ESA Climate Change Initiative
aerosol_cci

Deliverable D3.1c

System Specification Document issue 3

Version 1.2
Document reference: AEROSOL_CCI_SSD3_v1.2
### DOCUMENT STATUS SHEET

<table>
<thead>
<tr>
<th>FUNCTION</th>
<th>NAME</th>
<th>DATE</th>
<th>SIGNATURE</th>
</tr>
</thead>
<tbody>
<tr>
<td>LEAD AUTHOR</td>
<td>editor</td>
<td>Julian Meyer-Arnek</td>
<td>08.12.2015</td>
</tr>
</tbody>
</table>
| CONTRIBUTING AUTHORS| T. Popp  
Aerosol_cci Phase 2 System Engineering Working Group |           |           |
| REVIEWED BY         | Science leader                                             | T. Popp   | 09.12.2015 |
| APPROVED BY         | Technical officer (ESA)                                    |           |           |
| ISSUED BY           | Project manager                                            |           |           |
EXECUTIVE SUMMARY

Within this document a system specification of the operational Aerosol_CCI system is presented. This document describes an over-all system for the creation and repeated reprocessing of new satellite data based aerosol Essential Climate Variables (time series), which are accepted by the scientific community and used for IPCC assessments. It is based on the previously generated System Requirements Document which is deduced from the user requirements, product specifications and data access requirements. The scope of the underlying requirements extends over the next 15 years.

The system concerned here comprises a technical implementation to allow and support large-scale (re-)processing and analysis of the aerosol ECVs and a management mechanism to ensure the availability of scientific expertise for continuous algorithm improvement and product validation. In order to ensure the highest possible quality and acceptance in the community the overall principle for the whole system is the leadership by scientific expertise.

The described decentralised operational Aerosol retrieval processing system is now operated in the scope of the Copernicus Climate Change Service (C3S) 312a/Lot 5 (Aerosols).

After the introduction (Section 1) this document starts with a definition of the purpose of the system (Section 2) to specify subsequently its context (Section 3). Tasks of the Science Team are specified in chapter 4. In chapter 5, the operational scenarios for the ECV production system are discussed. Chapter 6 performs a trade-off analysis of different system design approaches (hardware characteristics and centralised vs. distributed approach). Chapter 7 describes the “Environmental Characteristics of the System” and chapter 8 its constraints. The concrete implementation of the system is given in chapter 9 “System Maintainance”. Part of this chapter is the “Description of the decentralized processing chains” operated by the partners. An analysis whether the existing documentation matched requirements of the ECSS is performed in Section 10. A matching matrix with the system requirements is made in Section 12 to assure traceability.
<table>
<thead>
<tr>
<th>Issue</th>
<th>Date</th>
<th>Modified Items / Reason for Change</th>
</tr>
</thead>
<tbody>
<tr>
<td>SSD0 - 1.0</td>
<td>17.07.2011</td>
<td>First Issue</td>
</tr>
<tr>
<td>SSD0 - 1.1</td>
<td>23.09.2011</td>
<td>Completely revised to correspond with System Requirements Document V1.6</td>
</tr>
<tr>
<td>SSD0 - 1.2</td>
<td>28.09.2011</td>
<td>Review by science leader; improved executive summary and section 2 (purpose of the system) and check of entire document; addition of organisation of science team and initial step of trade-off between hardware alternatives.</td>
</tr>
</tbody>
</table>
| SSD0 - 1.3 | 18.11.2011 | Revision of the document as an answers to Aerosol_CCI D5.2 SSD0 v1.2 RIDs  
Reviewed by co-science leader: adjustment of reference and applicable documents; final check with only minor corrections |
| SSD0 - 1.4 | 16.02.2012 | Re-working of the document by co-science leader and system experts based on new RIDs received for v1.3 aiming at a higher level point of view  
Added list of acronyms, cleaned up references  
Generally extended discussion and trade-off analysis  
Sec. 3: revised fig. 3.1-1 and associated text, archive discussion extended, interface discussion extended, updated fig. 3.4-1 (now fig 4.1-1), discuss intermediate products, several smaller corrections or additions  
Sec. 3.4 and 3.5 moved to new sections 4.1 and 4.2  
New sec. 5 moved from sec. 4.6; advisory board and science team merged into one, several smaller corrections and addition to science team discussion  
Sec. 6 (former 4): extended discussion on merging, production control, workflow management, revised discussion on network bandwidth estimation, several smaller corrections and additions, new fig. 6.2-1 added, new sec. 6.6 on interface to algorithm development added  
Sec. 7 (former 5): two new scenarios added and other scenarios written more generally at Aerosol_cci top level, dissemination discussion extended, several smaller corrections and additions  
Sec. 9 (former 7): largely re-written and trade-off analysis in sec. 9.3 much extended, sec. 7.1 on constraints extended and moved to new sec. 3.4, several smaller corrections and additions, first cost breakdown added (new sec. 9.4)  
New sec. 11 (funding) and requirement matching matrix (12) added |
## System Specification
### Document issue 3

<table>
<thead>
<tr>
<th>Issue</th>
<th>Date</th>
<th>Modified Items / Reason for Change</th>
</tr>
</thead>
<tbody>
<tr>
<td>SSD0 - 1.5</td>
<td>09.03.2012</td>
<td>Revision based on RIDs for v1.4: Corrections to fig. 3.1.1, sec 3, fig 4.1.1, sec 4, sec. 6 sec. 7, sec. 9</td>
</tr>
<tr>
<td>SSD1 - 1.0</td>
<td>06.08.2012</td>
<td>V1 of the document. Added new section 11 including trade off analysis for processing hardware organisation and processing distribution.</td>
</tr>
<tr>
<td>SSD1 - 1.1</td>
<td>07.08.2012</td>
<td>Review by science leader, formal version and checking all intended extensions</td>
</tr>
</tbody>
</table>
| SSD1 – 1.2| 12.09.2012| Revision based on RIDs for SSD1 – 1.1:  
  - 11.2.5: Addition of an evaluation  
  - 11.2.3: Comparison between cloud and own hardware cost  
  - 12: reference to sec. 9.4 and total cost included |
| SSD1 – 1.2|           | Science leader rewording to clarify content of addition in 11.2.3 |
| SSD2 - 1.0| 16.01.2015| New Release for Aerosol_cci Phase 2  
  Modifications:  
  - Chapter 5 “Matching Aerosol_cci documentation to ECSS documentation standard” added  
  - Re-numbering of previos chapter 5-13  
  - Chapter 6 “Science Team”: “Concrete implementation (Aerosol_cci Phase 2, year 1)” added.  
  - Chapter 9.4 “Concrete Measures (System Maintenance, Year 1)” added  
  - Chapter 9.5 “Software Standards (System Maintenance, Year 1)” added  
  - Chapter 10.5 “Concrete Implementations for Phase 2 of Aerosol_cci (Year 1)” added  
  - Chapter 11.3 “Concrete Implementation Phase 2 (Year 1)” (regarding “Hardware Considerations”) added  
  - Requirements matching matrix (chapter 14) updated |
<table>
<thead>
<tr>
<th>Issue</th>
<th>Date</th>
<th>Modified Items / Reason for Change</th>
</tr>
</thead>
<tbody>
<tr>
<td>SSD3 – 1.0</td>
<td>14.01.2015</td>
<td>New release of the document with major modifications. Nearly all chapters have been altered. Among others, the following major modifications have been performed:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Former chapters 4 (“The operational ECV production SYSTEM”) and 7 (“Fundamental functions of the ECV production system”) have been combined, shortened and streamlined.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- The former “Architecture Analysis” has been shortened and has become part of chapter 6 “Trade off analysis for Operational System Design”.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Chapter “System Maintainance” now contains the “Description of decentralised processing chains” provided by the Aerosol_cci 2 partners. This chapter was identified as an requirement to become compatible with ECSS quality standards.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Requirement matching matrix updated. Moreover, the order of all chapters have been re-arranged to foster the readability of the document.</td>
</tr>
<tr>
<td>SSD3 – 1.1</td>
<td>30.04.2017</td>
<td>Minor updates of the document version 1.0. Updates are including the transition of the decentralized Aerosol_CCI2 processing system into the operational Copernicus Climate Change Service (C3S) architecture. In section 3 (“Context of the System”), the evolution of the Aerosol_CCI2 architecture into the C3S architecture is explained.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Section 9 (“System Maintainance”) reflects the current status of the decentralized Aerosol processing sites. Updates have been provided by the Aerosol processing partners.</td>
</tr>
<tr>
<td>SSD – 1.2</td>
<td>28.06.2017</td>
<td>Minor corrections of the previous version based on the Review Item Discrepancies as raised by project officer Simon Pinnock.</td>
</tr>
</tbody>
</table>
LIST OF FIGURES

Figure 3-1: Distributed Aerosol_cci system and its context (adapted from [RD4])..........................14
Figure 3-2: Operational distributed Aerosol_cci system and its context within Copernicus Climate
Change Service (C3S)..................................................................................................................16
Figure 5-1: Internal Interfaces of operational ECV Production System. See section 5.4 for detailed
measurement...............................................................................................................................26
Figure 5-2: Web-interface for access to FTP server. Service provided by icare (http://www.icare.univ-
lille1.fr/drupal/archive/?dir=CCI-Aerosols). ...........................................................................28
Figure 6-1: Processing Capacity usage of an operational ECV production system .......................34
Figure 6-2: Evaluation of different hardware scenarios ................................................................39
Figure 6-3: Service oriented deployment of an operational system .............................................40
Figure 6-4: Orchestration of a production task ............................................................................41
Figure 9-1: Software development cycle as proposed for this project .......................................52
Figure 10-1: Branches and disciplines of the ECSS standard. The E-40-discipline (Software
Engineering) which is most relevant for Aerosol_cci is highlighted. ......................................71
Figure 10-2: Processes covered by ECSS E-ST-40 and related standards ....................................72
Figure 10-3: Structure of the ECSS E-40 discipline (taken from [RD5], p. 31). Numbering refers to
ECSS-E-ST-40C document. ..........................................................................................................72
Figure 10-4: Overview of the ECSS-software lifecycle process. Numbering refers to the [RD5].
Viewgraph was taken from [RD5] and partly modified. ..............................................................74
Figure 10-5: Overview of the ECSS software lifecycle process. Aerosol_cci-specific components are
indicated in dark blue. Numbering refers to [RD5].................................................................75
Figure 10-6: Document requirements matrix. Matches between Aerosol_cci Phase 2 documentation
and ECSS are indicated with “X”..............................................................................................76
TABLE OF CONTENTS

DOCUMENT STATUS SHEET ........................................................................................................ 2
EXECUTIVE SUMMARY ............................................................................................................. 3
DOCUMENT CHANGE RECORD ................................................................................................. 4
LIST OF FIGURES ....................................................................................................................... 7

1 INTRODUCTION ...................................................................................................................... 11
   1.1 Scope ............................................................................................................................... 11
   1.2 References ....................................................................................................................... 11
      1.2.2 Applicable Documents ............................................................................................... 11
      1.2.3 Reference Documents ............................................................................................... 11
      1.2.4 Acronyms ................................................................................................................ 12

2 PURPOSE OF THE SYSTEM .................................................................................................... 13

3 CONTEXT OF THE SYSTEM .................................................................................................. 14
   3.1 System Overview .............................................................................................................. 14
   3.2 Interfaces ......................................................................................................................... 17

4 TASKS OF THE SCIENCE TEAM .......................................................................................... 19
   4.1 Fulfillment of the tasks .................................................................................................... 19
   4.2 Concrete implementation for Aerosol_cci Phase 2, year 1 and 2 ................................... 21
   4.3 Prioritising and Handling Updates .................................................................................. 21
      4.3.1 Scenario Agreement on priorities for algorithm development ............................... 22
      4.3.2 Scenario Prototyping and Implementation of new algorithm or module .............. 24
      4.3.3 Scenario Addition_Of_New_Input_Data ................................................................... 25

5 OPERATIONAL SCENARIOS FOR ECV PRODUCTION SYSTEM ........................................ 26
   5.1 Interfaces within the operational ECV production system .............................................. 26
   5.2 User Interface .................................................................................................................... 26
   5.3 Dissemination .................................................................................................................... 27
   5.4 Processing ......................................................................................................................... 29
   5.5 Quality Control ................................................................................................................ 30
   5.6 Validation and Intercomparison ....................................................................................... 31
   5.7 Archiving .......................................................................................................................... 31
   5.8 Valid ranges of values, timing considerations and data sources .................................. 33

6 TRADE OFF ANALYSIS FOR OPERATIONAL SYSTEM DESIGN ............................. 34
6.1 CCI Project Characteristics ........................................................................................................ 34
6.2 Hardware usage strategies .......................................................................................................... 35
  6.2.1 GPU Usage .............................................................................................................................. 36
  6.2.2 Grid .......................................................................................................................................... 36
  6.2.3 Cloud ....................................................................................................................................... 36
  6.2.4 Cluster Computing .................................................................................................................. 38
  6.2.5 Summary ................................................................................................................................ 38
6.3 System deployment: Centralised vs. distributed system .............................................................. 39
  6.3.2 Monolithic system ................................................................................................................... 39
  6.3.1 Centralised approach ............................................................................................................... 39
  6.3.2 Distributed service oriented system ....................................................................................... 40
  6.3.3 Fixed distribution with focus on input data ........................................................................... 41
  6.3.4 Conclusion .............................................................................................................................. 42
6.4 Associated costs ............................................................................................................................ 44
  6.4.2 Centralised model ................................................................................................................... 44
  6.4.3 Decentralised model ............................................................................................................... 45
  6.4.4 Conclusion .............................................................................................................................. 46
7  ENVIRONMENTAL CHARACTERISTICS OF THE SYSTEM ................................................. 48
  7.1 Concrete Implementation of Aerosol_cci 2 (Year 2) ...................................................................... 48
  7.2 Initial Processing Hardware dimensioning .................................................................................. 50
  7.3 Initial Archive dimensioning ....................................................................................................... 50
8  SYSTEM CONSTRAINTS .................................................................................................................. 51
9  SYSTEM MAINTAINANCE .............................................................................................................. 52
  9.1 Overall system aspects: Software .............................................................................................. 52
  9.2 Hardware life cycle ...................................................................................................................... 53
  9.3 Description of decentralised processing chains ........................................................................... 53
    9.3.1 BIRA-IASB: AerGom / GOMOS .............................................................................................. 53
    9.3.2 BIRA-IASB: IASI ..................................................................................................................... 55
    9.3.3 DLR: ATSR L3 ensemble ........................................................................................................ 57
    9.3.4 DLR: IASI ............................................................................................................................... 57
    9.3.5 DLR: SYNAER ......................................................................................................................... 59
    9.3.6 FMI: ADV ((A)ATSR) ............................................................................................................. 60
    9.3.7 FMI: SLSTR ............................................................................................................................. 61
    9.3.8 LOA/ICARE: GRASP ............................................................................................................. 61
    9.3.9 LMD: IASI ............................................................................................................................... 62
    9.3.10 RAL: (A)TSR (ORAC) ........................................................................................................... 63
    9.3.11 SU: AATSR ............................................................................................................................ 65
    9.3.12 SU: SLSTR ............................................................................................................................. 66
    9.3.13 ULB: IASI .............................................................................................................................. 67
  9.4 Growing Amount of Input Data .................................................................................................. 68
  9.5 Availability of more consistent ancillary data .............................................................................. 69
10 MATCHING AEROSOL_CCI DOCUMENTATION TO ECSS ................................................. 71
11 FUNDING MECHANISMS........................................................................................................ 78

12 REQUIREMENT MATCHING MATRIX.................................................................................. 79
1 INTRODUCTION

1.1 Scope

The System Specification Document incorporates the requirements described in the System Requirements Document and specifies the characteristics of an operational ECV production system from a developer’s point of view. Each requirement of the System Requirements Document is tackled in the appropriate part of this document to show the system design answer to the requirement.

1.2 References

1.2.2 Applicable Documents

1.2.3 Reference Documents

1.2.4 Acronyms

The following list contains only system relevant acronyms - no satellite sensor or algorithm acronyms.

ADP Algorithm Development Plan
AEROCOM AERosol mOdell inter-COMparison initiative
ATBD Algorithm Theoretical Base Document
CCI Climate Change Initiative
DARD Data Access Requirements Document
ECSS European Cooperation for Space Standardization
ECV Essential Climate Variable
GCOS Global Climate Observing System
GPU Graphical Processing Unit
I/O Input / Output
ICARE Thematic centre of CNES, CNRS, Univ. of Lille, holds satellite aerosol data
Level1 calibrated satellite measurements in orbit projection
Level2 derived satellite geophysical parameters in orbit projection
Level3 geophysical parameters in gridded product
PSD Product Specification Document
PVIR Product Validation and Inter-comparison report
SRD System Requirements Document
SSD System Specification Document
URD User Requirements Document
WAN Wide area network
WMO World Meteorological Organisation
2 PURPOSE OF THE SYSTEM

The purpose of this operational ECV System is to produce and regularly reprocess satellite-based Essential Climate Variable time series of highest quality and credibility. The system specified in this document is based on a System Requirements analysis [RD4], which was derived from the User Requirements [RD2] and Product Specifications [RD3] of the Aerosol_cci project with its horizon encompassing the next 15 years (SR-OP-0320). It also includes considerations regarding a unified ECV production system not only for the Aerosol CCI Project but also for other CCI projects as well as consideration for different processing architectures. At this stage, data retrieved by ESA and other European satellite missions serve as a source for aerosol ECV products. As a starting point these comprise the instruments (A)ATSR(-2), GOMOS, MERIS, OMI, POLDER, AVHRR (METOP), GOME, GOME-2 and SCIAMACHY, but in the next one and a half decades further instruments onboard the various Sentinel and METOP platforms as well as recent and next generation instruments also on geostationary platforms shall be taken into account. In order to fulfil requirements SR-GE-0020 and SR-GE-0030 the system must be sized and remain scalable to allow adjustment to new input datasets and to allow substantial re-processing whenever a new product becomes available.

