to the Community" with **R-NF-3.1** and **R-NF-3.2**. 4. Added a number of new references specifying the algorithms to be used for the analysis tests.	NF
1.2	30.06.2010	Changes after 1.1 review by E. Kwiatkowska: **1.** Further clarified input data to be used in **R-SOW-2.4** **2.** Clarified user management in **R-FN-1.2** 3. Changed priority of **R-FN-2.3** and **R-FN-2.4, R-FN-2.5** from 'desired' to 'expected' 4. Extended **R-FN-3.1** by L3 processing 5. Completely revised **R-FN-3.2** 6. Changed priority of **R-FN-3.4** to 'essential' (processing parameters) and further clarified possible realisations. 7. Changed priority of **R-FN-3.4** to 'essential' (processing parameters) and further clarified possible realisations. 8. Changed priority of **R-FN-2.8** and **R-FN-3.9** from 'desired' to 'expected' (bulk processing) **9.** Fixed Test 3 description in **R-FN-4.3** and **R-FN-4.6**	NF

		10. Fixed priority of **R-FN-2.2** by changing it from 'desired' to 'essential' (future extension) 11.
1.2.1	16.07.2010	Clarified **R-SOW-2.4**, that the RB does not limit storage space anymore.

1.2 Purpose and Scope

The purpose of this document is to describe the requirements baseline of the *Calvados* project. The objective of the Requirements Baseline is to translate the general conception of the workflow outlined in the Statement of Work into a series of specifications describing the system to be implemented.

The requirements compiled in this document are both user and system requirements. They will serve as the primary source for the Calvados system technical specification and final acceptance testing. Particularly, the requirements baseline reflects ESA's and BC's common understanding of the project goal and describes the expectations on the outcome of phase 2.

1.3 Abbreviations

API	Application Programming Interface
BC	Brockmann Consult GmbH
(D)QWG	Data Quality Working Group
LET-SME	Leading Edge Technology - Small and Medium-sized Enterprises
MR	Map/reduce (programming model).
RB	Requirements baseline (document)
SOA	Service Oriented Architecture
SoW	Statement of Work
TS	Technical specification (document)

1.4 References

[RD-1]	Fomferra, N.: *The BEAM 3 Architecture*; http://www.brockmann-consult.de/beam/doc/BEAM-Architecture-1.2.pdf
[RD-2]	Brockmann, C., Fomferra, N., Peters, M., Zühlke, M., Regner, P., Doerffer, R.: *A Programming Environment for Prototyping New Algorithms for AATSR and MERIS – iBEAM*; ENVISAT Symposium 2007, ESRIN, Frascati, Italy; Proceedings 2007
[RD-3]	Fomferra, N., Brockmann C. and Regner, P.: *BEAM - the ENVISAT MERIS and AATSR Toolbox*; MEIRS-AATSR Workshop, ESRIN Frascati, Proceedings 2005
[RD-4]	Sanjay Ghemawat, Howard Gobioff, and Shun-Tak Leung: The Google File System; 19th ACM Symposium on Operating Systems Principles, Lake George, NY, October, 2003. (http://labs.google.com/papers/gfs.html)

[RD-5]	Jeffrey Dean and Sanjay Ghemawat: *MapReduce: Simplified Data Processing on Large Clusters*; OSDI'04: Sixth Symposium on Operating System Design and Implementation; San Francisco, CA, December 2004. (http://labs.google.com/papers/mapreduce.html)
[RD-6]	Ariel Cary, Zhengguo Sun, Vagelis Hristidis, Naphtali Rishe: *Experiences on Processing Spatial Data with MapReduce*; Lecture Notes In Computer Science; Vol. 5566 - Proceedings of the 21st International Conference on Scientific and Statistical Database Management - New Orleans, LA, USA, 2009. (http://users.cis.fiu.edu/~vagelis/publications/Spatial-MapReduce-SSDBM2009.pdf)
[RD-7]	R. Doerffer, H. Schiller: *MERIS Lake Water Algorithm for BEAM and MERIS Regional Coastal and Lake Case 2 Water Project*, Atmospheric Correction ATBD; ESRIN Contract No. 20436, June 2008
[RD-8]	Zhong Ping Lee, Kendall L. Carder, and Robert A. Arnone: *Deriving inherent optical properties from water color: A multiband quasi-analytical algorithm for optically deep waters;* APPLIED OPTICS, Vol.41,No.27, 20 September 2002
[RD-9]	Bryan A. Franz, Sean W. Bailey, P. Jeremy Werdell, and Charles R. McClain: *Sensor-independent approach to the vicarious calibration of satellite ocean color radiometry*; APPLIED OPTICS Vol.46,No.22, 1 August 2007
[RD-10]	Bryan Franz: *Methods for Assessing the Quality and Consistency of Ocean Color Products*; NASA Goddard Space Flight Center, Ocean Biology Processing Group, 18 January 2005, web page: http://oceancolor.gsfc.nasa.gov/DOCS/methods/sensor_analysis_methods.html
[RD-11]	Zibordi, G., Holben, B.N., Hooker, S.B., Mélin, F., Berthon, J.-F., Slutsker, I., Giles, D., Vandemark, D., Feng, H., Rutledge, K., Schuster, G., Al Mandoos, A.: *A network for standardized ocean color validation measurements;* EOS Trans. Am. Geophys. Union, 87, 30, pp. 293,297, 2006.
[RD-12]	Zibordi, G., Mélin, F., Hooker, S.B., D'Alimonte, D., Holben, B.N.: *An autonomous above-water system for the validation of ocean color radiance data*;IEEE Trans. Geosci. Remote Sens., 42, 401-415, 2004.
[RD-13]	Stanford B. Hooker, Elaine R. Firestone, and James Acker: *Level-3 SeaWiFS Data Products: Spatial and Temporal Binning Algorithms*; NASA Technical Memorandum 104566, Vol. 32, August 1995
[RD-14]	Kathryn Barker, Constant Mazeran, Christophe Lerebourg, Marc Bouvet, David Antoine, Michael Ondrusek, Guiseppe Zibordi, Samantha Lavender: *MERMAID: The MEris MAtchup In-situ Database*; MERIS User Workshop 22-26 September 2008 (http://earth.esa.int/cgi-bin/confm8.pl?abstract=181)
[RD-15]	Bryan Franz: *OBPG l2gen User's Guide*; (http://oceancolor.gsfc.nasa.gov/seadas/doc/l2gen/l2gen.html)

2 Introduction

2.1 Overview

ESAs Earth Observation (EO) missions, starting with ERS1/2, followed by Envisat and now with a series of Earth Explorers and the Sentinel missions, provide a unique dataset of observational data of our environment. Calibration of the measured signal and validation of the derived products is an extremely important task for efficient exploitation of the data and the basis for reliable scientific conclusions. ESA has established the instrument of Data Quality Working Groups (DQWG), in charge of permanent control and revision of instrument calibration, product validation and algorithm improvement; at the international level this work is contributing to the tasks of the CEOS Working Group on Calibration and Validation. In spite of this importance, the cal/val work is often hindered by insufficient means to access data, time consuming work to identify suitable in-situ data matching the EO data, incompatible software and no possibility for rapid prototyping and testing of ideas. In view of the future fleet of satellites and the largely increasing amount of data produced, a very efficient technological backbone is required to maintain the ability of ensuring data quality and algorithm performance.

