ESA Climate Change Initiative CCI+

System Specification Document

Issue 3 / Revision 0

07 June 2021

Deliverable No.:	D3.2
ESA Contract No.:	4000124098/18/I-NB
Science Lead & Prime:	Thomas Nagler, ENVEO IT GmbH, thomas.nagler@enveo.at
Technical Officer:	Anna-Maria Trofaier, ESA-ECSAT, anna.maria.trofaier@esa.int

To be cited as:

Wiesmann, A., K. Luojus, M. Moisander, G. Schwaizer, M. Hetzenecker, T. Nagler, K. Naegeli, R. Solberg and M. Takala (2021) ESA CCI+ Snow ECV: System Specification Document, version 3.0, June 2021.

	Name	*Date*
Checked by	Gabriele Schwaizer / ENVEO, Project Manager	26.05.2021
Authorized by	Thomas Nagler / ENVEO, Science Leader	27.05.2021
Accepted by	Anna Maria Trofaier / ECSAT, ESA Technical Officer	07.06.2021

This document is not signed. It is provided as electronic copy.

ESA STUDY CONTRACT REPORT

ESA Contract No: 4000124098/18/I-NB	**Subject: ESA Climate Change Initiative CCI+ - SNOW** System Specification Document		**Contractor:** ENVEO
ESA CR ( ) No:	**Star Code:**	**No of volumes:** 3 **This is volume no:** 3	**Contractor's Ref:** Deliverable D3.2

Abstract:

The European Space Agency (ESA) Climate Change Initiative aims to generate high quality Essential Climate Variables (ECVs) derived from long-term satellite data records to meet the needs of climate research and monitoring activities.

This document outlines the system specifications for the *snow_cci* processing capable of producing the *snow_cci* ECV (snow cover fraction – SCF; snow water equivalent – SWE). This document provides details on the *snow_cci* processing system, system scenarios and workflows. Furthermore, it discusses the functional design and the development and lifecycle strategies and gives a detailed traceability matrix with the system requirements.

Authors: Andreas Wiesmann;Kari Luojus;Mikko Moisander;Gabriele Schwaizer;Markus Hetzenecker;Thomas Nagler;Kathrin Naegeli;Mathias Takala;Rune Solberg

ESA Study Manager: Anna Maria Trofaier / ECSAT	**ESA Budget Heading:**

Document Change Record

Version	Date	Changes	Originator
1.0 draft	16 / 09 / 2019	First version	Wiesmann et al.
1.0	02 / 10 / 2019	First version approved by ESA	Schwaizer, G.
2.0 draft	04 / 06 / 2020	Second version	Wiesmann et al.
2.0	25 / 06 / 2020	Second version approved by ESA	Schwaizer, G.
3.0 draft	25 / 05 / 2021	Third version	Wiesmann et al.
3.0	07 / 06 / 2021	Third version approved by ESA	Schwaizer, G.

TABLE OF CONTENTS

1. INTRODUCTION

The European Space Agency (ESA) Climate Change Initiative aims to generate high quality Essential Climate Variables (ECVs) derived from long-term satellite data records to meet the needs of climate research and monitoring activities, including the detection of variability and trends, climate modelling, and aspects of hydrology and meteorology.

1.1. Purpose

This document is deliverable D3.2 System Specification Document (SSD) of the *snow_cci* project requested in the Statement of Work (SoW) [AD-1]. The SSD incorporates the requirements described in the System Requirements Document (SRD) [RD-4] and specifies the characteristics of an operational ECV production system from a system engineering point of view.

1.2. Scope

The system design is based on experience developed and applied in previous projects such as ESA DUE GlobSnow. The prototypes are referenced in this document. The processing system (PS) is further developed in the years to come. Consequently, this SSD is a living document and will be complemented when necessary.

1.3. Document Structure

Section 2 gives an overview of the *snow_cci* processing system. It describes its purpose and intended use as well as the main requirements, functions, and components. Section 3 shows the main operational scenarios. Section 4 highlights the functional design from different perspectives, the users, system operators and developers view. Section 5 summarises information about the system life cycle design, implementation and maintenance costs and performance. Section 6 gives the system requirements of [RD-4] traceability and evolution.