The operational Aerosol_cci ECV production system intends to provide an infrastructure for the creation of several aerosol_CCI product types. The product types are described within [RD3]. The system uses a variety of input products specified in [RD1] to extract the desired parameters – the Essential Climate Variables (ECVs). These parameters are currently the multi-spectral Aerosol Optical (AOD) with derived products, the Absorption Aerosol Index (AAI) and stratospheric extinction profiles (SR-GE-0010).

The operational Aerosol_cci ECV production system consists of two major components: a technical implementation to allow the required production and reprocessing of the datasets, and a coordination mechanism to ensure that the Aerosol_cci system is under permanent leadership of scientific experts. The close interaction of these two essential components assures the production of highest quality and credible products, which are accepted by the science community (SR-GE-0050).

Therefore the system is designed to fulfil the general requirements and produce data defined in the regularly updated Product Specification Document - PSD (SR-GE-0010) from data specified in the regularly updated Data Access Requirements Document - DARD (SR-GE-0020). In order to achieve timely production of ECV products the initial version of the system must rely on existing algorithms and their gradual upgrades (SR-GE-0090). The system shall be flexible in a way that enables the incorporation of data from a new sensor as a new data source (SR-FU-0141).

The concrete realization of the final Aerosol_cci system and its successor - the future operational C3S_312a/Lot5 production and archive system – as based on the experiences and efforts taken during all project stages of Aerosol_cci. All these experiences are collected and discussed in this system specification document.The decentralized Aerosol_cci retrieval und production system is now transferred to the Copernicus Climate Change Service (C3S) project 312a Lot 5 (Aerosol ECV production). Moreover, the system is extended to new Sentinel instruments delivering information on Aerosols.
3 CONTEXT OF THE SYSTEM

3.1 System Overview

In Figure 3-1 the Aerosol_cci system is depicted with all internal and external elements (within the blue box). Elements which could relatively easily be shared with other ECV operational systems in order to ensure consistency or user-friendliness are indicated as boxes in dashed lines at the edges of the blue box. Within the Aerosol_cci system box technical entities are depicted in blue, and coordination elements are depicted in black. The Figure 3-1 is adapted from [RD4].

Figure 3-1: Distributed Aerosol_cci system and its context (adapted from [RD4])

The operational distributed ECV production system consists firstly of hardware for the fast, complete and repeated reprocessing of large datasets. Secondly, it includes archiving functionality for the input data (satellite and ancillary datasets), for the Aerosol ECV products and for reference datasets used for validation and inter-comparison (ground-based datasets and external aerosol ECV datasets). Furthermore, it includes standard tools to support repeated testing and validation of new algorithms (SR-GE-0040) and capabilities for the dissemination of products and documentation to the users (SR-GE-0130). The distributed system is able to handle all relevant input data together with
required external reference data. A Science Team assures scientific fitness for purpose of the aerosol ECV products. Based on the joint expertise of climate users, validation and inter-comparison experts and algorithm developers its key responsibility is the definition of algorithm development priorities to steer the operational product evolution (SR-GE-0041). To fulfil this responsibility the science team coordinates external links with novel science, climate users, instrument teams and space agencies. A continuous algorithm development task implements the priorities defined by the science team in order to ensure product evolution to better meet end user needs and responding to their requests for improvements (SR-GE-0080). Validation and quality control for each new ECV dataset determines the error characteristics by comparison to external reference datasets (ground-based). In order to document the competitiveness or complementarity of the new ECV products, their statistical error characteristics must be related to other state of the art aerosol satellite datasets by inter-comparison with them (SR-DO-0160). In order to assure credibility, validation of the products and the uncertainty estimates provided as part of the products is conducted by experts who are independent from the algorithm developers (in line with SR-GE-0042). The requirements for the different actors and interactions within this system are described in more detail in [RD4].

Besides the technical elements purely dedicated to the automated creation of ECV datasets strong organisational interfaces and interactions are implemented to guarantee quality and credibility of the created products (SR-GE-0041) as well as give the user community the opportunity to influence development processes and processing priorities (SR-MA-230).

The project team is convinced that the best coordination mechanism to ensure the scientific quality of the ECV products is guaranteed through a science team with dedicated expertise for aerosol retrieval algorithms, validation and use of aerosol ECV products. This therefore is an answer to the requirements for both, internal coordination (SR-GE-0100) and external coordination (SR-GE-0110). As the science team organises the coordination mechanisms themselves these mechanisms can be flexible (SR-GE-0120). The science team also coordinates communication with other projects to maximise consistency of ancillary data used by other projects (SR-FU-0150). This is also the most efficient way to assure swift and continuous implementation of cutting edge research results such as new algorithms and upgrades (SR-QU-0180).

Due to this nature of the system we have to define the system interfaces in three layers:

- The systems external interfaces, where system means the whole structure depicted within the blue box in Figure 3-1.

- The internal interfaces within the system between the operational ECV production system and the other (coordination) elements within the blue box in Figure 3-1

- The interfaces within the operational ECV production system

In the scope of the Copernicus Climate Change Service (C3S) project 312a/Lot 5, the decentralized Aerosol retrieval and processing system is further developed. A central “triangle” consisting of FMI, ICARE and DLR will ensure a persistent machine-to-machine-interface towards the Copernicus Climate Data Store (CDS).

The decentralized processing partners will push their results to one of the central partners (DLR, FMI or ICARE). This includes all necessary metadata. Metadata is then provided to the CDS via open interfaces (such as OGC’s CSW) to allow the CDS building up a metadata catalogue.
Datasets will reside either at DLR (with will operate the central node) or at FMI or ICARE. Once a CDS-user wants to download a certain dataset, the CDS will contact DLR. If the dataset is not stored at DLR, the CDS-request is forwarded to the data holder transparently.

This planned setup is essentially an evolution of the Aerosol_CCI2 setup.

![Diagram](https://example.com/diagram.png)

**Figure 3-2: Operational distributed Aerosol_cci system and its context within Copernicus Climate Change Service (C3S).**

The next sub section briefly mentions these external and internal interfaces (3.2), before system constraints are summarized (see section8). Interfaces within the operational ECV production system and the operational ECV processing are discussed further in section 5.1. The science team is then discussed in section 4.
3.2 Interfaces

External ECV users

This is the most important external interface since it provides the main dissemination of the Aerosol ECVs (External ECV Users downloading ECV data products and documentation from the archive and interacting with any online tools provided by the archive) and the key feedback regarding the quality and possible utilizations of the produced ECV data to the science team (SR-OP-0460). This input is then translated into changed development priorities, which are documented by the science team in an updated user requirement document, an updated product specification document and an updated data access requirement document and released regularly (SR-DO-0375, SR-DO-0376, SR-DO-0377). In addition to this the science team shall also publish a collection of the utilisation made with the ECV products regularly (SR-DO-0379). In addition, external ECV users can request changes within algorithms to the Science Team (SR-MA-0230). The Science team converts these inputs into changes of the development priorities.

In the context of Copernicus Climate Change Service (C3S), the Climate Data Store becomes a major external user of Aerosol_CCI2 datasets. The OPEnDAP interface which will be operated at DLR will be tailored to meet all relevant requirements arising from the Copernicus Climate Change Service.

Science team

Within the Aerosol_cci system the science team is responsible for the algorithm development. This means it converts the validation and user feedback into evaluated development priorities (SR-OP-0480) regarding algorithm development. To achieve this goal and thus to fulfil SR-GE-0042 (scientific credibility) and SR-GE-0050 (science team composition) the science team has to be composed of algorithm experts, independent validation experts (SR-OP-0470) and ECV use experts.

Another internal interface of the science team is the external reference input needed for repeated quality checking, validation and inter-comparison of ECV products and thus fulfill requirement SR-GE-0040. All documents issued by the science team describing the input, the products, their validation and their utilization are stored and distributed.

Algorithm development

The algorithm development function of the Aerosol_cci system has to interface with the operational ECV production system to integrate new algorithm implementations into the operational ECV production system. The major benefit is that algorithm development can focus on the core retrieval issues since it can fully rely on the operational production system for handling of all input and output datasets, formats, metadata, etc.

The algorithm development task receives its priorities from the science team which has iterated these between user needs, validation results and algorithm possibilities. Algorithm development is then conducted by EO experts within the science team.

Validation / QA

Supervision and interpretation of the statistical analysis performed for validation is conducted by the validation experts in the science team and is supported by standard tools. The output of validation and inter-comparison exercises is delivered to the entire science team.
The validation task has interfaces to the archive (access the ECV products and reference datasets), to the operational ECV production system (standard tools for statistical validation), and to the science team (delivery of validation reports and recommendations).
4 TASKS OF THE SCIENCE TEAM

In Aerosol_cci, a science team was established with the main aim to

Assure scientific leadership of ECV development and processing in order to achieve highest quality and credibility of aerosol ECVs in the climate scientific community.

4.1 Fulfillment of the tasks

In order to fulfil the functions described in [RD4] the science team uses an efficient but comprehensive means of work. We consider as best suited an organisation with

- a science team leader,
- a flexible team of scientific experts to conduct scientific-technical work,
- space agency representatives as observers

The science team leader (part time to assure personal involvement with related scientific work and rotated among the leading members of the science community) manages all practical tasks conducted by the science team members (coordination between scientific experts and the operational system, coordination and quality assurance of documentation, preparation of decisions, representation at relevant user communities, interaction with space agencies, cal/val teams, international retrieval teams and other ECVs).

The science team assembles several times a year. For the time of their participation science team members concentrate on the relevant tasks (e.g. for the life cycle of their algorithm or validation source). The entire science team works as one team and has to endorse major decisions (algorithm development priorities, ECV validation acceptance). The science team must also include international experts.

To conduct scientific work, the science team must comprise of experts from three sub domains focusing on algorithm development, product validation and user interaction.

- Algorithm experts contribute on case-by-case basis (for dedicated limited time and specific task) regarding the priority development plan endorsed by the science team.
- Validation experts conduct validation and inter-comparison tasks on a long-term basis to ensure stability of quality control and reference datasets (possibly partly replaced every few years).
- User interaction includes updating user requirements every two years and test utilization and endorsement of new ECV products in conjunction with climate models and / or in collaboration with other ECV teams.

The science team bases its work on the following terms of reference:

1. To define algorithm development priorities the science team shall prepare an annual algorithm development (priority) plan based on newly arising user requirements (summarized
into an updated user requirements document by the user experts), latest ECV product validation and assessment results (summarized in validation and assessment reports by the validation and user experts) and proposed new or improved algorithms (new or updated algorithm theoretical baseline documents prepared by the algorithm experts or external retrieval scientists) after consultation with the entire science team and of recent scientific literature. This priority plan is discussed, altered where deemed necessary and endorsed by the whole science team at an annual review meeting and implemented afterwards by the science team. It is the overarching goal of this task to base algorithm development priorities on consensus within the international aerosol retrieval community – however focusing on European instruments. In order to meet SR-OP-0170, a schedule on the availability of ECVs must be published (SR-OP-0170). This activity intends to foster the ECV’s use.

**To formally approve the quality level of new ECV datasets** the science team will conduct validation of the new datasets by the related experts, who can rely on automatic tools for extraction of validation statistics to derive and judge the respective validation (validation report). Additionally test utilization may be organized. The science team will review at its annual review meetings these validation and assessment reports for any new ECV dataset and prepare a release note for this new dataset which summarises in condensed manner (typically 1 page) the overall validation results, the major use demonstrated, the added value over earlier datasets and recommendations and limitations for the use of the dataset. Experts for product validation and inter-comparison shall be consulted to validate ECV product quality (SR-OP-0470).

2. **To assure the use of state-of-the-art algorithms, input data and auxiliary data** the science team leader and the scientific experts shall rely on several means

   i. be involved in aerosol satellite retrieval activities which will allow them to regularly screen novel related science

   ii. actively participate to relevant conferences and organize specific workshops to discuss potentially beneficial improvements of algorithms and ECV products

   iii. actively contribute to international coordination (retrieval and climate science communities, space agencies, international scientific programs and bodies such as AEROCOM, GCOS, WMO, etc.) via bi-lateral exchange and conference participation

   iv. to assure the contribution of Aerosol cci to the international scientific dialogue by maintaining the communication mechanisms (web portal, newsletter, training courses and materials, etc.)

   v. pro-actively seek continuous flow of knowledge with instrument cal/val teams to assure proper understanding of calibration uncertainties and their improvements

   vi. collaborate with validation data originators (in situ, airborne) in order to understand their limitations and assure that all necessary external reference data are available to allow proper validation and improvement of algorithm parameterisations
vii. collaborate with the aerosol climate model community to promote the use and assure proper use of aerosol ECV datasets and obtain feedback on their added value, limitations and required improvements.

viii. use the relevant information and recommendations for updating the underlying specification documents (data access requirements, product specifications) which will then feed into the algorithm development priority plan with one particular focus on identifying gaps and possible solutions (e.g. new parameters, regional, …)

ix. To contribute to definition of mission requirements in order to assure the long-term continuity of (European) aerosol ECV measurements

x. To promote open data access for any input and auxiliary datasets needed and negotiate data access where necessary

xi. To act as an interface to the outside world for open and transparent communication of issues and potential solutions in order to gradually reduce uncertainties and errors of the ECV datasets

3. **To lead the operational system by scientific needs** the science team leader consulted by the experts will define any necessary upgrades of hardware, technical software for ECV processing, tools for validation, dissemination, and visualisation and with regard to new algorithms or input data.

### 4.2 Concrete implementation for Aerosol_cci

During the ESA-funded Aerosol_cci, Phase 2 project (2014-2017), the science team as defined in this section is realized by the project consortium partners. Annual review meetings backed up with the associated deliverable documents (user requirements, validation and assessment reports, annual algorithm development report and plan, …) provide the concrete implementation of the priority decisions. The project science leaders act as science team leaders who organize the internal and external coordination defined for the science team. The international embedding of the project-internal science team is achieved by AEROSAT, the open International Satellite Aerosol Science Network, which is co-chaired by one of the Aerosol_cci science leaders and a NASA colleague. AEROSAT holds annual workshops and disseminates its discussions and conclusions through its website (aerosat.org). Through their leading activities in AEROSAT Aerosol_cci team members take part in the international discussions on retrieval improvements and new developments and they expose the Aerosol_cci results to feedback in this forum. Where appropriate, working groups of Aerosol_cci act jointly with AEROSAT working groups or other CCI project groups. Through participation to science conferences and the organisation of user workshops Aerosol_cci strengthens up its feedback loops.

### 4.3 Prioritising and Handling Updates

The operational system has to be able to implement new or revised algorithms (or modules) following the agreement on algorithm development priorities coordinated within the science team between end users, validation experts and algorithm experts.
In order to facilitate this implementation the algorithm experts will have access to (modular) algorithm code implemented into the operational system so that they can easily make their changes and the subsequent re-implementation can be facilitated easily. Before re-implementation the science team conducts sensitivity studies, algorithm experiments and validation studies with a limited amount of data outside the operational system. Validation and inter-comparison experts shall have the possibility to request reprocessing of a specific time series of a ECV product to validate results of an algorithmic change (SR-OP-0400).

Following the first implementation verification within the operational system of a data amount agreed to be sufficient within the science team has to be made.

One of the main requirements to the aerosol_cci system as well as to all other CCI systems is a strong link to the science community. Therefore, the operational scenarios include scenarios for changes in algorithms or addition of new algorithms, and the addition of new input data.

### 4.3.1 Scenario Agreement on priorities for algorithm development

The process of identification of priorities for algorithm development is a key necessity to ensure cutting edge research addressing the user needs (SR-MA-0240, SR-MA-0250). These needs are rapidly identified and implemented, to focus science community algorithm development expertise on ECV requirements, and to justify funding for algorithm development – this implements SR-QU-0180.

To support an easy exchangeability of processing algorithms the system has to be modular in a way, which allows the exchange of separate modules in a processing algorithm – this is also an answer to SR-GE-0080, SR-OP-0420 and SR-MA-0250.

Starting from the validation of existing ECV products disseminated to the user community user feedback is obtained on existing ECV products, their validation results, and new ECV product requirements from the wider user community by the Science Team User Experts. On the basis of this user feedback, the Science Team Algorithm Experts identify potentially feasible areas of algorithm development, based on *e.g.* a knowledge of the potential information content of the satellite sensor data, known algorithm weaknesses, *etc.* (*e.g.* opportunity to use sun glint for retrieval of absorption properties, use of IR channels for better characterisation of desert dust, improvement of AATSR dual view collocation for improved retrieval performance, improvements to uncertainty characterisation, improvements to cloud masking, *etc.*). The Science Team produces an annual report on algorithm development priorities, including requirements for improved space agency (and ancillary data) products. This annual report is then disseminated to Space Agencies (formal request and prioritisation of improvements to standard products) and the EO science community (help direct and justify proposals for algorithm development funding). The Science Team then reviews new algorithm developments appearing in peer reviewed publications, and selects those which address documented high priority user needs and performs a trade-off analysis to identify the most promising or urgent algorithm upgrades, taking into account their potential impact, technical maturity, feasibility and cost. Finally, the Science Team recommends these priorities for implementation – this implements the requirement SR-DO-374 for a priority development plan based on broad scientific consensus.
<table>
<thead>
<tr>
<th>Scenario Name:</th>
<th>Agreement on priorities for algorithm development</th>
</tr>
</thead>
<tbody>
<tr>
<td>Start up Preconditions</td>
<td>New ECV requirement or new algorithm or module available</td>
</tr>
</tbody>
</table>
| Actors | Space_agencies, EO_science_community  
Science_team in all its capacities:  
Climate_research_ECV_users,  
Algorithm_experts,  
Validation_and_intercomparison_experts |

Science_team Validation_and_intercomparison_experts perform validation of existing ECV products and provide these results to the user community  
Deliverable: Product Validation and Intercomparison Report (PVIR)  

User feedback on existing ECV products, their validation results, and new ECV product requirements is collected from the wider user community by the Science_team Climate_research_ECV_users  
Deliverable: User Requirements Document (URD)  

The Science_team Algorithm_experts identify potentially feasible areas of algorithm development.  
Input: URD, knowledge of the potential information content of the satellite sensor data, known algorithm weaknesses  

Science_team produces an annual report on algorithm development priorities, including requirements for improved space agency (and ancillary data) products  
Deliverable: Algorithm Development Plan (ADP) – chapter 1.  