This document compiles the baseline requirements for a demonstration prototype that will demonstrate how cal/val and user services can be significantly **improved** by adopting a **leading edge technology** providing highly interactive, desktop-like online visualisation, analysis, extraction and processing services operating on-the-fly on terabytes of EO data. The technology has been designed for the processing of ultra-large amounts of data and is based on a **massive parallelisation of tasks** and a **distributed file system**, both running on extendible commodity **hardware**. Well known online services as provided by *Google*, *Yahoo*, *Amazon*, *Facebook* and *Last.fm* rely on this technology and its **spin-in application** to space born, spatial data is very eligible and feasible [RD-6].

The outcome of the project will be an EO data processing portal demonstrating to ESA a number of **value-adding** features and new capabilities for CalVal and User Services. The interesting features to be implemented are compute and data intensive and the goal is to make them capable to complete in minutes if not even seconds. Examples are:

- Merging of matching in-situ and EO data at given validation site and time
- Extraction of time series and product subsets of arbitrary variables, locations, time periods
- On-demand processing of higher level (L2, L3) from lower level (L1, L2) products
- Creation of RGB browse images of arbitrary variables or combination of them
- Creation of animation sequences for the 0^{th} and 1^{st} order derivatives of a variable

The actual features to be implemented within the scope of this project are compiled in this document. The selected features will be a trade-off of the benefit they have for users, their demonstration means and the implementation effort.

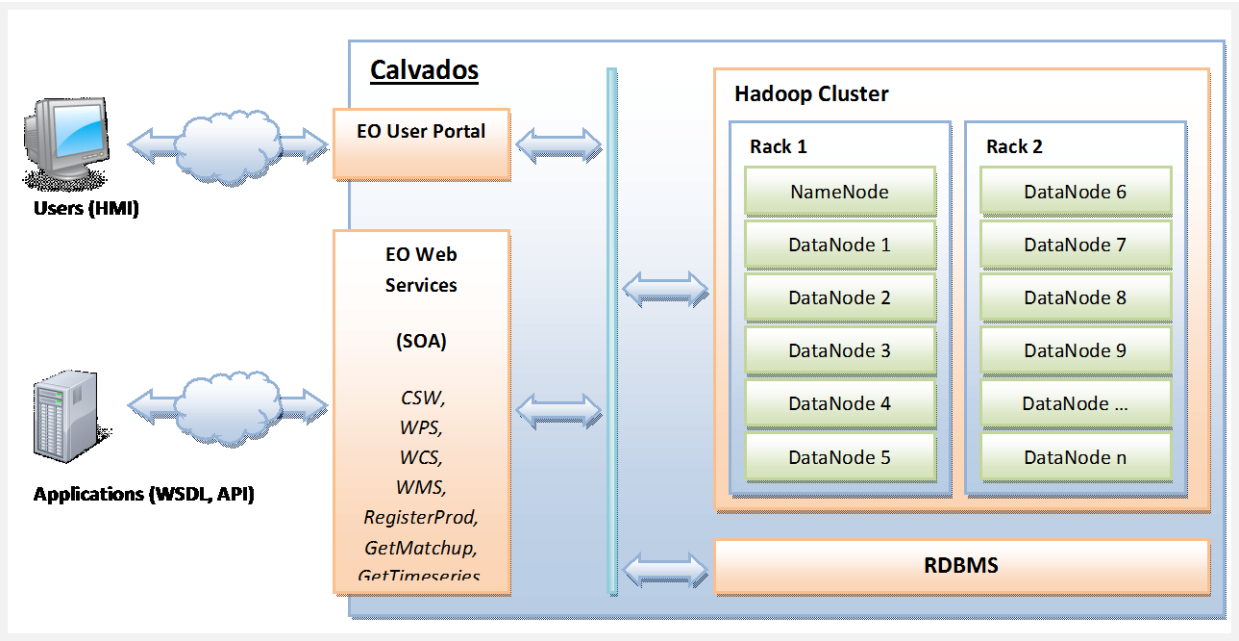
The intended work to be done is both a combination of research studies, **early stage developments** with respect to the new technology and the adoption and reuse of a number of existing software components developed at Brockmann Consult under ESA contract. We would like to point out that the parallelisation technology is on hand in form of stable and reliable software developed in a well maintained, industry sponsored **open-source** project.

In this sense, the ultimate objective of the Calvados project can be summarized as follows

Calvados is a RTD project under the LET-SME initiative that aims to demonstrate that Hadoop technology can be applied to EO data by implementing a demonstration system for CalVal and User Services.

2.2 Proposed Calvados Concept

In the following, a high level specification of Calvados, the demonstration system providing the CalVal and User Services, is given. The innovation in Calvados' architecture is the use of a Hadoop cluster. The following shows the main components of the Calvados demonstrator system.



EO User Portal: The human machine interface used by EO data users offering and demonstrating highly responsive online EO data query, analysis and processing. The EO user portal will be realised using components from MERCI and the CEOS CalVal Portal. The user portal technology is based on the *Liferay* Portlet Server and the *OpenLayers* JavaScript API. The portal will comprise a selection of the following yet non-existing features whose short response time will be achieved by a number of web services profiting from massive parallelization through the Hadoop cluster.

EO User Services: A number of web services through which other applications may access the system by dedicated (XML) protocols. Services include product registration, match-up generation, time series extraction, RGB image generation, as well as L2 and L3 processing. Encapsulating these system capabilities in separate web services allows for a Service Oriented Architecture (SOA) of the system.

Hadoop cluster: A cluster running the latest stable version of *Apache Hadoop*. The implementation of most of the MR tasks running on this cluster (raster data processing) will be based on BEAM technology. Especially the existing BEAM graph processing framework is ideally suited to be parallelized and adopted to the Hadoop MR paradigm.

RDBMS: The relational database management system used to store product metadata. The metadata is stored during product registration and comprises geographical boundaries, number of bands, resolution, etc. The RDBMS comprises a spatial extension capable of storing geometry data and

allowing for fast spatial queries. The database has already been implemented for MERCI and the CalVal Portal.

3 Background of Requirements

3.1 Statement of Work and Proposal, and Proposal Clarification

The primary source for the high-level requirements and therefore the benchmark for the project's success is the Statement of Work (SoW), in this case the ESA LET-SME 2009 call regarding the Announcement of Opportunities "Cal/Val and User Services". LET stands for Leading Edge Technology and SME indicates that Small and Medium-sized Enterprises are the main target of this call. Specifically,

> *LET-SME is a **spin-in** instrument encouraging the participation of SMEs to ESA technology. The LET-SME focuses on early stage development of "Leading Edge Technologies", i.e. the ones likely to become the reference technologies for the near future, and have good chances of being infused into ESA projects and missions.*

In accordance with the SoW, Calvados is a system that has been proposed to fully support the idea of LET-SME, thus with a strong focus on a selected LET, that may have the potential to become a reference in large scale EO data processing.