1.4. Document Status

This document is based on version 3.0 of the Data Access Requirements Document (DARD), version 3.0 of the Product Specification Document (PSD), and version 3.0 of the User Requirements Document (URD); refinement of this document will be necessary of catchment of future versions of these documents.

1.5. Applicable and Reference Documents

[AD-1] Climate Change Initiative Extension (CCI+) Phase 1 – New Essential Climate Variables (2017). Statement of Work (Sow) v1.3, ESA-CCI-PRGM-EOPS-SW-17-0032.

[AD-2] ESA CCI Data Engineering Working Group (2019). CCI Data Standards, v2.1. CCI-PRGM-EOPS-TN-13-0009. http://cci.esa.int/sites/default/files/CCIDataStandards_v2-1_CCI-PRGM-EOPS-TN-13-0009.pdf

[RD-1] Derksen, C. and Nagler, T. (2021). ESA CCI+ Snow ECV: User Requirements Document, v3.0.

[RD-2] Derksen, C., Wiesmann, A., Nagler, T., Schwaizer, G. (2021). ESA CCI+ Snow ECV: Product Specification Document, v3.0

[RD-3] Wunderle, S., Nägeli, K., Schwaizer, G., Nagler, T., Marin, C., Notarnicola, C., Derksen, C., Luojus, K., Metsämäki, S., Solberg, R. (2021). ESA CCI+ Snow ECV: Data Access Requirements Document, v3.0.

[RD-4] Wiesmann, A., Hetzenecker, M., Schwaizer, G., Nagler, T., Takala, M., Luojus, K. (2021). ESA CCI+ Snow ECV: System Requirements Document, v3.0.

[RD-5] Schwaizer, G., Ossowska, J., Nagler, T., Luojus, K., Nägeli, K., Wunderle, S., Metsämäki, S. (2020). ESA CCI+ Snow ECV: Product Validation and Algorithm Selection Report, v2.0.

[RD-6] Schwaizer, G., Luojus, K., Nägeli, K., Wunderle, S. (2020). ESA CCI+ Snow ECV: Algorithm Theoretical Basis Document, v2.0.

[RD-7] Schwaizer, G., Luojus, K., Nägeli, K., Wunderle, S. (2020). ESA CCI+ Snow ECV: Algorithm Development Plan, v2.0.

[RD-8] Notarnicola, C., Marin, C., Schwaizer, G., Nagler, T., Luojus, K., Derksen, C., Mortimer, C., Wunderle, S., Nägeli, K. (2020). ESA CCI+ Snow ECV: Product Validation Plan, v2.0.

1.6. Acronyms

AD	Applicable Document
ATBD	Algorithm Theoretical Basis Document
ATSR	Along Track Scanning Radiometers
AVHRR	Advanced Very High Resolution Radiometer
CCI	Climate Change Initiative
CDR	Climate Data Record
CPU	Central Processing Unit
DARD	Data Access Requirements Document
DEM	Digital Elevation Model
ECV	Essential Climate Variable

ESA	European Space Agency
EO	Earth Observation
GAC	Global Area Coverage
GCOS	Global Climate Observing System
HPC	High Performance Computing
MODIS	Moderate Resolution Imaging Spectroradiometer
NSIDC	National Snow and Ice Data Centre
PS	Processing System
PSD	Product Specification Document
RD	Reference Document
SCF	Snow Cover Fraction
SLSTR	Sea and Land Surface Temperature Radiometer
SoW	Statement of Work
SRD	System Requirements Document
SSM/I	Special Sensor Microwave/Imager
SSMR	Special Sensor Microwave Radiometer
URD	User Requirements Document
UUID	Universally Unique Identifier

2. SNOW_CCI PROCESSING SYSTEM OVERVIEW

This section gives an overview of the processing system (PS) with its main modules, functions, and components. It also summarises its designated use and the system requirements.