Annual report is disseminated to:  
(a) Space_agencies to formally request and prioritise improvements to standard products,  
(b) EO_science_community to help direct and justify proposals for algorithm development funding.  

The Science_team annually reviews new algorithm developments appearing in peer reviewed publications, and selects those which address documented high priority user needs.  

The Science_team performs a trade-off analysis to identify the most promising or urgent algorithm upgrades.  
Criteria applied to judge proposed algorithm upgrades in the trade-off analysis: potential impact, technical maturity, feasibility and cost.  
Deliverable: Algorithm Development Plan (ADP) - chapter 2  

In conclusion the Science_team recommends these priorities for implementation. For each identified change, upgrade or new development task the Science_team states a specified goal (e.g. improved accuracy, coverage, new output variable, …)  
Deliverable: Algorithm Development Plan (ADP) - chapter 3.
4.3.2 Scenario Prototyping and Implementation of new algorithm or module

Once an algorithm development plan has been worked out in the previous scenario and the funding for the activity has been secured the science team members (possibly extended by new relevant experts needed in this work) conduct all necessary steps to prototype and implement the new algorithm(s) or module(s). The algorithm experts start with coding the new algorithm based on its underlying scientific publication and scientifically testing it with a small amount of data. This step includes acquisition of any needed new ancillary data or validation reference datasets. Once this scientific case study code is regarded ready it is implemented into the ECV production system by the processor experts together with support by the algorithm experts as new prototype (separated from the existing operational algorithms) and used for production of a substantial data amount which allows statistically significant evaluation. Then the automatic standard tools for statistical analysis are invoked to prepare the statistical analysis for validation to ground-based reference datasets and for inter-comparison to other datasets (satellites, models). On the basis of this validation output the science team validation and inter-comparison experts review the results and the entire science team draws a final conclusion which takes into account statistical validation / inter-comparison quantities, coverage achieved and spatial and temporal correlation with observed aerosol features. By comparing the new products with the existing operational products the increment is assessed and compared to the underlying user requirements and the specified improvement goal which had been defined in the algorithm development plan - this is also an answer to SR-DO-0160. When the new product satisfies the specified goal its prototype implementation is switched free for operational production and the respective (re-)processing and subsequent validation of the new ECV dataset are initiated. The sequence of these steps has to be described within a manual (SR-DO-0440).
Scenario Name: Prototyping and implementation of new algorithms or modules

Start up Preconditions
Science_team has agreed on new algorithm or module to be developed and implemented
Funding for this activity is secured

Actors
Science_team: Algorithm_experts, Validation_and_intercomparison_experts, Climate_research_ECV_users
Operational_ECV_production_system, Processor_experts

The Algorithm_experts code the new algorithm and test it scientifically with a small amount of data. Acquisition of any needed new ancillary data or validation reference datasets is made.
Input: ADP, scientific publications describing new algorithm

The new algorithm / module is implemented into the operational_ECV_production_system by the Processor_experts together with the algorithm experts as new prototype (separated from the existing operational algorithms).

The prototype is used for production of a substantial data amount which allows statistically significant evaluation.

The automatic standard tools for statistical analysis are invoked to prepare the statistical analysis for validation to ground-based reference datasets and for inter-comparison to other datasets (satellites, models).

The Science_team Validation_and_intercomparison_experts review the results and the entire Science_team draws a final conclusion on the value of the new algorithm which takes into account statistical validation / inter-comparison quantities, coverage achieved and spatial and temporal correlation with observed aerosol features.
By comparing the new products with the existing operational products the increment is assessed and compared to the underlying user requirements and the specified improvement goal which had been defined in the algorithm development plan.

When the new product satisfies the specified goal its prototype implementation is switched free for operational production and the respective (re-)processing and subsequent validation of the new ECV dataset can be initiated (see operational (re-)processing scenario which is then invoked).

Result New algorithm or module is included into the operational_ECV_production_system

### 4.3.3 Scenario Addition_Of_New_Input_Data

If a new satellite mission produces new relevant data for aerosol ECVs this data has to be included into the Aerosol_cci system (SR-OP-0430). To facilitate this step easily the algorithm implementation has to be modular to allow swift replacement of sensor specific pre-processing tasks (SR-FU-0141). Trigger of this scenario will be an input of the Climate research ECV users within the Science Team who require the usage of these new data (SR-MA-0240). The Algorithm experts will develop or adjust the affected modules of an appropriate algorithm which will be implemented by the system experts. The workflow rule sets of the Aerosol_cci system are then extended to include the new input data and the matching processor into the workflows. Also the new sensor input products have to be added to the Aerosol_cci system inputs as well as the ancillary data needed by the algorithm implementation.
5 OPERATIONAL SCENARIOS FOR ECV PRODUCTION SYSTEM

5.1 Interfaces within the operational ECV production system

Figure 5-1: Internal Interfaces of operational ECV Production System. See section 5.4 for detailed explanation.

Within Fig. 4.1-1 a first functional decomposition of the operational ECV production system is performed which allows fulfilling all technical functionalities required in [RD4]. This chapter shall describe these functions and the interfaces between them. The functions are grouped into four building blocks user interface, dissemination, processing and archiving.

5.2 User Interface

The User Interface is the front-end to internal (Science Team with all its sub groups) and external ECV users. As simple backbone the end user interface must contain an online ftp archive to provide access to all ECV products. To maximise the use of the ECV products, the End user interface has to be documented online and designed taking into account intuitive user interface principles (SR-FU-0190). The system shall support planning of the ECV product use by publishing a schedule when new Aerosol_cci products will become available (SR-OP-0170). Finally, the user interface shall support users to explore the ECV products themselves. In order to achieve this goal with minimal resource
need this shall be achieved by making the standard validation tools inside the operational production system publically available, by providing selected further tools (e.g. MPI scoring) or by linking to further existing and maintained external tools (e.g. AEROCOM, ICARE tools).

A separate interface handles input from the system operator. This interface has to be described within an operator manual (SR-DO-0360) in order to facilitate its utilization by technical staff which will handle the operational production. Each participating organisation who is member of the distributed system must prepare an operation manual for each ECVs.

**Functions**

**End User Interface:** In addition to an online ftp archive this component provides access to the ECV products to all users (the Science Team, external users).

**Operator Interface:** This component is used for the operational production handling by an ECV production system operator. Therefore it is the wrapper of the whole operational ECV production system regarding ECV data handling by dedicated product orders.

**Interfaces**

**Order:** This interface provides access to the production control and allows checking order status via the Operator Interface. The operator uses this interface also to place and discard orders. In order to assure the integrity of the ECV datasets, end users can not utilize this functionality.

**Dissemination** This interface allows access to already created ECV products, after release by the science team. This does not depend on anything else then the existence of the product and its release by the science team, which can also be done on partial datasets to fulfil SR-FU-0490.

**Metadata:** Information regarding available products is provided to the user interface via this interface. It directly interfaces with the central component of the archiving function.

### 5.3 Dissemination

**Functions**

**Dissemination Control:** This function handles the delivery of ordered products via a simple open public ftp archive or advanced web services (SR-FU-0190 and SR-FU-500).

**Ftp archive:** The simple open public ftp archive makes all output products accessible all time. Accordingly, the dissemination function needs to upload all respective product and metadata files for each newly released dataset to this ftp archive. The installed hardware solution and connectivity targets requirement SR-OP-0560.
Interfaces

Dissemination: This interface exchanges data regarding ordering and delivery location of archived products with the user interface function.

Get Product: The Dissemination Control has this interface to get meta information regarding products and to get products from the archiving function.

Product: If a product is requested from the dissemination function which is not available on the ftp archive the dissemination function fetches the product from the archive and stores it on the ftp archive. This interface is only needed if all output products are not permanently stored on a public ftp archive.
5.4 Processing

Functions

Production Control: Production Control handles the overall workflows for the production of all products and generates processing assignments for the processing workflow management.

Processing Workflow Management: This function receives assignments for product generation from the production control function. It then allocates processing resources, fetches needed input products and starts an appropriate processor. It includes an intelligent mechanism to determine daughter products dependant on a new input dataset and assure reprocessing only of these (SR-GE-0030) After processing it hands back control to the production control. Intermediate products will be used for the next processing steps if they have been stored earlier. Final products will be archived. For archiving of the intermediate products two alternatives exist. They can be stored in the product archive (then allowing also distribution of these to users for scientific analysis) but this would increase I/O operations between processor and archive and archive size. Or they could be stored in a separate storage under control of the processor thus increasing I/O efficiency. In the latter case the sequence of product processing may be affected. However, the current principle of level2 retrieval algorithms is operation on units of complete orbits, so that independence between the major processing units is given. For level3 products based on ECV level2 products the dependence on availability of all level2 products remains for both alternatives. Thus the choice made is for storing intermediate products under control of the processors (not a permanent archive) and archiving a limited amount of intermediate products defined by the science team in the product archive (see 4.1.4)

Processor: This component is the implementation of a scientific algorithm which creates ECV products. It also can be a tailoring or quality control processor.

The system must be able to perform a processing of the complete Aerosol ECV products (SR-FU-0360).

The repeated ingestion of new and improved input data into the aerosol_CCI production system raises the need for a frequent reprocessing of data which has to be organized (SR-OP-0390).

The production control has to contain information on inter-dependencies between products and processors in order to assure complete reprocessing (if ordered) of all dependant products when one processor or one input dataset is updated (SR-GE-0030).

To achieve a minimal need of processing during an update-reprocessing the system has to keep a selection of the intermediate products available (SR-OP-0380). Here a trade-off between archive costs and flexibility needs to be made, which means that not all intermediate products can be stored permanently. (The System shall keep intermediate products of the most recent ECV products: SR-OP-380).

Validation and inter-comparison experts shall have the possibility to request reprocessing of a specific time series of a ECV product to validate results of an algorithmic change (SR-OP-0400).

Interfaces
Options for processor integration depend on the processor’s implementation:

Direct inclusion into workflow management can be achieved if the processor is implemented in the same language as the workflow management system. Then, parameter handover will be handled directly. Implementation of that option has to take special care for error handling as the processor implementation could directly interfere with the workflow management system. This would increase the efforts for re-coding of algorithms.

If a processor is delivered as an executable (preferably static linked) package the processor has to be executed by the workflow management system through a system call. Handover of input product locations and processing parameters can be handled through command line parameters (possible problem if a high number of parameters has to be handed over) or through an options file.

The last option is to be preferred, as it is much easier for an operator to track back any errors.

For handing back results the options file can define a “result file” where the processor can write back result parameters.

The following table provides a list of interfaces within the operational ECV production system. The naming schema used is IF-OPE-<0000>, with XX = OPE (operational ECV production), and 0000 = unique number for interface. All interfaces are depicted in Figure 5-1.

<table>
<thead>
<tr>
<th>ID</th>
<th>Short Description</th>
<th>Interface Partners</th>
<th>Data type</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>IF-OPE-0230</td>
<td>Dissemination of Product</td>
<td>ECV Production System ↔ Archive</td>
<td>Archived Products</td>
<td>Dissemination IF</td>
</tr>
<tr>
<td>IF-OPE-0240</td>
<td>Metadata for Products</td>
<td>ECV Production System ↔ Archive</td>
<td>Meta information</td>
<td>Metadata IF</td>
</tr>
<tr>
<td>IF-OPE-0250</td>
<td>Request Product from Archive and deliver Product</td>
<td>Receiver ↔ Archive</td>
<td>Request for product</td>
<td>Archive IF</td>
</tr>
<tr>
<td>IF-OPE-0280</td>
<td>Archive new Product, Change product metadata, Get Product</td>
<td>Archive ↔ ECV Production System</td>
<td>ECV Product</td>
<td>Archive IF</td>
</tr>
<tr>
<td>IF-OPE-0290</td>
<td>Get product metadata</td>
<td>Receiver ↔ Archive</td>
<td>Product Metadata</td>
<td>Metadata IF</td>
</tr>
</tbody>
</table>

5.5 Quality Control

To guarantee a certain degree of product quality an automated estimation of quality parameters is done as one step of the processing. Quality parameters generated during processing are automatically evaluated and generate a quality remark.

After creation of a new set of ECV data validation and inter-comparison experts have to be consulted to evaluate the new products and document the evaluation results in the form of an uncertainty characterisation report (SR-DO-0372, SR-DO-0373). The report then shall be endorsed by the science team and released to the public (SR-DO-0378). This especially means that during a reprocessing
campaign experts have to be consulted early to check quality of new products, e.g. by first conducting a partial ECV processing. If quality cannot be approved by the experts algorithms have to be improved further and reprocessing has to be done again. Every change within the algorithm or within algorithmic parameters result in the necessity for a product quality assessment.

A validation report has to be available for each ECV product version which gives the ECV user information regarding used algorithms, input products and regarding validation and inter-comparison reports (SR-DO-0372). This activity is documented in the PVIR.

Experts for product validation and inter-comparison shall be consulted to validate ECV product quality (SR-OP-0470).

“The aerosol_cci-System shall perform an automatic quality check, inter-comparison and validation against available validation tools for each ECV product which is created”, SR-QU-370: Validation of produced ECVs is performed within workpackages 3100 (“Level 2 validation”), 3200 (“Level 3 validation and scoring against AERONET”) and 3300 (“Uncertainty validation”).

### 5.6 Validation and Intercomparison

The validation component consists of 4 independent lines, each using their specific validation module and applying a work sharing (e.g. ICARE focusing on lv2, MPI focusing on spatial/temporal correlations, MetNo focusing on lv3 and NILU working on stratospheric extinction and uncertainties of nadir products). All relevant reference data are available at the validation modules: mainly AERONET (MAN) sun photometer multi-spectral AOD, NDACC lidar stratospheric profiles and derived products, but also inter-comparison data from other satellites / algorithms such as MODIS, MISR, SEAWIFS, OSIRIS, GlobAerosol and model data from AEROCOM median or MACC. The aerosol_cci ECV products are downloaded from the ICARE server, where all EO partners store their products. The validation output is integrated into the validation reports and disseminated via the Project web portal.

### 5.7 Archiving

**Functions**

**Archive Control:** The archiving control interfaces with other entities of the operational ECV system to archive and deliver all products, to search products and to change metadata.

**Ingestion:** The ingestion function interfaces with external entities to ingest new products (satellite input, ancillary input, and reference datasets) into the operational ECV production system (SR-FU-0290, SR-FU-0300).

**Archive:** The archive stores mass data. This includes all input data of the operational ECV production system as well as all output data (SR-OP-0270, SR-OP-0280).

In the current implementation, intermediate products are not archived at all since their amount and diversity is large. However, if they were archived it would facilitate their validation (as is seen in the current experiments this can be highly relevant). It is thus concluded that the system design should allow archiving of a limited amount of intermediate data for a limited time period (where the science
team puts a focus on assessing a certain intermediate module output). It is not planned to archive all intermediate data of all product versions in order to limit the archive size. For optimizing I/O efficiency a separate storage (not a permanent archive) is used for intermediate products during processing.

A further point to be assured is that the hard disks used during data processing (part of the operational ECV production system) should be separate from the archive used for serving the ECV products to the users. The user ftp archive needs to be a big disk without high performance requirements. The processing disks need to be fast, but only sized to store a fraction of the total input, intermediate and output data e.g. one orbit, one day or one month of data at a time. Using the same physical disks would enlarge costs (a very large set of hard disks of very high performance) and user access speed would be compromised whenever the products are being reprocessed.

Metadata: Metadata of input and output products are stored in a metadata database.

Interfaces

Archive Product: This interface of the archive control allows internal entities to hand over products including meta information to the archiving function.

Change Metadata: By using this interface meta information connected to a product can be changed.

Get Product: Internal entities can get products from the archive by using this interface.

Product: This archiving-internal interface hands over products between the archive and the archive control.

Product (Metadata): This archiving internal interface hands over metadata between archiving control and archive.

Furthermore, related documentation specified in [RD4] needs to be available and again to simplify work with the products needs to be linked to the products through their metadata. To facilitate and promote use of new ECV data sets, a short overview release note (SR-DO-0341) highlighting new features and added value together with limitations and recommendations for the product use shall be provided together with a more detailed product user guide (SR-DO-0342). Also the Algorithm Theoretical Baseline Document - ATBD shall be referenced within each product (SR-DO-0371, SR-QU-0450).

It is an absolute must that all level2 and all level3 data are constantly available via an online ftp access or from a web catalogue service. Also appropriate collections must be available “via one click” (e.g. the entire level3 daily products of one ECV for one year in one zip archive or all years of level3 monthly data in another zip archive).