3.2 Perspective of Validation Activities at ESA

3.2.1 Data Quality Working Group

ESA has established the Data Quality Working Groups (DQWG) after the completion of Envisat's commissioning phase in fall 2002. The mission of the data quality working group is to monitor the quality of the instrument products as produced by the satellite ground segments, and to recommend algorithm improvements including suggestions for new products. DQWGs exist for MERIS, AATSR and Atmospheric Composition instruments.

The DQWGs are composed of scientists being expert (or even developer) of the science algorithms, and technical experts on algorithm implementation. Main tool of the DQWGs is the instrument data processor prototype. Algorithm changes, auxiliary data changes and ideas for new products are prototyped in this environment and tested before proposed for implementation in the operational ground segment processor.

3.2.2 Instrument Validation Teams

A larger group of scientists form the instrument validation teams. Experts submit a proposal for a validation activity which are then formalized in a project under which data are made available free of charge already during commissioning phase to the scientists. A limited number of such projects is funded; mostly the project are financed through other than ESA sources. After launch of Envisat, the MERIS and AATSR Validation Team (MAVT) and the Atmospheric Chemistry Validation Team (ACVT) were established. The activities and lifetime of these validation teams were linked with the Envisat commissioning phase. However, the MERIS Validation Team was reactivated in 2009 in order to further support the work of the MERIS DQWG, in particular for the validation of the Case2 Water processing. A calibration and validation team was also implemented for the SMOS mission. This group has currently started its activities with the advent of SMOS data products in late 2009.

3.2.3 Perspective

The concept of a DQWG is closely linked to the concept that a satellite system is composed of a space and a ground segment delivering higher level data products, and that the space agency operating the satellite is delivering the higher level products only through this path.

This concept has already broken up in the past where higher level products are generated by the Agency by different means than the ground segment processor. Examples are the ESA GlobColour project which delivered the same kind of ocean colour products, but with different quality, error indicators and a validation report, and which is continued now in the GMES myOcean Marine Core Service (also an activity supported by ESA). As a consequence, operational usage of ocean colour data in Europe for GMES and coming from MERIS will be taken from myOcean and not from the ESA ground segment. The base products for myOcean are the standard MERIS Level 2 ocean colour products, but due to additional processing (filtering, averaging, merging of Case1 and Case2, and also with MODIS) these products are finally different from the ground segment products. Another example would be ESA GPOD through which the MGVI is made available (the same product as provided by the ground segment, but including more bands and L3 products) as well as new products such as the MERIS cloud fraction in SCIAMACHI.

Another evolutionary step is to give software and L1 data to the user, instead of processed L2 and L3 data. Also this path is supported by ESA through the ESA instrument toolboxes. Examples are the BEAM processors for MERIS and AATSR, or NEST processors for ASAR. NASA has realised this concept from the beginning through the SeaDAS toolbox (NASA is providing both, the software but also the L2/L3 products). This concept provides more flexibility to the user to configure her/his products, and reduces the processing and distribution work load for the agency.

From a user point of view all products and services including the ground segment processor, the Glob-projects, the GPOD or the BEAM processor come from ESA and should be of best quality.

3.2.4 ESA Climate Change Initiative

In the future this situation will become more complex, or possibly simplified, through the ESA Climate Change Initiative (CCI). One key objective of the CCI is to provide best quality long term time series, and the CCI projects on Essential Climate Variables have the task to review L1 processing including calibration and geo-location, and all Level 2 processing algorithms. Where necessary and possible new and better algorithms than the standard (=those used in the ground segment processors) shall be deployed. Assuming that the ECV teams will do their job properly, their result should be fed back into the ground segment processors. The role of the DQWGs is not clear in this context and should be re-defined for CCI and the operational phase of GMES (=Sentinels). Task 5 of the ECV contracts will be dealing with the future (=Phase 2 of CCI) operational implementation of the ECV processors. This task will also deal with the (systematic and automated) validation of the products including reprocessing.

3.2.5 ECV-OC and Coastcolour

The most important products from MERIS are the ocean colour products, and the MERIS DQWG is concerned most of its time with this set of algorithms and products. In the context of the ECV-Ocean Colour (ECV-OC) the algorithms and products will be reviewed. Error products shall accompany each derived variable. The Ocean Colour ECV project will start in late 2010 and results concerning product quality and validation activities will not be available before late 2011.

However, the Coastcolour project has been kicked-off in January 2010 and will last until 2011. This project is contributing the coastal component also to the CCI; coastal waters are excluded from the ECV-OC statement of work with reference to Coastcolour. The requirements on product quality and the task to critically review L1 and L2 processing is identical in Coastcolour and ECV-OC.

There are several key requirements in Coastcolour on validation:

- Definition of a standard set of products from different satellite missions; primary interest is MERIS, but for comparison also MODIS and SeaWiFS will be considered
- Compilation of an in-situ database with reference data to be used for validation of the standard products (point 1 of this list)
- Definition of standard tests to be applied to the standard products after algorithm changes, and for inter-comparison of different products and algorithms
- Frequent repetition of the tests upon algorithm changes
- Keeping history of algorithm changes and processing version
- Automated processing of the tests and evaluation
- Transparency of the process through an open, web based system

This list is not innovative since it describes best practice for validation. The important aspects here are

- the instrument concerned: MERIS,
- being an internal activity of highest awareness,
- linked with the IOCCG, NASA and CEOS WGCV,
- the perspective of this activity: continuation in the CCI and hence preparing the future activity of ESA DQWGs
- timing: the project runs in parallel with Calvados, and results are expected also in line with the Calvados schedule

Coastcolour Level 1 to Level 2 processing, or part of this, is therefore an ideal candidate to be linked with Calvados. The Level 1 to Level 2 processing consists of an atmospheric correction based on a neural network inversion of the radiative transfer equation, and the calculation inherent optical properties using also neural network inversion techniques as well as a semi-analytical approach. The neural network inversion technique will be realised by the GKSS Case2R processing scheme, and the semi-analytical algorithm by the Quasi-Analytical Algorithm (QAA) from Mississippi State University (ZhongPing Lee).

An important component of the Coastcolour project is an inter-comparison with standard MERIS processing, SeaWiFS and MODIS products, as well as a comparison with in-situ data. This links the Coastcolour processing with standard MERIS processing, as well as with NASA standard processing. Scientists from MERIS QWG (Doerffer, Fischer, Brockmann) and from NASA (Franz, Feldman) are contributing to this inter-comparison.

3.2.6 ESA Sentinel Missions and the Future

ESA's living planet programme is currently in a transitional phase, characterised by the end of the lifetime of ENVISAT, the preparation of the future operational Sentinel missions, and the growing number of Earth Explorer missions. As described above it is quite likely that the successful work of the DQWG will evolve as well into a concept that is currently not clear. The European FP7 programme includes in its R&D projects validation activities (e.g. myOcean validation, Aquamar Downstream project has two work packages on validation and validation technique evolution). The

recently extended ESA GSE projects (e.g. Marcoast) include also important activities on validation. All these are preparing the future for the operational calibration and validation of the ESA Sentinel missions, and the scientific Earth Explorer missions. The next step in this future will be through the ESA CCI initiative and the projects contributing to this today, such as Coastcolour.