In support of process studies a tailoring mechanism (to extract regional sub samples from global products) must be available (either for the operator or in the user interface – depending on the frequency of this utilization type). If this type of use becomes more frequent a shopping basket shall be offered for access to individual files, where a user can select a set of individual products.

For access to documentation an online content management system should be provided which offers transparent access to version controlled documents. Respective documents must be linked from each product (ATBD, release note, validation report, and metadata).
Both, documents and product catalogue / simple ftp access have to be directly available from a web portal, which offers also the possibilities to announce new product releases and disseminate user feedback and publications evaluating or exploiting the ECV datasets.

Another functionality of the user interface should be the online access to tools for e.g. manual visualisation, data inter-comparison, statistical analysis, difference plots, seasonal / annual averaging, or regional averaging as they are currently available from the AEROCOM and ICARE browser tools (SR-FU-0191). The required functionalities for these tools need to be defined in more detail and associated costs for their implementation need to be assessed.

### 5.8 Valid ranges of values, timing considerations and data sources

Valid ranges of values, timing considerations and applied data sources are discussed in section 9.3.
6 TRADE OFF ANALYSIS FOR OPERATIONAL SYSTEM DESIGN

This chapter shall have a closer look onto the operational ECV system requirements in correlation with the possible design and deployment of the operational ECV system. Also the context of the operational system will be considered.

Based on the system constraints (section 8), the architecture analysis shall discuss several questions:

- How far can the usage of operational ECV production system components by several CCI projects increase the implementation efficiency? (section 6.1)
- Which hardware design strategy is reasonable (section 6.2)
- How efficient is one centralized design for the entire CCI program operational ECV production system as compared to a distributed approach, where each individual ECV (or group of ECVs) has a separate node (section 6.3).

In conclusion a trade-off analysis between the two extreme concepts (fully central versus fully decentralized architecture) is made (section 6.3.4) which leads to several key recommendations.

Finally, an initial cost analysis for the two alternative architecture approaches is conducted (section 6.4).

6.1 CCI Project Characteristics

To consider the right circumstances for this trade off analysis regarding hardware and structure decisions for the Aerosol CCI project this chapter shall include a short summary of the relevant characteristics of the CCI projects. It is important to find characteristics not only for the Aerosol CCI project but also for other CCI projects to take a possible synergetic use of resources into account.

![Figure 6-1: Processing Capacity usage of an operational ECV production system](image-url)
The first key characteristic to be taken into account is the main purpose of the operational ECV production systems, which is a regular reprocessing of satellite data and leads to several requirements. At first the question is to be answered if reprocessing campaigns will succeed directly to other reprocessing campaigns which means a continuous usage of the system or if the hardware capacity will be used for a specified amount of time only to reprocess data and then “run idle” the rest of the time as in Figure 6-1 where partial use may happen for limited data amounts (testing of new algorithms) in between.

A major goal of the project is to keep the operational ECV production near to science. This document contains several strategies regarding the credibility of the produced datasets thus achieved. Another aspect of the science-driven concept is the accessibility of operational structures for scientists: The system has to provide an opportunity for scientists inside Aerosol_cci to check, if their new / improved algorithm works well, thus science code has to fit into the system in an easy and efficient way. This means, an industrialisation of the algorithms should not be followed up. Only the interfaces of the scientific codes have to be partially reengineered to fit into the operational system. This has an impact on the design of the operational ECV production since scientific algorithm implementations are not unified in respect to the required platform.

Another relevant aspect for the operational ECV production system is the distributed availability of input products. This raises the necessity for either a central collection of input products or a distributed availability of input products as well as a handling of the dataflow between archives and processing environments.

Different approaches for the design of the operational ECV production system have to be evaluated regarding afore mentioned aspects as well as regarding cost efficiency and complexity.

Within the operational ECV production system there are several components which are good candidates to be shared between several CCI projects. These candidates are:

- archive
- user interface
- processing resources

### 6.2 Hardware usage strategies

The processing hardware used within Aerosol_cci production system should be scalable and easily extendable. Additional considerations can be space requirement and energy efficiency. This section will only describe the usage of the processing hardware.

The following hardware environments were considered:

- GPU Usage
- Grid computing
- Cloud computing
- Cluster computing
6.2.1 GPU Usage

Recently available graphics processing units (GPU) can perform several hundred calculations in parallel. This has made this hardware interesting for high performance computing. The high degree of parallelisation, a fast memory connection and a relatively cheap price are the main reasons for this development. The high degree of parallelisation is due to the fact of reduction of processing capability to simple instructions.

Use of GPU-based processing requires all the programmes to be implemented within special programming languages. This contradicts the requirement that already implemented scientific algorithms shall be used and therefore makes GPU-usage an option, which cannot be considered within the scope of this project.

6.2.2 Grid

A grid consists of a number of mostly heterogeneous computers which are connected. A big advantage lies especially in the opportunity to use almost any hardware to form a grid. This advantage also can be a disadvantage when considering maintenance. If the grid consists of a whole variety of computers the administration effort is much higher than within a homogenous system.

Within a grid an application can be allocated almost any processing power. The borders of a computer system form the border in flexibility here. Regarding the afore mentioned characteristics of an operational ECV production system the usage of grid computing can be evaluated as following.

Regarding a requirement of regular reprocessing grid computing provides no special advantages or disadvantages. If processing is done only during a specific period the heterogeneous grid system could be allocated to other projects in between.

Regarding a diversity of scientific algorithm implementations a grid system could be in an advantage. The system could, provided a processing framework supporting this option is available, facilitate matching processing hardware and platforms for each algorithm implementation and therefore can provide an adequate allocation of hardware.

Regarding space and power requirements a grid system can be very efficient, if current hardware is used. But the strength of a grid system lies within the usage of a variety of hardware and therefore could be difficult to handle in respect to complexity.

6.2.3 Cloud

Cloud computing is the usage of distributed computing resources provided by mostly commercial suppliers. As the service offered is only rented big investment has not to be done. On the other hand renting will only be cost effective if the processing power is not needed permanently or only a marginal amount of processing power is needed.

A big disadvantage is the need of high-capacity networking connection to connect to the cloud. Also one loses control over its data. The usage of cloud computing raises severe security issues which can be a cost driver when dealt with properly.
Regarding peak load usage cloud computing has an advantage to other structures. As hardware can be rented “within the cloud” as used, cost will only be generated if processing power is used and therefore no idle hardware has to be paid. On the other hand cloud hardware usage is more expensive than hosting the hardware in one’s own processing centre or processing centres which makes it in the end almost equally expensive.

As cloud processing power is provided as a service, no considerations of space or power requirements have to be done. Complexity of the overall system will be handled by a huge portion by the cloud provider but the complexity and cost of the contractual situation is increased as a commercial cloud provider has to be handled.

A cloud structure can usually provide, by virtualising processing hardware, different platforms and therefore provide matching environments for most of the scientific algorithm implementations.

**Example Calculation**

To process a 10 year record of AATSR data, the ORAC-processor would need hardware with a price of approx. 6 k€ and take

\[
14,4 \frac{orbits}{day} \times 365,25 \frac{days}{year} \times 20 \frac{min}{orbit} \times 10 years = 715 days
\]

To process this 10 year record within 2 months therefore would require a hardware investment of approx. 12 times the afore mentioned hardware (i.e. 6 x 6k€ = 72 k€) In addition annual maintenance of the own hardware is considered as 10% of the investment per year.

If processed “in the cloud” the required hardware would cost (taken from amazons pricing table \(^1\) High Memory Double Extra Large cost of 1,012 $/h) 12 x 1,012 $/h = 12,144 $/h – this amounts to a cost of approx. 14 k€ for this single processing campaign.

<table>
<thead>
<tr>
<th>Processing Campaigns</th>
<th>Timeframe</th>
<th>Cost in Cloud</th>
<th>Cost in own hardware</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1 year</td>
<td>14 k€</td>
<td>79 k€</td>
</tr>
<tr>
<td>2</td>
<td>1 year</td>
<td>28 k€</td>
<td>86 k€</td>
</tr>
<tr>
<td>10</td>
<td>5 years</td>
<td>140 k€</td>
<td>108 k€</td>
</tr>
</tbody>
</table>

This table does only include pure hardware! It does not include the cost of data transfer into the cloud. The above calculation is done under the assumption of only partial usage of the hardware and the hardware running idle rest of the time. If hardware is used 80% of the year the cost is as following:

<table>
<thead>
<tr>
<th>Processing Campaigns</th>
<th>Timeframe</th>
<th>Cost in Cloud</th>
<th>Cost in own hardware</th>
</tr>
</thead>
<tbody>
<tr>
<td>4,8 (equivalent to processing for 80% of a 1 year)</td>
<td>1 year</td>
<td>68 k€</td>
<td>79 k€</td>
</tr>
</tbody>
</table>

\(^1\) [http://aws.amazon.com/de/ec2/#instance](http://aws.amazon.com/de/ec2/#instance)
In conclusion it is obvious that own hardware which is used 80% of the year for reprocessing (by e.g. scheduling different datasets reprocessing spread over the year) is significantly cheaper over 5 years than cloud computing, whereas rarely used own hardware (only 2 of 12 months) is more expensive. So the dimensioning of the hardware and the scheduling over the year of reprocessing campaigns for several datasets should take this fact into account.

### 6.2.4 Cluster Computing

A computer cluster is formed by several, mostly identical, computer devices. Each of the nodes is connected to the other nodes with a high performance network connection (e.g. Infiniband). A cluster of computers therefore is an expensive capital asset. By forming a homogenous resource the administration effort is less than e.g. the effort for a grid.

The homogenous hardware allows a good scalability within the high performance network and the similar hardware makes it easy to clone installations onto new hardware. High availability can be easily achieved as hardware is identical when using a matching virtualisation environment.

Regarding the first requirement, a peak load of processing during a part time only cluster computing does not provide any advantages – nevertheless it does not prevent usage of the hardware for another project during idle time.

Regarding the need of different platforms for different scientific algorithm implementations the cluster computing can provide, as cloud computing, by virtualisation different platforms for different implementations.

Space and power can be handled optimally within a cluster of homogenous hardware – complexity of administration is decreased regarding hardware administration – Handling of virtualisation has to be added to the complexity of the overall system.

### 6.2.5 Summary

As could be seen in the previous sections no way of organising hardware usage can be seen as a definite winner. The recommendation as a result of this analysis – taking into account also a possible distributed processing approach as well as sharing of capacity with other projects – are the following items

- homogenous hardware in combination with
- virtualised environment

The four hardware scenarios could be ranked regarding their fit with these key characteristics. The following figure shows on the x-axis homogenousness of the hardware usually used within a scenario and on the y-axis the fitness of a hardware scenario for virtualisation.
This means the hardware scenarios cluster computing and cloud computing are equally usable for the purposes of the CCI production system. Grid computing could be used for these purposes – but usually is working in other scenarios. GPU usage should not be used as it would not allow – by providing a virtualised environment – to reuse already developed scientific algorithms.

### 6.3 System deployment: Centralised vs. distributed system

An aspect which cannot be handled solely by the processing hardware organisation is the overall system distribution. Within this document some ideas for organising a distributed system are given. Within this chapter an analysis of different approaches shall be given. Therefore some general concepts are presented:

#### 6.3.2 Monolithic system

This concept relies on collocating the whole system hardware in one place. This includes processing hardware as well as archive and dissemination functions. An analysis of the characteristics of this system comes to the following conclusions:

A major drawback of a monolithic system is the generation of an entity which concentrates all knowledge needed for operating the operational system in only one place and therefore generates a certain distance between operational system and scientists developing new algorithms.

#### 6.3.1 Centralised approach

A centralised approach has several advantages as:

- fewer WAN connections have to be considered: A local network can be used and transfer times minimised
- service teams have to be only available at one single point

but this approach has also disadvantages as

- processor specialists at different facilities will not have direct access to the environment
- if the system drops out the system drops out as a whole
- if a sharing of resources between all ECV occurs an almost unmanageable huge system could be created

A major drawback of a monolithic system is the generation of an entity which concentrates all knowledge needed for operating the operational system in only one place and therefore generates a certain distance between operational system and scientists developing new algorithms.

### 6.3.2 Distributed service oriented system

Service orientation in this context means the provision of several services as the processing data from one level to another level or the dissemination of products to end users as well as the provision of archiving services or input data.

![Service oriented deployment of an operational system](image)

Figure 6-3: Service oriented deployment of an operational system

Within a complete service oriented architecture (Figure 6-3) there would be a need for an orchestration unit to control the appropriate services in a way, that processing from an archived input product to a higher level output product is possible. The orchestration would have to take into account cost of processing in every aspect. That would be transfer time, processing time, et cetera. Also it would be possible to identify hardware where processors could run on and use that hardware service to process. Figure 6-4 shows an example of processing from one level to another.
The use of a completely service oriented architecture would increase largely the flexibility of the operational ECV production system. Service orientation is compatible with any concept of hardware usage strategy. If used in several CCI projects cost of processing and archiving hardware can be shared to a large degree. It would even be possible to give scientists the chance to join the service oriented system and provide an experimental processing service to test a new algorithm within the system without deploying it into one of the other facilities.

As the science community is distributed this seems a natural approach to design the operational ECV production system. Advantages of these solutions are:

- this approach allows highest flexibility in sharing resources
- Processors could be operated at the premises of each processor expert and thus nearest to the science
- Politically beneficial more facilities can be involved into the processing

Also some disadvantages have to be taken into account:

- If the system is distributed between several locations WAN-connection bandwidth can become an issue
- Reaction and communication time between different facilities which operate parts of the system can become an issue if the workflows and responsibilities are not optimized for the decentralised approach
- Expertise on system maintenance has to be available at several locations

### 6.3.3 Fixed distribution with focus on input data

This concept tries to remove some complexity from the service oriented concept presented in the last section. This concept assumes that processors and required (high volume) input products are collocated. No concertation would here be needed with respect to archive or processing hardware. The whole system would be designed such, that a fixed configuration of hardware/ software allocation reaches an optimum (e.g. one sensor per location).

Archive reachability by external entities assumed it also would be possible to include scientific processor entities into such a concept. These would then be configured into the system and provide
scientific algorithm developers with the opportunity to test the algorithms within the operational ECV production environment.

### 6.3.4 Conclusion

When weighting the arguments provided in the last chapters one has to take into account the main goals of the system: to be near to the science and flexible in response to continuously evolving user requirements and algorithms. A first trade-off analysis between a de-centralized and a centralized system design is summarized in the following table. In this analysis the extreme points stand for one system encompassing all 13 CCI ECVs (or a subset as for example the atmosphere ECVs) versus a set of 13 separate systems for each ECV. At the same time the extremes point to one central system for the aerosol ECV versus a decentralized system within Aerosol_cci itself. In applying this self-similarity on different scales of the envisaged system it is assumed that the conclusions between the two extremes on each level tend to be similar.

The following table lists arguments in favour of each architecture extreme in columns 2 and 3 and a trade-off and possible optimal solution in the last column.

<table>
<thead>
<tr>
<th>Trade-off aspect</th>
<th>Centralized system</th>
<th>De-centralized system</th>
<th>Trade-off discussion</th>
</tr>
</thead>
<tbody>
<tr>
<td>User-friendliness</td>
<td><strong>One access point</strong></td>
<td>Tailored ECV access point</td>
<td><strong>Use of a common web portal is recommended</strong></td>
</tr>
<tr>
<td></td>
<td><strong>Multi-ECV access support</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Scientific credibility</td>
<td><strong>De-central elements near to algorithm experts</strong></td>
<td></td>
<td><strong>Algorithm development and processing should be de-central</strong></td>
</tr>
<tr>
<td>Product quality</td>
<td>Common quality standards</td>
<td>Tailored ECV quality standards</td>
<td><strong>An appropriate balance between common and specific standards needs to be coordinated</strong></td>
</tr>
<tr>
<td>Consistency</td>
<td>Apart from data access the system architecture has no impact on consistency – however a major coordination effort is necessary in all cases to assure consistency and its clear documentation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Flexibility &amp; scalability</td>
<td>Support by shared resources</td>
<td>Support by lower complexity</td>
<td></td>
</tr>
<tr>
<td>Coordination efforts</td>
<td>Focus on consistency</td>
<td><strong>Lower complexity and specific ECV focus</strong></td>
<td></td>
</tr>
<tr>
<td>Community integration</td>
<td>Could be more objective to wider algorithm community</td>
<td><strong>Closer identification and lead by algorithm experts</strong></td>
<td><strong>Lead by algorithm experts is regarded highest priority</strong></td>
</tr>
<tr>
<td>Failure safety</td>
<td>Higher back-up capabilities</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Definition of priorities | Lower likelihood of conflicts |
--- | --- |
Complexity of workflows | Lower complexity of workflows | Complexity of de-central system can be limited by organising the distributed components with lowest number of sub level interfaces (e.g. all lv2 algorithms for one sensor at one node) |

External interfaces | Total bandwidth of interface(s) would be similar |
Internal interfaces | Clearly less WAN interfaces needed |
Processing cost | Total processing cost of both system architectures are assumed similar when optimal scheduling is achieved |
Output archive cost | Total archive cost of both system architectures are assumed similar |
Input archive cost | Duplication of input data can be avoided | Back-up and faster access to distributed input data | Trade-off between archive and interface costs needs to be made |
Interface cost | Smaller cost for internal interfaces between processing nodes | smaller internal interface costs for multiple transfer of large lv1 datasets |
Maintenance cost | Smaller cost by more effective maintenance is expected |
Metadata & catalogue | Coordination of metadata and a central frontend to the CCI catalogue are necessary anyway |
Documentation | Coordination of a central frontend CCI documentation are necessary anyway |

According to this analysis the major arguments for a central system are its support of user-friendliness, its smaller input archive costs and its lower WAN interface costs between processing nodes. On the other hand the de-centralised architecture is strongly favoured with its scientific credibility and community integration by nearness to algorithm experts as well as its lower complexity and its tailored focus to each ECV. We therefore recommend a compromise, where components for input data access and for the user interface are shared between many ECV systems, but the components for archiving and processing are implemented in a de-centralized architecture (tailored to each ECV and where possible broken further down within one ECV per sensor line).