The aim of Calvados is also to prepare the future and to test novel concepts to support calibration and validation activities, which will take in the future. The project should learn from today's obstacles for CalVal and should link with future Cal/Val activities in ESA in order to prepare a technological environment for their work. This could be best achieved by putting the CCI into its focus and developing its concepts at the example of the Coastcolour project.

4 Requirement Conventions

4.1 Notation

Each requirement is, whenever applicable, specified using the following notation scheme:

R-<group>-<number> Here the detailed description of the requirement.

Priority: *Here one of* **Essential** | **Expected** | **Desired** | **Optional**	**Risk:** *Here one of* **Dangerous** | **Risky** | **Fair** | **Safe**	**Effort:** *Here one of* **Months** | **Weeks** | **Days** | **Hours**
Here an optional explanation of the selected priority rating	*Here an optional explanation of the risk selected rating*	*Here an optional explanation of the selected effort rating*

Realisation: *Here a brief description of how the requirement is planned to be realised.*

4.2 Terminology

A number of the technical requirements in this document refer to the following common terms and concepts.

Product query: Defines all the inputs required to perform a product data query on the EO data currently stored in Calvados. Product queries can be stored for later use.

Product set: A set of persistent (L1) or temporary (L2, L3) EO data products. Either a result of a submitted product query, a manual selection of EO data products, or a processing job.

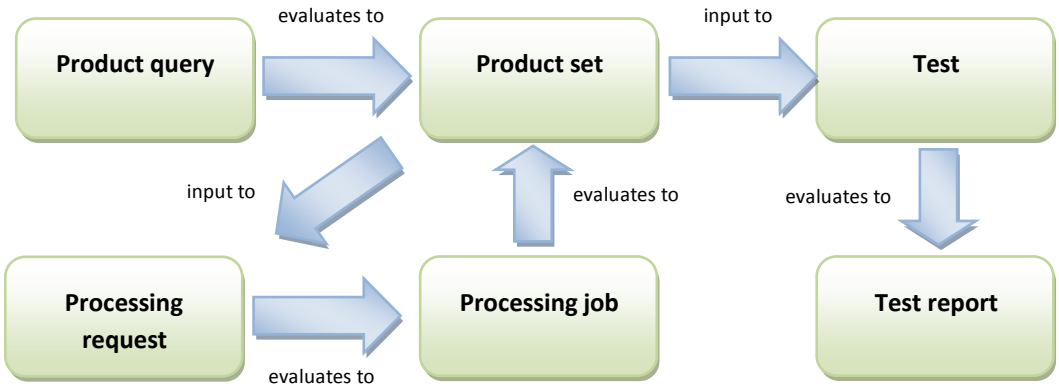
Processing request: Defines all the inputs required to perform a certain EO data processing. Inputs are the type of processing to be performed, the (input) product set and optionally specific, user -provided processing parameters.

Processing job: Result of a submitted processing request. Processing jobs represent server-side, temporary processes. A job has a unique ID, is owned by the user who has invoked it and has a visible, queryable status.

Test: A specific (statistic) analysis that is performed on a specific product set. Certain tests may require additional user input, see also [RD-8].

Test report: The result of a submitted test. The report is a document which may comprise various diagrams and data tables.

The relationship of these concepts are shown in the following diagram:



4.3 Users and Roles

The following user roles are used for the definition of the Calvados requirements:

Guest: Users with no special privileges, e.g. visitor of the Calvados project page.

User: Users registered with the Calvados system, such as scientists in charge of specific cal/val topics, with special Calvados privileges allowing them do perform operations a **Guest** usually cannot perform or cannot even see.

Administrator: Equal to **User**, but can additionally perform some Calvados set up, configuration and maintenance tasks.

Developer: Equal to **User**, but also in the position to extend or modify the Calvados processing and analysis services by developing new or change existing algorithm source codes.

5 Requirements

This chapter compiles the high-level project, functional and non-functional requirements for the Calvados project.

5.1 High-Level Project Requirements

All requirements in this chapter are considered to be **Essential** because they are requirements directly extracted from the SoW and as such, they directly formulate the project goals.

5.1.1 General LET-SME Requirements

R-SOW-1.1 It is up to the bidder to propose the adequate technology that fulfils the described needs.

Priority: **Essential** **Risk:** **Fair**

Not safe, but fair, because the technology might not be directly applicable all domain-specific problems.

Realisation: *The proposed technology is Apache Hadoop.*

R-SOW-1.2 Although LET-SME seeks innovation, technologies or processes proposing a significant improvement of existing ones will be considered, if this improvement is clearly substantiated. However, proposals overlapping or extending other on-going ESA activities will not be accepted.

Priority: **Essential** **Risk:** **Safe**

Safe, if differences with e.g. ODESA can be clearly substantiated.

Realisation: *There is no ESA project known, which proposes using Hadoop as spin-in technology in the domain of cal/val and user services. On-line data processing and cal/val and user services are addressed in the project ODESA.*

R-SOW-1.3 The proposed work shall be product/process oriented and shall include as much as possible, hardware and software feasibility demonstration, prototyping and validation.

Priority: **Essential** **Risk:** **Safe**

Realisation: *See **R-SOW-1.2**.*

R-SOW-1.4 The proposed programme of work shall aim at resolving the uncertainties on the concept feasibility and applicability to space.

Priority: **Essential** **Risk:** **Fair**

Project resources might not be sufficient to address all possible applications.

Realisation: *The outcome of the project will be a prototype system which will clearly demonstrate its applicability to cal/val and user services. The features of resulting prototype will be carefully selected so that the concept feasibility and applicability to EO data are addressed to a maximum degree.*

R-SOW-1.5 The added value of the proposed technology and its innovative content with respect to space shall be highlighted, and shall give first indications on any possibility of on-ground commercial exploitation.

Priority: **Essential** **Risk:** **Fair**

See R-SOW-1.1, the technology might not be directly applicable to all domain-specific problems.

Realisation: *Hadoop, the MapReduce programming model and EO data locality are expected to boost the performance of processing systems so that highly interactive cal/val and user services become possible. Once the prototype shows its applicability it may be extended to other sensors, missions, data models and processing chains.*

5.1.2 Specific SoW Requirements (LET-SME Solicitation Topic)

R-SOW-2.1 In order to support ESA calibration and validation activities and advance the exploitation of current and future optical missions, there is a need to develop efficient quality control, data extraction and analysis services. The services to be developed under this activity will support DQWGs in evaluating instrument calibration and algorithms, and will provide data assistance to the user community. The system shall have two important links: to the Earth Observation (EO) data files; and to level 2 and 3 processing algorithms. The service will have to handle in a highly parallel fashion data acquisition, storage, management, and processing. The system shall be capable of serving QWGs with specific cal/val tests (e.g. mission-long time series, match-ups), analyses, validations, a data merger. At the same time, it will perform standard and non-standard data extraction and processing requests from users.

Priority: **Essential** **Risk:** **Fair**

See R-SOW-1.1.

Realisation: *BC has a long-term experience in developing and providing such systems and services. However the usage of Hadoop as the backbone processing system is new.*

R-SOW-2.2 The list of desired features and services and the resulting system architecture shall be linked to the goals of ESA data cal/val, Global Monitoring for the Environment and Security (GMES), and Climate Change Initiative (CCI) programs.

Priority: **Essential** **Risk:** **Safe**

Realisation: *The list of desired features will be linked to the perspective of validation activities at ESA as outlined in chapter 2.2.*

R-SOW-2.3 The system will comprise an efficient web-based front-end and an anticipated physical infrastructure. A complete system design is anticipated which will serve the selected mission. At the same time, the design should be general, so as to be readily adaptable to other EO missions.

Priority: **Essential** **Risk:** **Fair**

Not safe, but fair, because Hadoop technology is new to EO data application and new to BC.

Realisation: *The web-based frontend technology is well known (Frontend: Liferay Portal-Server, Portlets and Apache Wicket. Backend: Apache Hadoop and JAX-WS, SOAP-based web-services).*

BEAM is considered to provide the mission neutral EO data processing framework.

R-SOW-2.4 The practical design of the services shall consider a single Envisat imager, i.e. MERIS (Medium Resolution Imaging Spectrometer). The size of the dataset will depend on the demonstration analyses that shall be implemented (refer to chapter 5.2.4) in the demonstration system. The input data products shall cover

1. At least 5 years MERIS L1b RR data, limited to the first 4 days of each month. This data will be mostly used to perform Test 2, see **R-FN-4.5**.
2. Greatest possible amount of MERIS L1P FR data subsets in order to reasonable perform Test 1, see **R-FN-4.4.**

Priority: **Essential** **Risk:** **Safe**

Realisation: *Extending the data set to the full MERIS mission would be beneficial for an operational application, but not necessary for a demonstration system. The hardware has been dimensioned accordingly. The idea to work on the full MERIS mission is fully supported once the required extra resources could be made available.*

R-SOW-2.7 The potential interface with GECA should be addressed in the context of validation with ground measurements (http://earth.esa.int/cgi-bin/confatmos09.pl?abstract=380).

Priority: **Optional**

Calvados communicating with GECA web services is definitely a reasonable and also feasible feature.

Risk: **Risky**

Incorporation of data sources, other than the ones required for the prototype system, produces an inacceptable overhead which does not improve the demonstration of applicability the proposed technology.

Realisation*: A link to the GECA development team has been established. For a Calvados evolution, GECA could serve as satellite and reference data provider.*

5.2 Functional Requirements

5.2.1 General EO-User Portal Requirements

R-FN-1.1 Calvados shall comprise an EO-User portal offering as the frontend for various Calvados services. The portal will offer at least access to services like the management of users, data product queries, and processing requests.

Priority: **Essential** **Risk:** **Safe** **Effort:** **Months**

Realisation: *The portal server Liferay will be used to implement the Calvados EO-User portal.*

R-FN-1.2 Guests shall be able to visit the Calvados portal and inform themselves about the Calvados project, its goals and capabilities. Guests will only be allowed to access static information.

Priority: **Desired** **Risk:** **Safe** **Effort:** **Days**

Realisation: *The portal server Liferay handles different user privileges and also provides a fully featured user management. It will be very easy to become a user – just to give an e-mail address and establish a username and password. Access to Calvados services should not be limited. For demonstration, only a few people will be allowed to use the system – so there is no problem. When one day Calvados becomes operational and works on a large cluster, anybody should be able to put in a request. It will then be up to Calvados scheduling to prioritise the jobs and not to overwhelm the cluster.*

R-FN-1.3 Users shall be able to sign into the Calvados system. Only registered users shall be permitted to sign in. Signed-in users can access the dynamic content and features of the Calvados portal. Admins can additionally perform configuration and maintenance tasks.

Priority: **Expected** **Risk:** **Safe** **Effort:** **Days**

Realisation: *See **R-FN-1.2**.*

5.2.2 Data Management Requirements

R-FN-2.1 Users shall be able to define data product queries on the data product archive. The query definition shall at least specify

- the product types (e.g. MERIS L1),
- the time period,
- the region either given by a selected in-situ site (see **R-FN-4.1**) or a user defined polygon.

The query user interface comprises components for the query parameter and includes a dynamic, multi-layer World map component where user can draw polygons.

Priority: **Expected** **Risk:** **Safe** **Effort:** **Weeks**

Realisation: *Similar to product query pages in CalVal Portal or MERCI.*

R-FN-2.2 After submitting a query, users shall be able to browse the resulting product set including image thumbnails, basic quality information, statistics and metadata. The data products swath boundary shall be displayed in the World map component.

Priority: **Desired** **Risk:** **Safe** **Effort:** **Days**

Realisation: *Similar to product query pages in CalVal Portal or MERCI.*

R-FN-2.3 After submitting a query, users shall be able to select the data products for downloading. The download may be online (data streaming) or offline (ftp delivery). A download quota is not neither forseen for the demonstration system nor for an operational system. Cavados should implement staging strategies that allow the generation of product sets of virtually any size. In the demo system, these strat. need not to be considered.

Priority: **Expected** **Risk:** **Safe** **Effort:** **Days**

Realisation: *Similar to product query pages in CalVal Portal or MERCI.*

R-FN-2.4 Users shall be able to name a specific data product query and store it for later use. Users shall be able manage their data product queries. The required operations are: add, display, edit, list, select, remove queries. Listed data products can be selected, so that bulk operations, such as removing all selected items, can be performed (see also **R-FN-3.1**).

Priority: **Desired** **Risk:** **Safe** **Effort:** **Weeks**

Realisation: *Data product queries associated with a particular user will be stored in a relational database.*

5.2.3 Data Processing Requirements

R-FN-3.1 Users shall be able to define L2 and L3 processing requests. For L2 processing, the inputs are

- the L2 processor type,
- the L1 input product(s),
- the processing parameters.

For L3 processing the inputs are

- the L2 input product(s),
- the processing parameters.

A processing request may be defined for single, selected input products or for a set of input products (bulk-processing). The set of input products shall either be defined by the current selection set or the result set of a named, user-defined data product query (see **R-FN-2.4**).

Priority: **Expected** **Risk:** **Fair**. **Effort:** **Weeks**

Realisation*: This is the actual challenge in this project: How to adopt existing EO data processors to MR. This can be a more or less risky tasks depending on the implementation language, processor architecture, documentation, coding style, etc.*

R-FN-3.2 Calvados demonstration shall implement two or more different L2 processor types. Different processors may output different geophysical parameters or output same geophysical quantities but using different algorithms. The selection of data processors to be integrated shall be based on the following criteria:

- the algorithms shall be recognized by the community, peer-reviewed and well documented,
- their source code shall be comprehensible and easy to change (see R-FN-3.14),
- their source code shall input and return default parameters and have an option to compute and output only the parameters selected by the user in order to reduce the CPU and I/O time as well as disk space available to the Calvados demonstration.

Priority: **Expected** **Risk:** **Fair** to **Dangerous** **Effort:** **Weeks** to **Months**

The risk strongly depends on selected processor code base.