The overall operational ECV production system will then share as much resources amongst similar projects to reduce cost and engage communication between different science communities. Also a distributed approach should be chosen to allow the involvement of a widespread community directly into the operational production and to encourage an active participation in the projects further development.
To guarantee a science-driven system a distributed approach for the operational ECV production system has to be chosen. To avoid too high complexity a distributed system with a fixed configuration as discussed in section 6.3.1 seems to be the best choice. Between separate processing facilities an agreement on a minimalistic set of common interfaces has to be established; especially with respect to dissemination. Processing within the separate facilities does not have to be unified. To handle large amounts of data and also provide exchangeability of processors between different processing facilities a common processing framework should be used.

6.4 Associated costs

To evaluate the different design approaches an estimation of the costs associated with the different scenarios is provided. At first the two main alternatives shall be shortly analysed with respect to the main cost drivers, which have to be taken into account. The following cost estimates encompass the aerosol ECV products only in both centralised and decentralised design.

6.4.2 Centralised model

**Archive:** In a centralised system input, intermediate and output data archives have to be available, which is able to handle the full amount of data needed by the project. All input data has firstly to be stored within the archive. The cost to maintain such an archive can be calculated using the amount of data to be stored (186TB) and the cost per GByte (1.-€). This cost for storing 1 GB seem very high but one has to consider, that several copies of one dataset have to be stored to ensure operational long-term availability of data. Cost depends also on the type of medium the data is stored on. If a high-end LTDP (long term data preservation) medium is used cost differs extremely too much cheaper end-user hardware (NAS, LTO). Another factor is the migration which has to be performed regularly from one media generation to the next one. Usually one calculates the need for migration at the latest every 7 years. Nevertheless these migrations will in the coming years result in a cost reduction by a factor of up to 4. In addition an archive has high start up costs if not already available within a multi mission environment and maintenance cost of 15% of the start up cost has to be taken into account per year of operational usage.

**Processing:** When choosing a centralised approach the processing hardware will be located at one single facility. The cost for this hardware investment can easily be calculated using the figures for computing power provided in the requirements. In addition to the investment cost an overhead of 15% of this cost per operational year for maintaining the system has to be taken into account.

**WAN Data Transfer:** For a centralised approach we only have to take transfer of data to users into account. We assume a 1.5 MByte/sec network connection is used.

**System operator cost:** We assume that the Aerosol_cci operational costs come on top of an existing data center and thus half a person will be sufficient (estimated average cost per person month 10k€)

**Science team:** We assume on average the following resources split over the science team: leader (one third of a person), user experts (half a person), validation experts (half a person), algorithm experts (2 persons), thus a total of 3.3 persons.

The overall cost of the centralized approach consist of

**Archive:**
Start up investment: 500k€ (minimum)
Archive cost per year: 186k€ / year
Maintenance cost per year: 75k€ / year

**Processing**
Start up investment 40k€
Maintenance cost per year 6k€ / year

**Data Transfer**
Cost per Year for a 1.5 MByte/sec connection 4k€ /year

**System operator**
Cost per year: 60k€ / year

**Science team**
Cost per year: 400k€ / year

**Total cost**
Start up investment 540k€ (minimum)
System operations cost per year 331k€ / year
Science team cost per year 400k€ / year

### 6.4.3 Decentralised model

**Archive**: In a decentralized approach data will be stored redundantly on the system. We estimate that 50% of the data is stored twice within the system and two archives are present.

**Processing**: We estimate the same amount of processing power is available as in a centralised approach.

**WAN Data Transfer**: Depending on the structure of the decentralised system data has to be transferred between different processing and archiving locations. Assuming an efficient distribution (level2 processors for one instrument are all at one node or level1 input datasets are copied once) we estimate that 50% of the input and output data has to be transferred during the process of ECV production using WAN connections. This would accumulate to 93 TB data per year which is approximately 3 MByte/sec. In addition to this, the delivery of products to users needs 1.5 MByte/sec band width.

**System operator cost**: We assume similar operator cost distributed over the different nodes.
**Science team:** We assume a similar cost for the science team work on algorithms. However, it is expected that the level of formalism and the effort for harmonized documentation is decreased if each decentral node (e.g. separate ECVs or products per algorithm) can focus on specific type of algorithm. Thus we reduce the science team effort by 25%

The overall cost of the decentralized approach consist of:

**Archive:**
- Start up investment 2 x 375k€ 750k€ (minimum)
- Archive cost per year 2 x 139k€ / year 278k€ / year
- Maintenance cost per year 2 x 56k€ / year 112k€ / year

**Processing**
- Start up investment 2 x 20k€ 40k€
- Maintenance cost per year 2 x 3k€ 6k€ / year

**Data Transfer**
- Cost per Year for a 4.5 MByte/sec connection 10k€ / year

**System operator**
- Cost per year 2 x 30k€ / year 60k€ / year

**Science team**
- Cost per year 300k€ / year

**Total cost**
- Start up investment 790k€ (minimum)
- System operations cost per year 466k€ / year
- Science team cost per year 300k€ / year

### 6.4.4 Conclusion

Looking at these cost estimates a central system has about a 30% advantage on start up investment and operations cost, but larger science team cost by about 25%. In toto, there will be almost equal operations costs for a centralised approach and a decentralised approach.

One of the main drivers for the CCI program is the production of credible high quality ECV products, which is mainly achieve if a high level of scientific expert involvement can be assured. We think that it is impossible to translate this benefit of a decentralised approach into a cost estimate.
Furthermore, the two approaches, centralistic and decentralised, have to be evaluated with respect to the actual situation of the participating facility or facilities. It is very likely, that participating facilities will already have an archive in place which means no or only an upgrade start up investment will have to be made. Also it is very probable, that participating facilities will already have a part of the input data available in their archive (e.g. if they operate the level1 processing on behalf of ESA), which means the costs for data storage (especially the decentralised copying) will not be fully born by the CCI program.

In summary, the cost estimate shows that there are cost savings by a centralised approach (especially regarding investment), but they depend very much on the choice of operational nodes and their underlying tasks with regard to ESA sensors (e.g. operating PAC services). On the other hand, the decentralised approach allows focusing the required expertise on a specific type of algorithms and assures best the strong involvement of all relevant experts. It is in the light of this benefit, that we advocate the decentralised implementation. This statement refers to the CCI program as a whole (no ONE system implementation), but also to potential sub groups of ECVs (e.g. all atmosphere or clouds and aerosols integrated) and furthermore within one ECV (e.g. splitting to different nodes for different instruments within one ECV).

Finally it should be explored, whether some elements (e.g. input data archive, user interface) can be shared between the distributed ECV systems. We do not expect a major cost saving by this approach (as discussed before), but we expect benefits on consistency in input data and output to users (formats, variable names, etc.). Furthermore, the science teams of each ECV need to interact pro-actively to ensure scientific consistency (e.g. a common cloud masking working group for all ECVs which need to avoid cloud contamination or interaction between 2 ECVs such as cloud and aerosol ECV products to assure that no double counting of pixels in both classes is made).
7 ENVIRONMENTAL CHARACTERISTICS OF THE SYSTEM

7.1 Concrete Implementation of Aerosol_cci 2

This section gives an overview of the current status of the major system components operated and maintained by the partners.

<table>
<thead>
<tr>
<th>DLR</th>
<th>Operating Content Management System (<a href="http://www.esa-aerosol-cci.org">http://www.esa-aerosol-cci.org</a>)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>- The ownership of domain name (esa-aerosol-cci.org) is renewed by DLR on regular basis</td>
</tr>
<tr>
<td></td>
<td>- Hosting platform (virtual machine) with performant connection to the internet is operated and maintained. Moreover, sufficient storage capacity for Aerosol_cci web content and documents is available. Aerosol_CCI2 CMS is based on Drupal 6.</td>
</tr>
<tr>
<td></td>
<td>- Security updates (maintenance) are performed as needed. Since 24.02.2016 Drupal 6 is no longer supported. No security updates will be released.</td>
</tr>
<tr>
<td></td>
<td>- A backup of the CMS’s content is run on a daily basis.</td>
</tr>
<tr>
<td></td>
<td>- Drupal version 6: Since Drupal only provides security updates for the most recent two versions of Drupal, ESA’s cci-projects will encounter severe problems. Drupal version 8 was released on 19 November 2015. So we have unpatched Drupal 6 instances running since then. A migration of the current Drupal 6-website to Drupal 7 or 8 will not easily be possible (according to our analyses).</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ICARE</th>
<th>Data archive service</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>- Disk archive to centralize all versions of all Aerosol-cci data sets online. Includes a private workspace for internal Aerosol cci use and a public area for data dissemination (one public version for each data set)</td>
</tr>
<tr>
<td></td>
<td>- Upload mechanism accessible to Aerosol cci EO data providers to let them freely upload data sets produced in their own facility. Uploaded data sets must follow product guidelines (format, naming convention,</td>
</tr>
</tbody>
</table>
| ICARE   | Data access service | Access to ACCI data sets for both project internal use and data dissemination:  
|         |                     | - Restricted access to internal ACCI workspace (access to all versions of all data sets, restricted to project partners using specific identification)  
|         |                     | - Open access to public ACCI area (one public version for each data set, access with generic identification (username: cci, password: cci, no registration required))  
|         | FTP access:         | - Internal workspace:  
|         |                     | - Public area:  
|         | FTP-like web access:| - Internal workspace:  
|         |                     |   http://www.icare.univ-lille1.fr/archive/?dir=CCI-Aerosols_internal  
|         |                     | - Public area:  
|         |                     |   http://www.icare.univ-lille1.fr/archive/?dir=CCI-Aerosols  
|         | SSH access:         | Direct access to disk space (private workspace and public area) and to computing resources for data processing, subsetting, analysis, etc.  
| ICARE   | Web tools           | Web tools (browse, extract) support selected parameters of all CCI EO data sets, along with other existing data sets archived at ICARE.  
|         |                     | Data sets uploaded by ACCI EO data providers are automatically discovered and supported as long as they follow the product guidelines.  
|         | Browse interface:   | - Graphical web interface for visualisation,
analysis, evaluation purposes, and comparison to other CCI and non-CCI data sets (satellite, ground-based, model analyses)

- Private data sets are accessible upon identification:
  http://www.icare.univ-lille1.fr/browse/?project=cci

- Public data sets are freely accessible (CCI-Aerosols):
  http://www.icare.univ-lille1.fr/browse

Extract tool:

- Web interface for interactive subsetting around validation sites

- Private data sets are accessible upon identification:
  http://www.icare.univ-lille1.fr/extract/?project=cci

- Public data sets are freely accessible (CCI-Aerosols):
  http://www.icare.univ-lille1.fr/extract

7.2 Processing Hardware dimensioning

Hardware dimensioning is directly linked to processing algorithm performance as well as the amount of input data. For creation of the input data currently addressed within [RD1] the equivalent of 3 8-core x86 machines are needed to fulfil the requirements regarding processing power. Each of the cores must have 2GB RAM accessible. For storing intermediate products the processing systems have to have access to 6 TB online storage. This dimensioning also is an answer to SR-OP-0310, SR-OP-0540, SR-OP-0541 and SR-OP-0542.

7.3 Archive dimensioning

To fulfil requirements regarding storage of input, output and intermediate products the archive has to have a size of 186TB (SR-OP-0340, SR-OP-0510, SR-OP-0520, SR-OP-0551 and SR-OP-0550).
8 SYSTEM CONSTRAINTS

Aerosol_cci Phase 2 is realised as a distributed system: Processing capacities are located at the partners’ facilities. Therefore the processing environment likely becomes quite heterogeneous, since each partner applies different solutions. Moreover, the knowledge on how to operate an individual processor remains at one place, the partner’s facility. In case the processor-developer leaves the Aerosol_cci partner institution, there is the risk that the capability to run the processor on short notice is reduced. However, all aspects of each processor need to be documented in the appropriate document (e.g. ATBD for algorithm-relevant information, PSD for product specification, PVIR for product validation and intercomparison, etc.). Since the system is distributed over a number of partners, there is the rest that product quality flags or uncertainty information differs within Aerosol_cci 2 partners.

Users working with these long time series for climate research do not want to spend or cannot afford to conduct intensive pre-processing of the ECV datasets. In order to make the system as user friendly as possible products should thus be consistent in their quality flags and each product should contain a “best” variable, which is recommended for users without further utilization of quality flags. For expert users with more detailed requirements clearly documented quality flags and additional data fields which contain more pixels to achieve better coverage but at the cost of reduced accuracy may be provided, but need to be clearly identified.

A very important aspect is the timeliness of ECV production, which severely depends on the capabilities of the individual partners. Although all Aerosol_cci partners ensured to meet ESA’s requirements, a significant uncertainty in reprocessing timeliness arises.

Since no monolithic implementation at one of the partners is conducted, the effort (costs) of maintaining the system is quite equally distributed among all partners. So each partner is responsible for the costs to produce and disseminate his ECV(s).

Conclusively it has to be stated that – when being aware of the mentioned system constraints – the layout of the Aerosol_cci ECV processing system provides a high degree of robustness and proved to be capable to meet ESA’s requirements with respect to timeliness and quality of the results.
9 SYSTEM MAINTAINANCE

To maintain operability during the 15 years of anticipated system life time of the Aerosol_cci production system several factors have to be considered. These factors are:

- Overall system aspects: Software and Hardware
- Description of the elements of the decentralised processing chains
- Growing amount of input data

These three factors lead to several precautions, which have to be taken.

9.1 Overall system aspects: Software

During the lifetime of the Aerosol_CCI system several changes within the software components used mainly driven by progress in science and development will occur. To maintain operability of the system for software changes on the one hand test suites have to be designed which can verify and validate the changes within software which occur. On the other hand also the platforms, the software is running on will change during the operational period, which generates the need for software adaptations.

As the system is designed for constant change, the software life cycle management has to answer this need for constant change. This means, the software development cycle has to be ongoing during the system’s life time. A design for constant change has to rely on the constant availability of scientists, system and software engineers to answer immediately to changing requirements. Figure 9-1 describes the software development cycle as proposed for this project. System requirements of the figure are the
development priorities of this project, user requirements in the figure are the user feedback collected by the science team within this project.

The gradual change is an aspect which is also needed to fulfil SR-GE-0090.

### 9.2 Hardware life cycle

As hardware today has relatively short lifetime its maintenance includes the exchange of hardware against new hardware during the lifetime of the Aerosol_cci production system of 15 years (SR-OP-0320). To allow growing processing throughput hardware upgrades within its lifetime have to be assured with respect to growing amount of input and output data and to growing algorithm needs (answer to SR-OP-0330 and SR-OP-0340). To fulfil this requirement, the system performance needs have to be considered regularly every 6 month (SR-OP-0552).

### 9.3 Description of decentralised processing chains

Each processing chain should – according to the ECSS-requirements analysis conducted in section 10—be characterized by the following categories:

- Overall software architecture
- Memory and CPU budget
- Design standards, conventions and procedures
- Internal interface design

This section intends to achieve full transparency about the demands with respect to hardware requirements and time consumption of the individual distributed processing chains. This information should provide the science team, ESA and other relevant stakeholder with information about feasibility of timely ECV production by the team.