Realisation*: This is the actual challenge in this project: How to adopt existing EO data processors to MR. This can be a more or less risky tasks depending on the implementation language, processor architecture, documentation, coding style, etc.*

R-FN-3.3 Regarding **R-FN-3.2**, the following L2 shall be considered to be implemented in the demonstration system (converted to MapReduce):

- Processor 1: MERIS Case2R L2 Processor by Roland Doerffer [RD-7] (essential)
- Processor 2: MERIS QAA by Zhong Ping Lee [RD-8] (essential)
- Processor 3: *l2gen* by Bryan Franz [RD-15] (optional)

Any future improvement coming from Coastcolour shall be integrated in Case2R and QAA when available. *l2gen* is an important processor, particularly in the context of ECV-OC. The actual integration of l2gen in Calvados is optional in the Calvados project.

Priority: **Essential** **Risk:** **Fair** **Effort:** **Weeks**

Realisation: *Reuse the BEAM Case2R and QAA processors and adapt them to MR, see also note on **R-FN-3.2**. Since l2gen is implemented in C, the C-API as described in **R-FN-3.13** will be used to integrate l2gen.*

R-FN-3.4 Users shall be able to alter the processing parameters for a processing request. There are several options to realise this. Each processor type could have its own user interface for changing its particular processing parameter values. The user interface may be generated using the description (metadata) provided with each of the parameters associated with a processor type. Alternatively, users have the possibility to upload a processor-specific parameter (key-value) file and/or have the option to edit the parameter table directly.

Priority: **Essential** **Risk:** **Safe** **Effort:** **Weeks**

Realisation: *The BEAM team is experienced in creating such generic systems. The BEAM GPF implements similar features.*

R-FN-3.5 Users shall be able to submit their processing requests which will translate into a server-side processing job. A job comprises a highly distributed number of tasks running in parallel. After a job has been successfully invoked its status is set to "Running". If it cannot be invoked, the user will see a detailed error message.

Priority: **Essential** **Risk:** **Safe** **Effort:** **Weeks**

Realisation: *The processing service will translate processing requests into processing (Hadoop) jobs. Hadoop then translates jobs into distributed tasks. Processing jobs will be uniquely identified by their ID.*

R-FN-3.6 Users shall be able to cancel running processing jobs. Calvados will make best efforts to immediately terminate the related server-side processes in order to switch the state from "Running" to "Cancelled".

Priority: **Expected** **Risk:** **Safe** **Effort:** **Days**

Realisation: *The processing service will have the capability to cancel jobs for a given job ID.*

R-FN-3.7 Users shall be able to monitor the state and progress of processing jobs. After a processing job finalizes on the server side, it changes its state to either "Completed" or "Failed". In the case of success, the return value is a new, named set of output products (see **R-FN-2.1**). In the case of failure, a detailed error message will be shown to users.

Priority: **Expected** **Risk:** **Safe** **Effort:** **Days**

Realisation: *The Hadoop API allows for querying a job's status.*

R-FN-3.8 Users shall be able to aggregate a set of L2 input products originating from a data product query (**R-FN-2.1**) or a L2 processing request (**R-FN-3.1**) (satellite coordinate system) into L3 products (geographical coordinate system). The reference aggregation algorithm shall be the L3 binning scheme implemented in SeaDAS (which is also implemented in BEAM), see [RD-13]. The L3 binning shall also run in fully distributed mode. For the L3 product generation, the period of *p* days for which the L2 output data is available may be split into *n*-day (e.g. n=1...30) intervals leading to a time series of *floor(p/n)* L3 output products.

Default L2 mask definitions shall be associated with each geophysical L2 variable to be processed to L3. The L2 masks used by the binning should be an editable field of a processing request and may be changed by the user (e.g. Boolean flag expression).

Priority: **Expected** **Risk:** **Safe** **Effort:** **Weeks**

Realisation: *Reuse the BEAM L3 processor and adapt it to MR.*

R-FN-3.9 Bulk L2 processing requests shall provide the option to directly invoke the L3 binning for the generation of time series for further analysis.

Priority: **Expected** **Risk:** **Safe** **Effort:** **Days**

Realisation: *Reuse the GUI from **R-FN-3.8**. Invoke* the L3 service on the set of output products.

R-FN-3.10 Users shall be able to name a specific processing request and store it for later use. Users shall also be able manage their stored processing requests. The required operations are: new, delete, display, edit, list and select processing requests. Listed processing requests can be selected, so that bulk operations, such as removing all selected items, can be performed.

Priority: **Desired** **Risk:** **Safe** **Effort:** **Weeks**

Realisation: *Data product queries associated with a particular user will be stored in a relational database.*

R-FN-3.11 For developers, Calvados shall provide a dedicated EO data processing API for the implementation of new L2 data processors. Using this API, developers can concentrate on implementing their EO algorithms, without having to know about the details of the Hadoop MapReduce API. Processing modules developed using this API are plug-ins to the Calvados system: once they are deployed, users can choose the new or modified version of a particular processor type for defining their new processing requests (see **R-FN-3.1**). Deployment can be done either by

uploading the module to a dedicated location or simply by sending to an administrator who can manually do the integration.

Priority: **Expected** **Risk:** **Fair** **Effort:** **Weeks**

Realisation: *The BEAM development team at BC is very experienced in developing EO data processing APIs and designing plug-in-based systems (new modules are dynamically bound to the system, instead of linking them statically).*

R-FN-3.12 The Calvados data processing API shall be available for the Java programming language. More specifically, and in order to support a well established user community, it shall be compatible with the programming models exposed in the existing BEAM Core and the BEAM Graph Processing Framework APIs. Calvados shall provide a simple but functional code example for the Java API.

Priority: **Expected** **Risk:** **Fair** **Effort:** **Weeks**

Realisation: *There is out-of-the box support in Hadoop for distribution of MapReduce task to cluster nodes.*

R-FN-3.13 The Calvados data processing API shall also be available for the C or C++ programming languages. The API will be more simplistic than the Java API but its programming models should be similar and comparable. Calvados shall provide a simple but functional code example for the C/C++ API.

Priority: **Expected** **Risk:** **Fair** **Effort:** **Weeks**

Realisation: *See **R-FN-2.12**.*

R-FN-3.14 Developers shall be able to checkout a new branch for an existing Calvados processor type from a dedicated, publicly accessible version control system (VCS), modify the code in the branch, change the module's version, build it and deploy the compiled module.

Priority: **Desired** **Risk:** **Fair** **Effort:** **Days**

Realisation: *We consider Calvados being an open-source project (envisaged license is GPL 3). As such we provide a publicly accessible VCS repository at* Github *(http://github.com/bcdev/).*

R-FN-3.15 The output format of the product sets generated by a processing job shall be configurable. It is envisaged to use a format that is well-known and accepted by the targeted user community, for example HDF-4 (-EOS), NetCDF-3 (-CF), GeoTIFF or BEAM-DIMAP.

Priority: **Expected** **Risk:** **Safe** **Effort:** **Days**

Realisation: *The BEAM API abstracts from the physical output format. Product writers for certain output formats are plug-ins for BEAM.*

5.2.4 Analysis Requirements

R-FN-4.1 Users shall be able to perform well-known, specific analyses on processed EO data, referred to as *Test*. In general, a test takes as input one or more product sets and a number of control

parameters (optional). It then performs some statistical algorithms and outputs an analysis results comprising data tables and diagrams. The tests applied in Calvados shall apply to AC (water leaving reflectances and AOT) and water products (water leaving reflectances, IOP, concentration, whatever is available). The reference for the implementation of dedicated validations shall be the algorithms described in [RD-9] to [RD-14]. The tests shall provide algorithm inter-comparison and tuning relevant for CoastColour and shall taking into account ECV-OC requirements.

Priority: **Essential** **Risk:** **Safe** **Effort:** **Days**

Realisation: *For the selected MERIS RR data set, the MERMAID in-situ database will be used.*

R-FN-4.2 Calvados shall be able to access selected reference data sets. A reference data set provides in-situ data for specific test sites and for specific geophysical parameters. The reference data sets provided with Calvados will have to provide in-situ data for the output parameters of the selected L2 processors (see **R-FN-3.2**). The primary source for the in-situ data shall be *MERMAID* [RD-14].

Priority: **Essential** **Risk:** **Safe** **Effort:** **Days**

Realisation: *For the selected MERIS RR data set, the MERMAID in-situ database will be used.*

R-FN-4.3 The candidates for the tests to be implemented are

- Test 1: L2 Match-up generation over CoastColour sites (essential)
- Test 2: L3 Time Series over stable waters (essential)
- Test 3: L2 to L3 Comparision (desired)

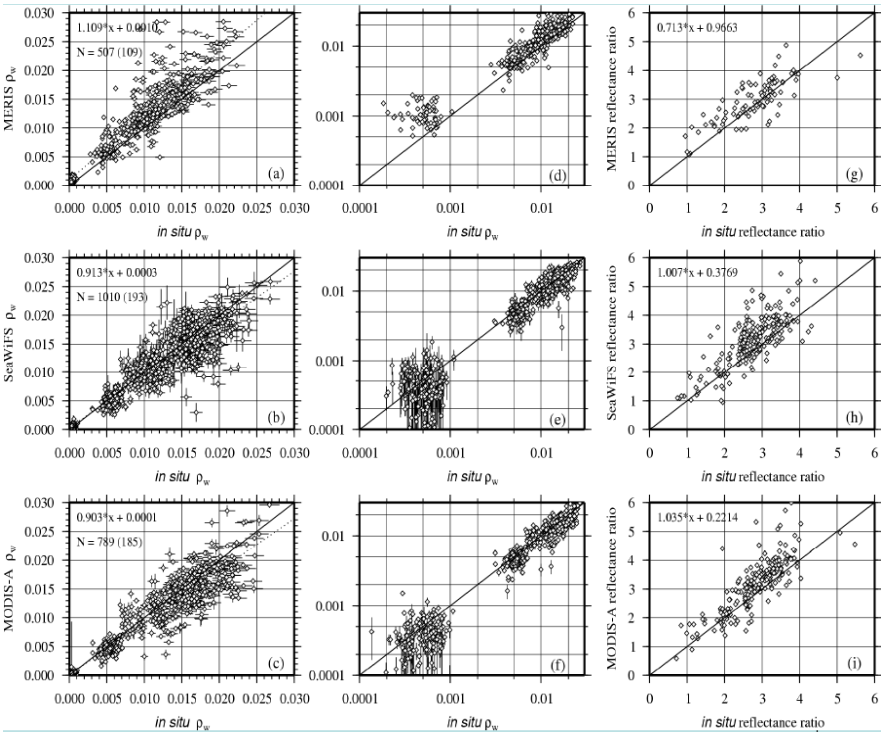
Calvados will provide at least Test 1 and 2 as built-in tests. Test 3 is optional and may also include testsfor statistical analysis of the vertical striping (to test reduced noise due to MB method) and other test on camera interfaces, Level 2 artefacts, detector-to-detector discontinuity etc.

Priority: **Essential** **Risk:** **Safe** **Effort:** **Weeks**

R-FN-4.4 Test 1 shall allow for generating match-ups of

- in-situ data with EO data subsets comprising
- MERIS from Calvados (CoastColour) processing and
- MERIS from standard processing,
- MODIS products,
- SeaWiFS products.

A single match-up file combines EO data and in-situ data (see **R-FN-4.2**) in coincident in space and time. The match-up files shall also comprise spatial EO-data statistics (min, max, mean, standard deviation, meridian) for the provided time period. The report generated by Test 1 includes a matrix of scatter plots created by comparing in-situ data and EO data data with each other source. An option shall exist to also include the extracted match-up files whose format shall be Excel, plain text or or NetCDF/CF.



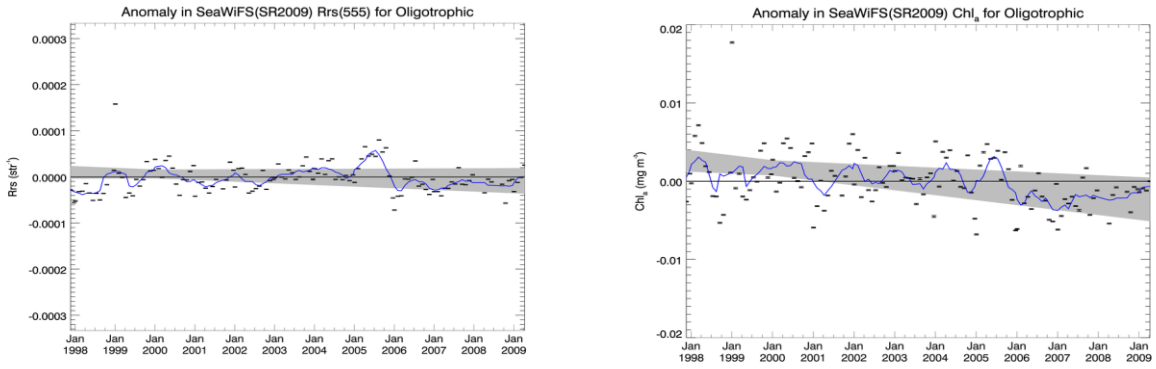
Example plots as part of the Test 1 report (from Boussole test site)

Priority: **Essential** **Risk:** **Safe** **Effort:** **Days**

R-FN-4.5 Test 2 shall allow for generating time-series. It shall run at first priority on a global dataset (>1km depth, first 4 days of a months --> 12x4=48 days/year). Second priority time-series shall be generated from the regions

1. South Pacific Gyre (SPG),
2. South Indian Ocean (SIO) and the
3. Eastern Mediterranean Sea (EMS).

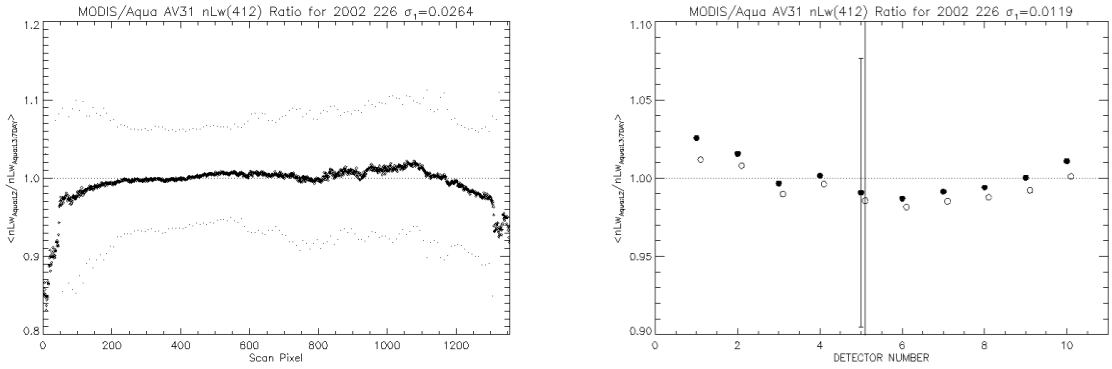
SPG, SIO and EMS will only need to be performed if the technical capabilities of the demonstration system turn out to be sufficient.