Each processing chain does assure configuration control, documentation, engineering standards, modularity, flexibility, transferability, maintainability, operability

#### 9.3.1 BIRA-IASB: AerGom / GOMOS

<table>
<thead>
<tr>
<th>Processing obligations</th>
<th>Y1 + Y2: Full reprocessing of 10 year global datasets in 3 months</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall software architecture</td>
<td>The AerGom code is developed within a science environment. Two versions of the retrieval code are available. Active development takes place on a Matlab version which is fully modifiable, and a compiled C-version also exist but because the source code is not available to us (developed by ACRI during the initial AerGom project), this limits its use for further improvement of the code (although many processing parameters can be changed without recompiling). The processing is</td>
</tr>
<tr>
<td>Design standards, conventions and procedures</td>
<td>There is currently no version control (e.g. SVN, Git) in place, although the way the algorithm is constructed in Matlab, different versions are available from within the algorithm simply by selecting the appropriate version before starting the processing. As the development is starting to generate a large amount of code which is difficult to keep track of manually, it is foreseeable that version control might be used for future development. Active development is carried out on Matlab, but a conversion to another programming language could be considered in the future to improve the processing time.</td>
</tr>
<tr>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>Internal interface design</td>
<td>Batch processing</td>
</tr>
<tr>
<td>Valid range of values</td>
<td>No enforced valid range of values during processing</td>
</tr>
<tr>
<td>Data sources</td>
<td>GOMOS residual extinctions (GOM_EXT files), ESA product version 6.01.</td>
</tr>
</tbody>
</table>
9.3.2 BIRA-IASB: IASI

| Processing obligations | Y1: Processing of 1y “extended Sahara” dataset  
Y2: Processing of 10yr “extended Sahara” dataset in 6 months | Overall software architecture | The MAPIR IASI dust retrievals are for the first time included in a full chain that allows processing large amounts of data. This chain has been used for the Y1 round robin processing and has been further improved. The processing chain has been adapted to allow the use of the new HPC cluster. The first processing (RR) has been undertaken on several Linux cluster(s), being selected within the compute servers available at BIRA-IASB. As scientific improvements were made during the round robin analysis, two reprocessings of the RR have been done using the new HPC cluster which was in final testing. The final 10 years processing has been undertaken on the new HPC cluster, using the PBS queue system. Here under is a graphical description of the MAPIR processing chain. Additional information is to be found in the ATBD. |

| Memory | The retrieval method is very time consuming (optimal estimation with in |
| **and CPU budget** | line RT) but the necessary compute power was available. Estimated cpu time for the final processing: 15 cpu-days per day
For future reprocessings, a new version of MAPIR is under development with a new data pre-filtering that will strongly reduce the number of retrievals to run.
Available compute power:
- 10 times 2x 6-core (in total 120 CPU) with hyperthreading Intel Xeon 2.66GHz, 24 GB RAM
- 6 times 2x 6-core (in total 72 CPU) with hyperthreading Intel Xeon 1.2GHz, 32 GB RAM
- 4 times 4x 8-core (in total 128 CPU) Intel Xeon 2.66GHz, 128 GB RAM
- 2 HPC clusters 44+12(fat) nodes * 2 Xeon E5-2680V3 CPU (12core, 2.5GHz); in total 2688 CPU; this cluster is shared with the Royal Observatory of Belgium and the Royal Meteorological Institute of Belgium and uses PBS Pro scheduling system |
| **Design standards, conventions and procedures** | - The different parts of the code are written in different languages (C++, f77, Matlab) but interfaces run smoothly
- Versioning (SVN) in place for the most important part: the ASILID code; the other codes do not have physical importance (mainly data handling and a few filtering) and are maintained locally with manual version control linked to the algorithm number.
Additional documentation (including ATBD) and versioning in case of major changes in the codes (due to changes in the retrieval strategy, or input data selection for example) |
| **Internal interface design** | Batch jobs submitted to the PBS queue on the HPC cluster. |
| **Valid range of values** | Dust AOD at 10µm and 0.55 µm:
Valid range of values 0 … 10 (dimensionless) |
| **Data sources** | EUMETSAT IASI l1c, in Principal Components Scores when available
EUMETSAT IASI I2 H2O and T profiles, surface Temperature
NOA CALIOP dust vertical profile climatology (Amiridis et al)
NASA IASI land surface emissivity (Zhou et al)
ECMWF ERA-interim Skin Temperature |
### 9.3.3 DLR: ATSR L3 ensemble

<table>
<thead>
<tr>
<th>Processing obligations</th>
<th>Y2: Full reprocessing of 17yr ensemble dataset in 3 months</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall software architecture</td>
<td>simple algorithm to calculate (weighted) means of the 3 ATSR algorithm values is processed for 17 years full mission ATSR2 and AATSR</td>
</tr>
<tr>
<td>Memory and CPU budget</td>
<td>Currently the L3 ensemble generation is operated within the “Geofarm”-Cloud-Computing facility at DLR-DFD, a suite of AMD machines with the following specification are applicable:</td>
</tr>
<tr>
<td></td>
<td>- 4 x 6 Cores 64bit-AMD OpteronI Processor 6176, up to 64 GB per virtual machine.</td>
</tr>
<tr>
<td></td>
<td>On all of the above mentioned machines, a Linux software environment (OpenSuSE or debian) is available.</td>
</tr>
<tr>
<td></td>
<td>Needed processing time:</td>
</tr>
<tr>
<td></td>
<td>The processor needs 2 days (wall clock time) per month of L2 (A)ATSR data.</td>
</tr>
<tr>
<td></td>
<td>When simultaneously running 12 instances: 2 days wall clock time per year of L2 data is realistic. This yields approximately 1 month wall clock time for 10 years of L2 data.</td>
</tr>
<tr>
<td>Design standards, conventions and procedures</td>
<td>(A)ATSR ensemble production is written in IDL. Source codes are under version control.</td>
</tr>
<tr>
<td>Internal interface design</td>
<td>Batch job (cshell-processing environment)</td>
</tr>
<tr>
<td>Valid range of values</td>
<td>Result: Gridded and area-averaged AOD with 0…2</td>
</tr>
<tr>
<td>Data sources</td>
<td>L2 + L3 of all three Aerosol_cci (A)ATSR-algorithms</td>
</tr>
</tbody>
</table>

### 9.3.4 DLR: IASI

| Processing obligations | Y2: Reprocessing of 10yr global dataset in 6 months |
Overall software architecture
IASI retrieval is developed in science environment which is already adequate for operational processing within Aerosol_cci 2. Separation of pre-processing (only executed once) and processing (repeatedly applied for improved versions) facilitates fast reprocessing capabilities.

The viewgraph below depicts an overview of the DLR IASI processing chain:

Memory and CPU budget
The available hardware includes two identical machines with the following specifications:

- 2 x 12 Cores 64bit-Intel “SandyBridge” CPU @ 2.7 GHz, 256 GB RAM, 40 TB hard drive.

For the duration of the processing, this hardware will exclusively be available for the purposes of the Aerosol_cci-project.

Currently the IMARS processing chain is developed and operated on the following machine:

- 32 x 64bit-AMD OpteronI Processor 8384 with 128 GB RAM

Moreover, within the “Geofarm”-Cloud-Computing facility at DLR-DFD, a suite of AMD machines with the following specification are applicable:

- 4 x 6 Cores 64bit-AMD OptI(tm) Processor 6176, up to
64 GB per virtual machine.

On all of the above mentioned machines, a Linux software environment (OpenSuSE or debian) is available.

**Estimation of required processing time:**
- Preprocessing takes 2 days/month (full spectra)
- Processing: 1 day/month using one instance of the processor. Speed-up is realized by increasing the number of instances.

<table>
<thead>
<tr>
<th>Design standards, conventions and procedures</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Version control system for IASI retrieval algorithm: Application of DLR-wide “SVN” solution</td>
</tr>
<tr>
<td>- Versioning and documentation outside SVN: Characterizing major changes in code by additional documentation and reported in the ATBD.</td>
</tr>
<tr>
<td>- Code of IASI retrieval is written in Fortran 90.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Internal interface design</th>
</tr>
</thead>
<tbody>
<tr>
<td>Batch job (cshell-processing environment)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Valid range of values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dust AOD at 10µm and 0.55 µm: Valid range of values 0 … 10 (dimensionless)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Data sources</th>
</tr>
</thead>
<tbody>
<tr>
<td>L1c IASI spectra (xxx/PCS) obtained from EUMETSAT/UMARF archive on tape</td>
</tr>
</tbody>
</table>

**9.3.5 DLR: SYNAER**

<table>
<thead>
<tr>
<th>Processing obligations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y3: Full reprocessing of 10yr dataset in 3 months</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Overall software architecture</th>
</tr>
</thead>
<tbody>
<tr>
<td>Major bugs were identified and corrected and the full error propagation between both sensors implemented in SYNAER within a science environment. For the processing several Linux clusters will be used (a total of 80 CPUs is available for the processing period),</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Memory and CPU budget</th>
</tr>
</thead>
<tbody>
<tr>
<td>With 80 CPUs preprocessing and processing of 10 years global data will take ~ 2 months.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Design standards, conventions and procedures</th>
</tr>
</thead>
<tbody>
<tr>
<td>Version control system for SYNAER retrieval algorithm: Application of DLR-wide “SVN” solution</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Internal interface</th>
</tr>
</thead>
<tbody>
<tr>
<td>Batch job (cshell-processing environment); condor parallel job</td>
</tr>
<tr>
<td>design</td>
</tr>
<tr>
<td>------------------------</td>
</tr>
<tr>
<td>Valid range of values</td>
</tr>
<tr>
<td>Data sources</td>
</tr>
</tbody>
</table>

**9.3.6 FMI: ADV ((A)ATSR)**

<table>
<thead>
<tr>
<th>Processing obligations</th>
<th>Y1 + Y2: Full reprocessing of 17 year global datasets in 3 months</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall software architecture</td>
<td>ADV has been further developed within a science environment. New approach for the k-ratio has been tested but not implemented, since the previous approach has been giving better results. The main focus is on cloud screening improvement. The technical transfer of ADV to a supercomputer has been completed. On the supercomputer, the operational processing scheduled for Years 2 and 3 will be conducted.</td>
</tr>
<tr>
<td>Memory and CPU budget</td>
<td>A supercomputer is a eowulf beowulf cluster running 64bit CentOS 5.9 linux environment. It has 32 compute nodes, each node has 2 intel xeon ES-2660 processors@ 2.2 GHz that supports hyper threading. Each processor has 8 cores, thus a total of 520 cores with Voltaire Infiniband interconnect. The computer interfaces with lustre file system, a high performance parallel distributed file system for large scale cluster computing that can provide fast, reliable data access and supports intensive IO processing from all nodes. For storage, StorNext File System a shared disk file system is used. Large files are shared without any network delays. The supercomputer has a dedicated 100TB of tape storage, and 60TB of disk storage.</td>
</tr>
<tr>
<td>Design standards, conventions and procedures</td>
<td>ADV retrieval is coded in C and Fortran 90: Usage of “gfortran”-compiler is envisaged. Currently, no version control system is used; Changes in the code are stored in additional documentation. Using SVN for version control and documentation control has been discussed and is advised for use by FMI ICT.</td>
</tr>
<tr>
<td>Internal interface design</td>
<td>Batch job / orchestration</td>
</tr>
<tr>
<td>Valid range of values</td>
<td>AOD, FMAOD at 0.55 µm, 0.65 µm, 0.87 µm (ocean only) and 1.6 µm: Valid range of values 0 … 5 (dimensionless)</td>
</tr>
<tr>
<td>Data sources</td>
<td>(A)ATSR Version 2.30_plume, years 1997-2012, is on ICARE</td>
</tr>
</tbody>
</table>
### 9.3.7 FMI: SLSTR

<table>
<thead>
<tr>
<th>Processing obligations</th>
<th>Y3: Reprocessing of 1yr global dataset in 2 months, if quality-controlled SLSTR data are available</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall software architecture</td>
<td>The SLSTR AOD retrieval algorithm (similar to AATSR ADV/ASV algorithm, adjusted to SLSTR instrument features) is currently under development. Several test runs have been performed.</td>
</tr>
<tr>
<td>Memory and CPU budget</td>
<td>Preparations for SLSTR retrieval is ongoing, no meaningful estimation can be given yet.</td>
</tr>
<tr>
<td>Design standards, conventions and procedures</td>
<td>Version control using Git with backup to GitHub publically maintained no-cost cloud service is envisaged. Main retrieval code is written in Fortran and processing chain scripting in Bash.</td>
</tr>
<tr>
<td>Internal interface design</td>
<td>Batch job</td>
</tr>
<tr>
<td>Valid range of values</td>
<td>AOD is constrained to 0…5 (dimensionless value)</td>
</tr>
<tr>
<td>Data sources</td>
<td>N/A yet</td>
</tr>
</tbody>
</table>

### 9.3.8 LOA/ICARE: GRASP

<table>
<thead>
<tr>
<th>Processing obligations</th>
<th>Produce POLDER regional output for up to four “diagnostic sites” time series (1200x1200 km2 over full PARASOL period) and one continental output (Africa, land only) for one year of PARASOL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall software architecture</td>
<td>The PARASOL/GRASP algorithm is developed jointly by LOA and ICARE at University of Lille. LOA is responsible for the science code and ICARE is responsible for the operational production framework. Currently an operational prototype of the code is running in the ICARE production system and is under evaluation.</td>
</tr>
<tr>
<td>Memory and CPU budget</td>
<td>GRASP is running on the ICARE production cluster, of which about 250 CPU cores and 500 GB of RAM can be dedicated to GRASP. With this configuration, one month’s worth of PARASOL data (global, land and ocean) can be processed in 13 days (i.e. 3200 core-days per month).</td>
</tr>
</tbody>
</table>
| Design standards, conventions and procedures | - Both the science code and the production framework are version-controlled with GIT  
- Some limited documentation is currently available |
## Production framework

Production framework composed of layers of Python scripts and C codes, science code in Fortran90 (compilers: gfortran and gcc)

### Internal interface design

- Scheduler and processing control interface integrated in information system (Oracle/PLSQL procedures)
- Fully automated procedures
- Processing codes handle exceptions and errors
- Nominal run configuration: 3 months/4 pixels per processing core

### Valid range of values

- AOD at 443 490 565 670 865 and 1020 nm (Total, fine and coarse), valid range of values 0 -10 (dimensionless)
- Angström exponent, valid range 0-3 (dimensionless)
- Lidar ratio (443 490 565 670 865 and 1020 nm), valid range 0-100 (dimensionless)
- Aerosol single scattering albedo (443 490 565 670 865 and 1020 nm), valid range 0-1 (dimensionless)

### Data sources

PARASOL Level-1 data provided by CNES and archived at ICARE

### 9.3.9 LMD: IASI

#### Processing obligations

Y2: Reprocessing of 10yr global dataset in 6 months

#### Overall software architecture

The IASI retrieval chain from LMD is already designed for operational processing within the aerosol_cci2. The retrieval is based on a Look-up Table (LUT) method. The computing of the LUT is performed once for all during pre-processing phase and is not supposed to be recomputed except if a change is made on the dust parametrization. The chart below describes the LMD processing chain:
**Memory and CPU budget**
Currently the processing requires 20gb of memory and 20hCPU for 1 month (for the two daily overpasses)

**Design standards, conventions and procedures**
IASI retrieval code is written in Fortran 90
Documentation is by comments within code, with major changes recorded at the level of ATBD updated versions.

**Internal interface design**
Batch job (bash-shell-processing environment)

**Valid range of values**
Dust AOD at 10µm and 0.55 µm:
Valid range of values 0 … 4 (dimensionless)

**Data sources**
L1c IASI spectra obtained from Ether (CNES and CNRS datacenter http://ether.ipsl.jussieu.fr/)

---

**9.3.10 RAL: (A)TSR (ORAC) + SLSTR**

**Processing obligations**
Y1 + Y2: Full reprocessing of 17-year global (A)TSR datasets in 3 months
Y3: Reprocessing of 1-year SLSTR global dataset in 6 months

**Overall software architecture**
The ORAC (A)ATSR aerosol retrieval scheme shares the same code base as the Cloud_cci “heritage channel” processing chain (know as the Community Cloud for Climate processor - CC4Cl). This code is a “community algorithm”, with the code being available to download the processor and registered developers from outside the core Oxford and RAL groups being able to contribute to the ongoing development of the code.

The ORAC processing chain is entirely written in Fortran, with
The processing chain relies on several external libraries, particularly for input and ancillary data reading:

- ecCodes: ECMWF data reading library.
- EMOS lib: ECMWF data regridding and interpolation library.
- EPR-API: ESA/Brookmann Consult library for reading ENVISAT format (A)ATSR L1b data.
- HDF libraries (HDF4, HDF-EOS, HDF5 and HDF5-EOS): Needed for reading NASA ancillary data. (Also used to read MODIS L1 data in Cloud_cci)
- JasPer library: needed to support JPEG compression in HDF libraries
- NetCDF-4 libraries: provides support for reading and writing data in NetCDF 3 or 4 format.
- RTTOV: Radiative transfer code used to calculate clear-sky transmissions for cloud retrievals (not used in aerosol_cci processing, but still required for compilation)
- SZIP and ZLib: Compression libraries required by HDF and NetCDF.

The aerosol_cci processing system consists of four stages:

1. Pre-processor (Fortran): Reads all required input and ancillary data, performs any required regridding and spatio-temporal interpolation. Outputs all ORAC processor inputs in NetCDF format
2. Processor (Fortran): Performs the retrieval itself, assuming a particular surface model and aerosol/cloud type. For Aerosol_cci processing this step is run 20 times: One for each of 10 aerosol models used, plus separately for land and ocean pixels.
3. Post-processor (Fortran): Combines the output of the processor into a single, speciated, 1 km product.
4. Aerosol_cci product generate (IDL): Averages the 1 km aerosol product onto the 10 km sinusoidal grid used for Aerosol_cci L2 products, and output in an Aerosol_cci format NetCDF file.

The processing system has been tested with SLSTR data, and
### Memory and CPU budget

ORAC is installed on the STFC “SCARF” computational facility and the JASMIN/CEMS “data super-cluster” facility provided by CEDA. Each of these systems provides on the order of 5000 processing cores (64bit Intel), with up to 16 GB of RAM per core. Both systems also have direct access to the CEDA (A)ATSR and ancillary data archives. Both JASMIN/CEMS and SCARF run a Red Hat Enterprise Linux system.

To produce a single orbit of L2 Aerosol_cci AATSR data, using 10 different aerosol types, requires approximately 7.6 hours of CPU time, and requires approximately 26 GB of temporary storage.

In practice the limiting factor in processing the full (A)ATSR (or SLSTR in the future) dataset is the temporary storage space needed by the processor running at full 1 km spatial resolution. A full processing of the 17-years of (A)ATSR data can be completed in approximately 6 weeks on JASMIN/CEMS.

### Design standards, conventions and procedures

The ORAC community code is managed with the subversion (SVN) program implemented within the TRAC system (see http://proj.badc.rl.ac.uk/orac/wiki). All code is Fortran 2008 compliant and common documentation and style formatting is used. A set of simple regression tests are used to validate code changes.

### Internal interface design

ORAC consists of three command-line programs. The pre and post-processors are run once each to prepare the input and output files for a single orbit or granule. The main processor is run once for each aerosol type to be investigated, enabling significant parallelisation. Processing is managed by a series of shell scripts to distribute the jobs into the JASMIN or SCARF batch queues.

### Valid range of values

AOD at 550 nm is constrained to the range 0.01 to 5.0. Effective radius at 550 nm is constrained to the range 0.01 to 10.0 µm.