Example plots as part of the Test 3 report

Priority: **Essential** **Risk:** **Safe** **Effort:** **Days**

R-FN-4.6: Test 3 shall allow for L2 to L3 inter-comparisons by collecting all L3 bins for a 15-day or a monthly period and comparing the values of matching bins with the across-track pixel values of a L2 product selected from the mid of the aggregation period.



Example plots as part of the Test 3 report

Priority: **Essential** **Risk:** **Safe** **Effort:** **Days**

R-FN-4.7: Calvados may also provide a simple query language which would allow for analyses running on the entire EO dataset. An analysis request could be submitted via a command line tool similar to the SQL console known for all widely known database management systems. This is probably not out of the scope of in this project but should be kept in mind.

Priority: **Optional** **Risk:** **Fair** **Effort:** **Weeks**

Realisation: *The Hadoop* pig *tool is a solution for a similar use case (http://hadoop.apache.org/pig/). The query language is named* pig-latin*. Pig translates a query request into a chain of MapReduce jobs submitted to the Hadoop cluster. Possibly pig can be customized for Calvados needs*.

5.3 Non-Functional Requirements

5.3.1 Performance Requirements

R-NF-1.1 The Calvados prototype system shall demonstrate that it is innovative in the sense that

- it can process EO data in a fractional amount of time than it would take on standard user workstation,
- its hardware and maintenance is cheap compared to standard operational processing systems,
- it is fault tolerant so that it can continue operation not only if one or more hard drives fail but also even if entire machines crash,
- its underlying hardware is cheap compared to standard operational processing systems, and can be extended easily at any point in time.

Priority: **Essential** **Risk:** **Risky** **Effort:** **n.a.**

Realisation: *Hadoop MapReduce and HDFS are the proposed spin-in technology. Naturally, there is a risk associated with spin-in technologies, i.e. technologies that have rarely or never been used in a certain problem domain such as EO cal/val and user services.*

R-NF-1.2 The Calvados prototype system expected that the system's runtime performance shall scale almost linearly with the number of machines in the cluster.

Priority: **Expected** **Risk:** **Risky** **Effort:** **n.a.**

Realisation: *It is a well-known fact that Hadoop MapReduce clusters have a performance that is proportional to the number of nodes in the cluster, but only provided that the algorithms in use allow for arbitrary parallelisation. So the L2 and L3, and analysis algorithms need to be carefully implemented by exploiting the MapReduce programming model to a maximum extend.*

R-NF-1.3 The Calvados prototype system shall provide storage for at least 2 years of MERIS RR data (~6 TB). At least the double of that space is required to hold temporary L2 and L3 processing results. The minimum data replication level of 3 shall be envisaged.

Priority: **Essential** **Risk:** **Safe** **Effort:** **n.a.**

Realisation: *The hardware used will have at least 6 TB x 2 x 3 = 36 TB.*

5.3.2 Architecture

R-NF-2.1 Calvados shall have a Service Oriented Architecture (SOA). It shall comprise a EO User Portal as frontend (Portal Server) and encapsulate backend services such as data product catalogue and the processing system (which again is utilising Hadoop) as web services.

Priority: **Expected** **Risk:** **Safe**

Realisation: *For the realisation of this architecture, BC utilises the following technologies:*
Frontend: Apache Tomcat, Apache Wicket, Liferay portal server, Portlet Specification

Backend: java.net Metro, JAX-WS Web Service Specification, Apache Hadoop MapReduce, -HDFS, and -HBase

---

R-NF-2.2 The Calvados data models and services shall be as generic as possible in order to allow for a future extension to data, algorithms and analyses dedicated to other sensors and missions.

Priority: **Essential** **Risk:** **Safe**

Realisation: *For processing, BEAM provides the core EO data models. Tests and data product services, will be implemented sensor-neutral.*

R-NF-2.3 Calvados shall use Hadoop MapReduce as primary technology providing the massive task parallelisation required for fast processing of large amounts of satellite data.

Priority: **Essential** **Risk:** **Risky**

MapReduce is the spin-in technology

*See **R-NF-1.2***

R-NF-2.4 Calvados shall use the Hadoop Distributed File System (HDFS) as primary file system for

- permanent and temporarily EO data storage

- addressing the locality problem (processing shall occur local to the data),
- implementing data redundancy.

Priority: **Essential**
Proposed spin-in technology.

Risk: **Risky**
*See **R-NF-1.2***

R-NF-2.5 The Calvados prototype system requires a cluster of computers capable of demonstrating the benefit of using massive parallelisation for cal/val and user services. All machines in the Hadoop cluster shall have identical specifications, at minimum

- quad-core CPU, 2.0 GHz
- 4 GB RAM
- 4 TB disk space

The number of machines in the cluster shall not be less than 12 in order to

- assess the system performance when changing the actual number of machines involved in the cluster, and
- to provide sufficient total storage space.

The total price for the hardware shall not exceed 25,000€.

Priority: **Essential**
This is the minimum required hardware for an efficient Hardware cluster.

Risk: **Risky**
The number of nodes may be to less to demonstrate the power of the Hadoop cluster

Setting up the hardware.

Realisation: *BC will procure a system of 12 machines, including network switches and two racks.*

R-NF-2.6 The target operation system for the backend services, such as the Hadoop processing system, shall be Unix.

Priority: **Essential**.

Risk: **Safe**

Effort: **Days**
Setting up the software.

Realisation: *BC will install Ubuntu Server Linux on all machines.*

5.3.3 Demonstration to the Community

R-NF-3.1 The requirements for processing and validation tests shall be derived from Coastcolour and CCI-OC requirements shall be taken into account. The ideal outcome of Calvados should be to have a demonstration system in place in one year time (~ July 2011), that should then be used and further developed by the CCI-OC and other CalVal users.

Priority: **Essential**.

Risk: **Fair**
The prototype's hardware set-up might not be sufficient to really demonstrate a hi-end runtime performance

Effort: **Months**

R-NF-3.2 BC shall keep a close link with the international ocean colour communities, through CoastColour and ECV-OC, and to proof the usefulness of the demonstration system by feedback from these groups. To the communities, the project shall be communicated as a technology study.

Priority: **Essential**. **Risk:** **Safe**