### Data sources

Inputs are drawn from the CEMS archive, including calibrated ATSR-2 and AATSR observations and ECMWF reanalysis products. All SLSTR data will be mirrored as they become available.

### 9.3.11 SU: AATSR

#### Processing

Y1 + Y2:
<table>
<thead>
<tr>
<th>obligations</th>
<th>Full reprocessing of 17 year global datasets in 3 months</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall software architecture</td>
<td>The SU_AATSR code is developed within a science environment. An existing version of the code was used to generate the 17 year dataset for ATSR2/AATSR CCI phase I. A modified version of this code is being implemented on a new system, Blue Ice 2. In the duration of the project successfully migrated to the JASMIN data analysis environment of the Centre for Environmental Data Analysis (CEDA) This step enables the direct processing of (A)ATSR data from the CEDA / NEODC archive.</td>
</tr>
<tr>
<td>Memory and CPU budget</td>
<td>The JASMIN analysis platform provides a high performance distributed computing system with about 3000 individual CPU cores. It operates a Red Hat Linux system with scheduling mechanisms for parallel computation enabling the SU-ATSR retrieval to reprocess the 17 year archive in about 7 days.</td>
</tr>
<tr>
<td>Design standards, conventions and procedures</td>
<td>Version control using Git with backup to GitHub publically maintained no-cost cloud service. Main retrieval code is written in C, with preprocessing is written in JAVA, and processing chain scripting in Bash. Documentation is by comment within code, with major changes recorded at the level of ATBD updated versions.</td>
</tr>
<tr>
<td>Internal interface design</td>
<td>The retrieval system consists of a small number of command line tools providing preprocessing, cloud screening, AOD retrieval post processing and conversion to required output format. These tools are chained by shell scripts to perform the entire retrieval from L1b obrit files to L2 and L3 AOD results.</td>
</tr>
<tr>
<td>Valid range of values</td>
<td>AOD at 550 nm is constrained to the range 0.01 to 5.0.</td>
</tr>
<tr>
<td>Data sources</td>
<td>Version 6.1, drawn from CEDA archive server during processing.</td>
</tr>
</tbody>
</table>

9.3.12 SU: SLSTR

<table>
<thead>
<tr>
<th>Processing obligations</th>
<th>Y3: Processing of 1yr global dataset in 6 months</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall software architecture</td>
<td>The processing of AATSR has been ported to retrieve AOD from SLSTR L1b input files. The processing chain has already been implemented and runs stable on the JASMIN/LOTUS platform. The current performance enables processing of one month of input data in about 6 hours using parallel processing available on JASMIN /LOTUS.</td>
</tr>
</tbody>
</table>
### Memory and CPU budget

The JASMIN analysis platform provides a high performance distributed computing system with about 3000 individual CPU cores. It operates a Red Hat Linux system with scheduling mechanisms for parallel computation enabling the SU-SLSTR retrieval to reprocess one month in about 6 hours.

### Design standards, conventions and procedures

Version control using Git with backup to GitHub publically maintained no-cost cloud service is envisaged. Main retrieval code is written in C and processing chain scripting in Bash.

### Internal interface design

The retrieval system consists of a small number of command line tools providing preprocessing, cloud screening, AOD retrieval post processing and conversion to required output format. These tools are chained by shell scripts to perform the entire retrieval from L1b obrit files to L2 and L3 AOD results.

### Valid range of values

AOD will be constrained to 0…3.

### Data sources

Not applicable.

### SLSTR L1b input data will be provided directly from the data mirror on the CEDA archive.

### 9.3.13 ULB: IASI

#### Processing obligations

Y2: Reprocessing of 10yr global dataset in 6 months

#### Overall software architecture

IASI retrieval is currently further developed in science environment. In its current status it already allows a sufficient processing capacity for the round robin analysis which is scheduled for Year 1. A single modern PC covers all the needs for processing of the dataset as the algorithm is being specifically developed to be light on computational resources. Both the in and output data are stored on external disks.

#### Memory and CPU budget

The most time consuming step for the ULB algorithm are for the moment reading/reformatting of the input files, but nothing which cannot be handled with very reasonable hardware requirements. Estimated CPU time: 4 hours/month

#### Design standards, conventions and procedures

The code is entirely written in Matlab

#### Internal interface design

Batch Job

#### Valid range of values

AOD output is constrained
Data sources | Raw IASI L1C/L2 is preprocessed at Ether/LATMOS server (selected channels/L2 data are exported to smaller binary files)

### 9.4 Growing Amount of Input Data

With the growth of the amount of input products several parts of the Aerosol_CCI Production system have to grow. These parts are especially the processing and the archiving capacity. During the operational period of the system the needs for these two factors in the near future have to be checked repeatedly and additional processing capacity and archiving capacity has to be added to the Aerosol_CCI Production System. Precautions to be taken to arm against the growing amount of data are scalability in design of archiving and processing functions. In this chapter, the partners describe their specific mitigation strategies.

**DLR (all processing chains):**
DLR operates (and owns) a growing compute infrastructure, the “Geofarm”-cloud. Currently the complete hardware system has a capacity of more than 2000 CPUs and 23 TB RAM. Its harddisk storage capacity is in the order of 1.6 PB. However, the system is shared among a number of projects, so this huge system will not be available exclusively for a single project. Additional IT-systems (e.g. operated by our department) are also depicting the growing demand regarding CPU- and memory usage. As a consequence, from the technical point of view, the system performance can be considered as sufficient at the moment. However, new challenges are arising for the algorithm development: New approaches e.g. regarding parallel processing have to be taken into consideration in order to speed up the processing.

**FMI:**
Currently, the ADV/ASV production chain uses the FMI computing facilities described in Sec.9.3.6. We have reserved 5Tb RAM, which is enough to store several versions of the products produced with the ADV/ASV algorithm. Thus, the system performance can be considered as sufficient at the moment. The facilities available are also enough for the SLSTR AOD data production. However, the SLSTR data storage is under construction, which needs sufficient amount of memory (harddisk and RAM) increasing yearly. Currently, we use EUMETSAT Copernicus Online Data Access ([https://coda.eumetsat.int/#/home](https://coda.eumetsat.int/#/home)) as SLSTR data base for AOD production.

**RAL/ORAC/SU:**
Two upgrades are planned and funded for the JASMIN cluster over the coming years to meet the increased demand for Earth observation data in the UK. This capacity will be more than sufficient to cover all products expected to be produced with the Swansea (SU) and ORAC algorithm. Currently all datasets are historical (satellite has stopped operations) with the exception of the IASI instrument. Thus no major growth of data has yet to be faced. However, with the launch of Sentinel-3 larger data amounts will become available after the end of the commissioning phase during 2016. However, full mission processing will only be activated after the end of the current project, so that preparations for inclusion of growing data amounts can take place in phase 3 of the current project.

**BIRA/GOMOS:**
Since Envisat was lost in 2011, there is no data production from new measurements, and currently, the amount of input data is rather stable. At this stage, we dispose of a large GOMOS dataset covering the Level 1B transmission product (GOM_TRA_1P) and the residual extinction (Level 2 product GOM_EXT_2P) which is the most important dataset for the processing of Aerosol_CCI products. It
seems that these products are sufficient, at this stage, to allow all considered further algorithmic developments in the framework of CCI. We are also not aware on ESA’s intentions to reprocess the GOMOS data, which would require the acquisition of an additional version. Hence, we expect no storage problem of input data in the future.

BIRA/IASI:
BIRA-IASB has acquired together with the Royal Observatory of Belgium and the Royal Meteorological Institute of Belgium a High Performance Computing system (2 HPC clusters 44+12(fat) nodes * 2 Xeon E5-2680V3 CPU (12core, 2.5GHz); in total 2688 CPU). This cluster uses PBS Pro scheduling system, and is used by a number of different projects from the 3 institutes. At this point, the cluster has revealed to be sufficiently available and the processing has not been an issue. For future processings, additional work is done on the algorithm to make it less CPU-consuming, and therefore no processing issue is foreseen.

The primary data storage infrastructure for science data at BIRA-IASB is currently based on 2 fully redundant NAS storage systems from Hitachi Data Systems (HDS HNAS 4080 clusters with 2 HNAS nodes each, 4 HDS VSP G400 Raid systems as disk back-end). These 2 storage clusters offer a combined online storage capacity of 2.4 Petabyte. Data can be stored either redundantly, with a copy in both data centers for rapid disaster recovery, or with an offline copy on the tape archiving system for backup purposes. All systems are fully redundant in all critical components and are situated in independent data centers with redundant power and cooling systems.

ULB:
Raw IASI L1C/L2 is preprocessed at LATMOS server. In this process, selected spectral channels and Eumetsat L2 data are exported to smaller binary files; which are then transferred via ftp onto external harddisk at the ULB. LATMOS relies on the data processing and storage of the PC cluster ciclad of the ESPRI mesocenter in Paris. This system is shared among a number of projects, so it is not available exclusively for a single project. It provides 1300 CPUs with 4G per core and 900T for the storage. IASI data already occupy 200T.

ICARE:
The PARASOL mission ended in 2013, so the POLDER/PARASOL input data set for GRASP will not grow any longer. The total volume of the POLDER/PARASOL input data set is 30 TBytes

Wrap up / conclusion:
All partners have strategies to cope with the growing amount of EO data at this current stage. In most cases, hardware is extended or access to high performance computer systems is established. However, a prediction for periods extending more than 2-3 years into the future is hard to give — especially when considering the typical 3-5 years lifetime of IT hardware.

9.5 Availability of more consistent ancillary data

During the continuous evolution of the Aerosol retrieval algorithms, new ancillary data may become available. In this chapter, the partners describe all aspects covering the identification of additional data sources, their implementation and their application during the processing.

RAL/ORAC:
The ORAC algorithm can draw numerous inputs constraining the spectral properties of the surface. To maximise the independence of the cci products, these are currently minimal, such as assuming constant chlorophyll concentrations. The quality of the retrieval can usually be improved by strongly
constraining the surface, provided the constraint is accurate and drawn from information independent of ORAC’s other inputs. As such, an independent ocean colour product could be very useful, but this would need to be extensively explored before application. A common land-sea mask provided from Land_cover_cci is recommended to be used in all CCI systems. Its potential impact is hard to judge before the implementation, but it has been preliminarily assessed for use in Aerosol_cci and no obvious problems were identified. The implementation is planned to be conducted in 2016.

BIRA/GOMOS:
Concerning GOMOS, as mentioned in chapter 9.4, as far as we are well informed, no further reprocessing is expected from GOMOS data by ESA. Since we are using ancillary data provided with the GOMOS dataset, we don’t expect to have to consider other sources for the ancillary data.
10 MATCHING AEROSOL_CCI DOCUMENTATION TO ECSS

In this section, the relationship between the current Aerosol_cci documentation with the ECSS (European Cooperation for Space Standardization, http://www.ecss.nl) standard is investigated. The goal is to determine, whether all relevant aspects of ECSS are covered by the currently existing Aerosol_cci documentation. As a result, potentially missing aspects of the documentation can be identified.

For this analysis it has to be kept in mind, that ECSS is focused on space system development and not on supporting science driven research projects. Therefore, no one-by-one relationship between project specific documents and pure ECSS documents is to be expected.

ECSS targets all aspects of space system engineering, such as project management, quality assurance, engineering and sustainability, see Figure 10-1.

![Figure 10-1: Branches and disciplines of the ECSS standard. The E-40-discipline (Software Engineering) which is most relevant for Aerosol_cci is highlighted.](http://www.ecss.nl/)

(As of 6 May 2014)
Since Aerosol_cci targets the definition and implementation of a software system, therefore “E-40” and related disciplines are to be regarded as relevant for the Aerosol_cci project. ECSS-E-ST-40 [RD5] covers life cycle processes such as development, operation, maintenance, verification, validation and review, see [RD6]. The related ECSS-Q-ST-80 standard targets “quality-related” supporting and organizational life cycles such as “improvement”, “training” of staff members, “problem resolution” and “quality assurance”.

![Figure 10-2: Processes covered by ECSS E-ST-40 and related standards](image)

![Figure 10-3: Structure of the ECSS E-40 discipline (taken from [RD5], p. 31). Numbering refers to ECSS-E-ST-40C document.](image)
Prerequisite of ECSS is the existence of a customer ⇔ supplier-relationship, see Figure 10-4. Accordingly any system engineering process starts with the customer defining and providing “software related system requirements” (process 5.2). Thereafter, ECSS defines numerous activities and processes representing the full software lifecycle. ECSS processes cover all aspects of system requirement, architecture and engineering, software design and implementation, validation and verification, software maintenance and software operation. The numbering of the individual processes refers to Figure 10-3, providing an overview of the different processes targeted by ECSS-E-ST-40.

Interaction between customer and supplier is organised by scheduled “reviews”. The basis for these reviews are files (e.g. RB, DJF, MGT, TS, etc.) which contain dedicated documents (e.g. “SSS - Software system specification”, “IRD – Software interface requirements document”, etc.). Titles and content of the needed documents are listed in [RD5], annex “A”. A detailed description of the content (“expected response”) is given in annex “B” to “P”.

Similar to the ECSS theoretical concept, Aerosol_cci Phase 2 can also be considered as a customer ⇔ supplier relationship, see Figure 10-5. At first, the user (the science community) specifies requirements (URD – User Requirements Document). Then the supplier (the project consortium) documents its approaches in providing satellite derived aerosol information and products. Product specification, retrieval algorithms, validation and product provision are documented by the Product Specification Document, ATBDs, System Specification Document (this document) and many further. Finally, the processed results (aerosol information), validation reports, user guides and additional documentation are made available to the customer (the user community).
Figure 10-4: Overview of the ECSS-software lifecycle process. Numbering refers to the [RD5]. Viewgraph was taken from [RD5] and partly modified.

In Figure 10-4, Aerosol_cci Phase 2 documents are placed close to their corresponding ECSS processes. However, only those documents which are strongly related to scientific work packages have been considered.

Project-specific documents which have not been considered in this analysis are:

- DARD (Data Access Requirements Document),
- User Workshop Summary Report,
- SLSTR NRT strategy report,
- ATSR-AERONET matchup tool, ATSR ensemble test report,
- CRDP (Climate Research Data Package),
- CAR (Climate Assessment Report).
The following section focuses the question whether the current Aerosol_cci Phase 2 documentation covers all important aspects of ECSS. Therefore, a document requirement matrix was developed (see Figure 10-6).

The document requirement matrix compares ECSS files and documents (upper row) with Aerosol_cci Phase 2 documents (left column). Documents which are specific for Aerosol_cci are indicated. This analysis concentrates on system engineering aspects in a science environment and therefore considers only a subset of 7 ECSS documents which has been considered as relevant. However, the relevant standards ECSS-E-ST-40 and ECESS-Q-ST-80 require up to 60 different documents (see the Document Requirement List (DRL) indicated in [RD5], page 78-82).

Please not, that a perfect one-by-one relationship between project specific documents and pure ECSS documents is not to be expected.
Figure 10-6: Document requirements matrix. Matches between Aerosol_cci Phase 2 documentation and ECSS are indicated with “X”.

In general, a successful matching of Aerosol_cci and ECSS documents is possible. Each row and each column contains at least one match.

However, based on the experiences during the system engineering telcos and the meeting conducted so far, DLR proposes one additional chapter within the Aerosol_cci-specific SSD (System Specification Document) which is equivalent to ECSS’s “Software Design Document (SDD)”. The SDD is part of
ECSS’s “Design Definition File (DDF)”. In the scope of Aerosol_cci Phase 2, this additional section should be named “Description of decentralized processing Chains”.

The proposed subsections of the “Description of decentralized processing Chains” should be named as follows:

- Processing Chain X
  - Overall software architecture
  - Memory and CPU budget
  - Design standards, conventions and procedures
  - Internal interface design

The covered aspects are compliant to compare to ECSS-E-ST-40C, annex F, “Software Design Document”.

The major goal of this additional section is to achieve transparency among the project partners about the individual computational effort which has to be undertaken by the specific partner to meet his processing obligations. It will help determining whether the processing plans of the individual teams can be considered as realistic. Moreover it facilitates a potentially necessary handover procedure during or after the project.

The aspects of the new chapter are currently covered by several subsections of this document, named “Concrete Measures”.
11 FUNDING MECHANISMS

This short chapter provides a beginning of this highly relevant discussion, which needs further extension in the future.

The easiest means of funding would be a regular budget provided from an international or European body, based on which the science leader could be selected and appointed for several years and then issue contracts to all relevant experts with shorter duration depending on their tasks and the need for their expertise. The funding source should then also include the resources for maintenance and operation of the operational ECV production system. Additional funding for use cases, larger algorithm development activities or larger system upgrades could be sought by proposals to different funding bodies.

An alternative funding mechanism would be an international or European funding for just the operational system and funding for the algorithm work via proposals. However, this would severely impact the continuity and strength of the algorithm work and probably lead to failure of the overarching goal that the production system should be led by scientific experts.

The total cost, which has to be funded, is calculated in section 6.4:

**Total cost**

<table>
<thead>
<tr>
<th>Description</th>
<th>Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>Start up investment</td>
<td>540k€ (minimum)</td>
</tr>
<tr>
<td>System operations cost per year</td>
<td>331k€ / year</td>
</tr>
<tr>
<td>Science team cost per year</td>
<td>400k€ / year</td>
</tr>
</tbody>
</table>
12 REQUIREMENT MATCHING MATRIX

The following table provides the matching of requirements in the system requirements document [RD4] to the sections of this system specification document.

<table>
<thead>
<tr>
<th>ID</th>
<th>Type</th>
<th>Description</th>
<th>SSD section</th>
</tr>
</thead>
<tbody>
<tr>
<td>SR-GE-0010</td>
<td>General</td>
<td>The system shall be able to produce aerosol ECV products as specified in the PSD. As the PSD will evolve over the 15 years the system must be scalable to this evolution.</td>
<td>2</td>
</tr>
<tr>
<td>SR-GE-0020</td>
<td>General</td>
<td>The system shall be able to produce aerosol ECV products from the input datasets specified in the DARD. As the DARD will evolve over the 15 years the system must be scalable to this evolution.</td>
<td>2</td>
</tr>
<tr>
<td>SR-GE-0030</td>
<td>General</td>
<td>The system shall be able to reprocess the entire aerosol ECV products when new input dataset versions become available. The reprocessing shall be organised intelligently, so that a new input dataset version for one sensor becoming available invokes exactly but no more that the reprocessing for all products (directly derived or merged with other sensor input) where this respective sensor contributes</td>
<td>2, 5.4</td>
</tr>
<tr>
<td>SR-GE-0040</td>
<td>General</td>
<td>The system shall support the repeated validation and inter-comparison of the aerosol ECV products produced with external reference and comparison datasets.</td>
<td>3.1, 3.2</td>
</tr>
<tr>
<td>SR-GE-0041</td>
<td>General</td>
<td>The ECV products produced by the aerosol_cci system must have scientific credibility from the users’ point of view, e.g. by close interaction with end users, by performing regular inter-comparisons and validation against datasets that the users already trust and by submitting aerosol_cci results in peer reviewed journals that the users read and respect.</td>
<td>3.1</td>
</tr>
<tr>
<td>SR-GE-0042</td>
<td>General</td>
<td>The team performing validation and inter-comparison tasks has to be independent from algorithm developers.</td>
<td>3.1, 3.2</td>
</tr>
<tr>
<td>SR-GE-0050</td>
<td>General</td>
<td>In order to ensure highest quality and credibility of ECV products the system shall be developed, maintained and operated under the control of an Aerosol Science Team containing internationally recognised aerosol experts in the following</td>
<td>2, 3.2</td>
</tr>
<tr>
<td>SR-GE-0051</td>
<td>General</td>
<td>The system must be open to assure free and open product availability and transparency of algorithm definitions.</td>
<td>3.2</td>
</tr>
<tr>
<td>SR-GE-0080</td>
<td>General</td>
<td>The system shall be able to easily integrate new algorithm versions or completely new algorithms for limited test purposes (to qualify algorithm improvements) and for complete reprocessing.</td>
<td>3.1, 4.3.1</td>
</tr>
<tr>
<td>SR-GE-0090</td>
<td>General</td>
<td>In order to achieve timely production of ECV products the initial version of the system must rely on existing algorithms and their gradual upgrades.</td>
<td>2, 9.1</td>
</tr>
<tr>
<td>SR-GE-0100</td>
<td>General</td>
<td>The system shall contain internal coordination mechanisms to assure the quality of algorithm development. This includes validation and inter-comparison expertise and agreement on algorithm development priorities between algorithm experts and validation experts / user representatives.</td>
<td>3.1</td>
</tr>
<tr>
<td>SR-GE-0110</td>
<td>General</td>
<td>The system shall contain external coordination mechanisms to assure international coordination, effective interfaces with instrument calibration teams, and close links with other ECV teams (consistency of products).</td>
<td>3.1</td>
</tr>
<tr>
<td>SR-GE-0120</td>
<td>General</td>
<td>The internal and external coordination mechanisms shall be flexible to allow adjustment based on experiences made throughout the continued work.</td>
<td>3.1</td>
</tr>
<tr>
<td>SR-GE-0130</td>
<td>General</td>
<td>The system shall contain an archive and dissemination facility including a repository for documentation on algorithms, validation and utilization of ECV products</td>
<td>3.1, 3.2</td>
</tr>
<tr>
<td>SR-FU-0141</td>
<td>Functional</td>
<td>aerosol_cci system shall be flexible in a way that enables the incorporation of data from a new sensor as a new data source. The current state of the DARD will reflect the need for new data</td>
<td>2, 4.3.3</td>
</tr>
<tr>
<td>SR-FU-0150</td>
<td>Functional</td>
<td>aerosol_cci system shall maximise the consistency of ancillary data used with other datasets used by the end users (e.g. from CCI Clouds and Fire projects, explicitly prioritized in the user</td>
<td>3.1</td>
</tr>
<tr>
<td>Documentation</td>
<td>The system shall assess competitiveness or complementarity of any new aerosol earth observation product to the Aerosol CCI product and this should be documented and proven at the end of the product development and evaluation cycle.</td>
<td>3.1, 4.3.2</td>
<td></td>
</tr>
<tr>
<td>----------------</td>
<td>-------------------------------------------------------------------------------------------------</td>
<td>------------</td>
<td></td>
</tr>
<tr>
<td>Operational</td>
<td>The system shall support availability of the Aerosol CCI products which allows planning of their use. Accordingly there is a need to publish a schedule when the Aerosol CCI products will become available.</td>
<td>4.1</td>
<td></td>
</tr>
<tr>
<td>Quality</td>
<td>The system shall assure swift and continuous implementation of cutting edge research results (new algorithms and upgrades).</td>
<td>3.1, 4.3.1</td>
<td></td>
</tr>
<tr>
<td>Functional</td>
<td>The system shall provide easy (“few-click-away”) data availability via an open website or ftp account to support the capacity of users to access satellite data.</td>
<td>5.2, 5.3</td>
<td></td>
</tr>
<tr>
<td>Functional</td>
<td>The system shall support users to explore the ECV products (e.g. via AEROCOM, MPI, ICARE tools).</td>
<td>5.6</td>
<td></td>
</tr>
<tr>
<td>Operational</td>
<td>The science Community and users shall be able to influence reprocessing frequency, parameters and input data</td>
<td>3.2</td>
<td></td>
</tr>
<tr>
<td>Operational</td>
<td>The science Community and users shall have the opportunity to include new data sources into aerosol_cci system</td>
<td>4.3.1, 4.3.3</td>
<td></td>
</tr>
<tr>
<td>Operational</td>
<td>The science Community and users shall have the opportunity to include new processors into aerosol_cci system</td>
<td>4.3.1</td>
<td></td>
</tr>
<tr>
<td>Quality</td>
<td>The science Community and users shall interface with the system for quality feedback</td>
<td>3.2</td>
<td></td>
</tr>
<tr>
<td>Operational</td>
<td>The aerosol_cci system shall include an archive which is able to collect, store, backup and deliver Input products for processing (see Figure 3-1, regarding products)</td>
<td>5.6</td>
<td></td>
</tr>
<tr>
<td>Operational</td>
<td>The aerosol_cci system shall include an input data archive which incorporates the needed ancillary input data for all used processors</td>
<td>5.6</td>
<td></td>
</tr>
<tr>
<td>ID</td>
<td>Type</td>
<td>Description</td>
<td>Page</td>
</tr>
<tr>
<td>---------</td>
<td>------------</td>
<td>-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>SR-FU-0290</td>
<td>Functional</td>
<td>The aerosol_cci system shall have ingestion abilities to insert necessary input satellite data for the ECV production</td>
<td>5.6</td>
</tr>
<tr>
<td>SR-FU-0300</td>
<td>Functional</td>
<td>The aerosol_cci system shall have ingestion ability to insert ancillary data</td>
<td>5.6</td>
</tr>
<tr>
<td>SR-OP-0310</td>
<td>Operational</td>
<td>The aerosol_cci system shall have the ability to reprocess all available input data every two years within 6 months. Additionally, the ECV product time series needs to be extended for another annual dataset due to the fact that new satellite data has been acquired.</td>
<td>7.2</td>
</tr>
<tr>
<td>SR-OP-0320</td>
<td>Operational</td>
<td>The aerosol_cci system shall be designed for an expected life time of 15 years</td>
<td>9.2</td>
</tr>
<tr>
<td>SR-OP-0330</td>
<td>Operational</td>
<td>The aerosol_cci system shall be implemented in a way which allows to increase the available processing power easily</td>
<td>9.2</td>
</tr>
<tr>
<td>SR-OP-0340</td>
<td>Operational</td>
<td>The aerosol_cci system shall include an archive which is capable to store all expected output data sets / be extendable to handle the amount of output data created during the 15 year life time</td>
<td>7.3, 9.2</td>
</tr>
<tr>
<td>SR-DO-0341</td>
<td>Documentation</td>
<td>Each product created by the aerosol CCI system shall be described within a release note which describes at least input products, ancillary products and algorithms used to create the output products and highlights the benefits and added value of this new product.</td>
<td>5.6</td>
</tr>
<tr>
<td>SR-DO-0342</td>
<td>Documentation</td>
<td>Each product created by the aerosol CCI system and its recommended use shall be described in more detail within a product user guide highlighting limitations of its use.</td>
<td>5.6</td>
</tr>
<tr>
<td>SR-DO-0360</td>
<td>Documentation</td>
<td>To introduce operators into the operation of the aerosol cci system an operator manual is needed which includes the operations needed to conduct the operational scenarios presented within this chapter</td>
<td>5.2</td>
</tr>
<tr>
<td>SR-FU-0360</td>
<td>Functional</td>
<td>The aerosol_cci system shall be able to perform a processing of the complete Aerosol ECV products</td>
<td>5.4</td>
</tr>
<tr>
<td>SR-QU-0370</td>
<td>Quality</td>
<td>The aerosol_cci-System shall perform an automatic quality check, inter-comparison and validation against available validation tools for each ECV</td>
<td>5.5</td>
</tr>
</tbody>
</table>
### aerosol_cci Phase 2
System Specification
Document issue 3

<table>
<thead>
<tr>
<th>Reference</th>
<th>Type</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>SR-DO-371</td>
<td>Documentation</td>
<td>Algorithms shall be described in ATBDs which are available for the end users. This document shall be endorsed by the science team on each major algorithm revision</td>
<td>5.6</td>
</tr>
<tr>
<td>SR-DO-372</td>
<td>Documentation</td>
<td>A validation report shall be produced for every processing run. The report shall be created by validation experts and endorsed by the science team</td>
<td>5.5</td>
</tr>
<tr>
<td>SR-DO-373</td>
<td>Documentation</td>
<td>An inter-comparison report shall be created by inter-comparison experts and endorsed by the science team</td>
<td>5.5</td>
</tr>
<tr>
<td>SR-DO-374</td>
<td>Documentation</td>
<td>A development priority list and work plan shall be created by the science team with the inputs of the end users.</td>
<td>4.3.1</td>
</tr>
<tr>
<td>SR-DO-375</td>
<td>Documentation</td>
<td>User requirements shall be described in a URD. This document shall be endorsed by the science team on bi-annual basis or release of major new user requirements (e.g. GCOS) to initiate the next algorithm development and production cycle.</td>
<td>3.2</td>
</tr>
<tr>
<td>SR-DO-376</td>
<td>Documentation</td>
<td>Product specifications shall be described in a PSD. This document shall be endorsed by the science team based on the URD on bi-annual basis to initiate the next algorithm development and production cycle.</td>
<td>3.2</td>
</tr>
<tr>
<td>SR-DO-377</td>
<td>Documentation</td>
<td>Data access requirements shall be described in a DARD. This document shall be endorsed by the science team based on the URD on bi-annual basis to initiate the next algorithm development and production cycle.</td>
<td>3.2</td>
</tr>
<tr>
<td>SR-DO-378</td>
<td>Documentation</td>
<td>An error characterization report (ECR) shall document the summary of findings on product uncertainties. It shall be endorsed by the science team for every completed processing run.</td>
<td>5.5</td>
</tr>
<tr>
<td>SR-DO-379</td>
<td>Documentation</td>
<td>The science team shall publish a collection of utilizations made and reported feedback by end users of the ECV products on bi-annual basis. Wherever possible users shall be encouraged to publish these results in peer-reviewed journals.</td>
<td>3.2</td>
</tr>
<tr>
<td>SR-OP-0390</td>
<td>Operational</td>
<td>New Versions of data from the DARD shall lead to a reprocessing of the products specified in the PSD</td>
<td>5.4</td>
</tr>
<tr>
<td>Code</td>
<td>Type</td>
<td>Description</td>
<td>Section</td>
</tr>
<tr>
<td>-----------</td>
<td>--------------</td>
<td>-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
<td>---------</td>
</tr>
<tr>
<td>SR-OP-0410</td>
<td>Functional</td>
<td>The aerosol_cci system shall be able to extract inter-comparison information out of comparable products to generate diagnostic information for validation and inter-comparison experts.</td>
<td>5.6</td>
</tr>
<tr>
<td>SR-OP-0460</td>
<td>Interface</td>
<td>Climate research ECV users shall have the opportunity to give feedback regarding ECV products quality.</td>
<td>3.2</td>
</tr>
<tr>
<td>SR-OP-0470</td>
<td>Interface</td>
<td>Experts for product validation and inter-comparison shall be consulted to validate ECV product quality.</td>
<td>4.1,5.5</td>
</tr>
<tr>
<td>SR-OP-0480</td>
<td>Interface</td>
<td>Feedback received from climate research ECV users and product validation and inter-comparison experts shall influence development priorities of the algorithm implementations.</td>
<td>3.2</td>
</tr>
<tr>
<td>SR-FU-0490</td>
<td>Functional</td>
<td>The aerosol_cci system shall be able to deliver already created data even if a running processing assignment is not already completed.</td>
<td>5.2</td>
</tr>
<tr>
<td>SR-FU-0500</td>
<td>Functional</td>
<td>The aerosol_cci system shall be able to deliver products online.</td>
<td>5.3</td>
</tr>
<tr>
<td>SR-OP-0380</td>
<td>Operational</td>
<td>The System shall keep intermediate products of the most recent ECV products.</td>
<td>5.4</td>
</tr>
<tr>
<td>SR-OP-0400</td>
<td>Interface</td>
<td>Validation and inter-comparison experts shall have the possibility to request reprocessing of a specific time series of a ECV product to validate results of an algorithmic change.</td>
<td>4.3</td>
</tr>
<tr>
<td>SR-OP-0420</td>
<td>Operational</td>
<td>The aerosol_cci system shall be modular in a way that allows the easy exchange of a processing module which includes the algorithm implementation.</td>
<td>4.3.1</td>
</tr>
<tr>
<td>SR-OP-0430</td>
<td>Operational</td>
<td>The aerosol_cci system shall be modular in a way that allows to add a new sensor as input to the workflows.</td>
<td>4.3.3</td>
</tr>
<tr>
<td>SR-DO-0440</td>
<td>Documentation</td>
<td>A manual has to be available which describes the steps to be performed to exchange the processing algorithm module.</td>
<td>4.3.2</td>
</tr>
<tr>
<td>SR-QU-0450</td>
<td>Quality</td>
<td>The characteristics of processing algorithms have to be annotated within the products meta data.</td>
<td>5.6</td>
</tr>
<tr>
<td>SR-OP-0510</td>
<td>Operational</td>
<td>The aerosol_cci system shall be able to store input products with a size of about 30TB data per year of</td>
<td>7.3</td>
</tr>
<tr>
<td>Requirement ID</td>
<td>Type</td>
<td>Description</td>
<td>Section</td>
</tr>
<tr>
<td>----------------</td>
<td>-------</td>
<td>-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
<td>---------</td>
</tr>
<tr>
<td>SR-OP-0520</td>
<td>Operational</td>
<td>The aerosol_cci system shall be able to store output products with a ratio of 10% to the input products for each reprocessing run.</td>
<td>7.3</td>
</tr>
<tr>
<td>SR-OP-0540</td>
<td>Operational</td>
<td>To produce the currently addressed two years of ECV the processing power equivalent of three 8 core x86 machines is needed. This requirement is only an initial requirement and not necessarily valid during the whole system life.</td>
<td>7.2</td>
</tr>
<tr>
<td>SR-OP-0541</td>
<td>Operational</td>
<td>Each processor thread on a processor core has to be able to allocate 2 GB RAM, That means an 8 core machine has to be equipped with more than 16 GB of RAM. This requirement is only an initial requirement and not necessarily valid during the whole system life.</td>
<td>7.2</td>
</tr>
<tr>
<td>SR-OP-542</td>
<td>Operational</td>
<td>6TB online storage shall be available to store input, intermediate and output products for one month of sensor data for 6 sensors. This requirement is only an initial requirement and not necessarily valid during the whole system life.</td>
<td>7.2</td>
</tr>
<tr>
<td>SR-OP-0550</td>
<td>Operational</td>
<td>To store the needed input and output data currently addressed in the DARD two years of production, the aerosol_cci system shall have the ability to store 66TB. This requirement is only an initial requirement and not valid during the whole system life.</td>
<td>7.3</td>
</tr>
<tr>
<td>SR-OP-0551</td>
<td>Operational</td>
<td>The aerosol_cci system shall be able to store intermediate products with a size of 2 times the size of the input products for each reprocessing run, this means 120 TB of intermediate products. This requirement still needs to be further validated!</td>
<td>7.3</td>
</tr>
<tr>
<td>SR-OP-0552</td>
<td>Operational</td>
<td>All requirements affecting sizing of the system have to be re-evaluated every 6 months to react fast enough to change requirements regarding • Input data size</td>
<td>9.2</td>
</tr>
</tbody>
</table>
### SR-OP-0560: Operational

To deliver the ECV data produced from the input data addressed in the DARD to the estimated 30 users in the user groups addressed in the URD in full resolution within 0.5 months the aerosol_cci system has to disseminate 180 TB data in 0.5 months which results in a data rate of ~150 MByte/Second.

## End of the document