2.1. Context

As depicted in the general overview of Figure 2.1, the *snow_cci* PS generates products and supports the process of algorithm improvement, reprocessing and validation. It provides products and services to the snow and hydrology community supporting their climate change impact assessment over a wide range of scales. The PS will be used by the *snow_cci* consortium but can also be applied by others as the overall workflow is very generic. A major difference to data production in other science projects is their often missing dissemination, i.e. the work ends with a publication and generated data products are not shared. The PS is specified to provide snow-related products such as Snow Cover Fraction (SCF) and Snow Water Equivalent (SWE) based on state-of-the-art technology using the best suited and available EO data and algorithms. The products are produced in a transparent and documented way, with accompanying meta-data, documentation, uncertainty, and validation reports (project outreach).

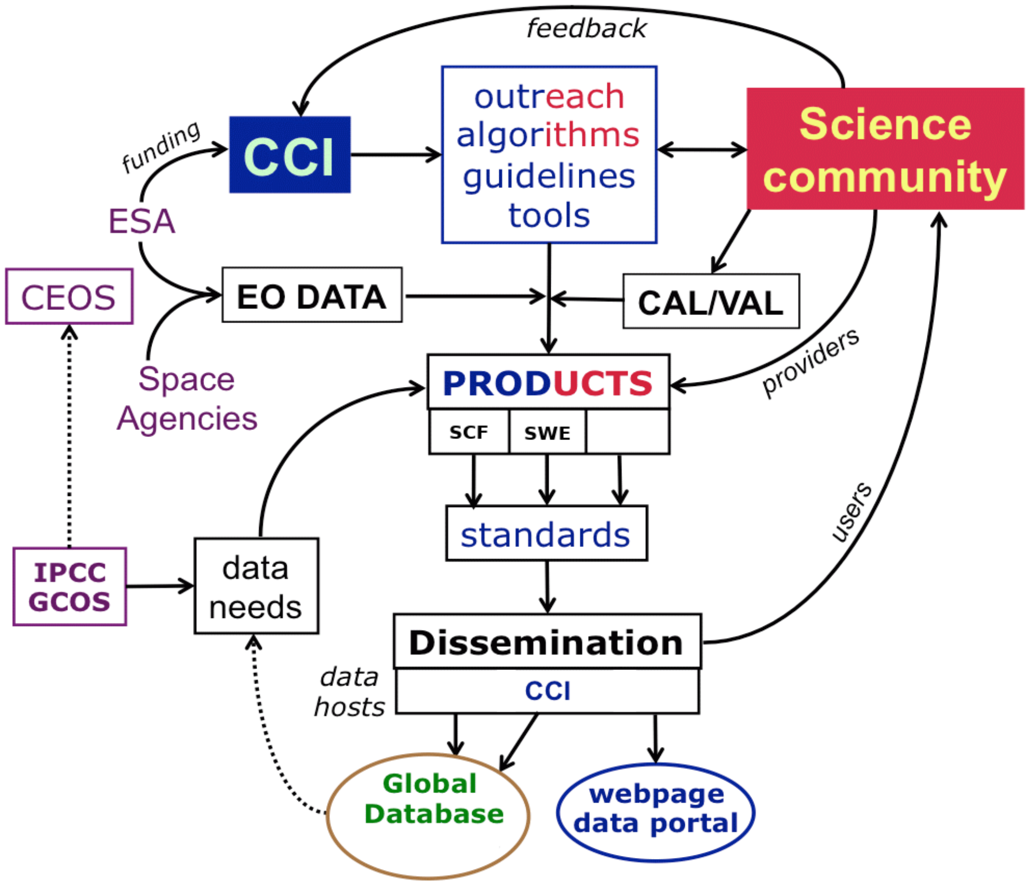

Figure 2.1: Schematic overview of the *snow_cci* processing system and its relation to developments by the science community for the individual products and their dissemination.

The PS can be understood as a value-adding layer between the data provider and the users. A high-level relation diagram of the PS is given in Figure 2.2. There are interfaces to the different user communities, which receive products and can provide feedback. Another interface is with the EO data providers. Depending on the module, EO data are obtained from the providers at CEOS level 1b or 2 and are ingested into the PS. Feedback is given to the providers about issues found with the data, processing improvements and requirements for the continuity of the service. Another interface is towards third-party sources to receive ancillary and validation datasets.

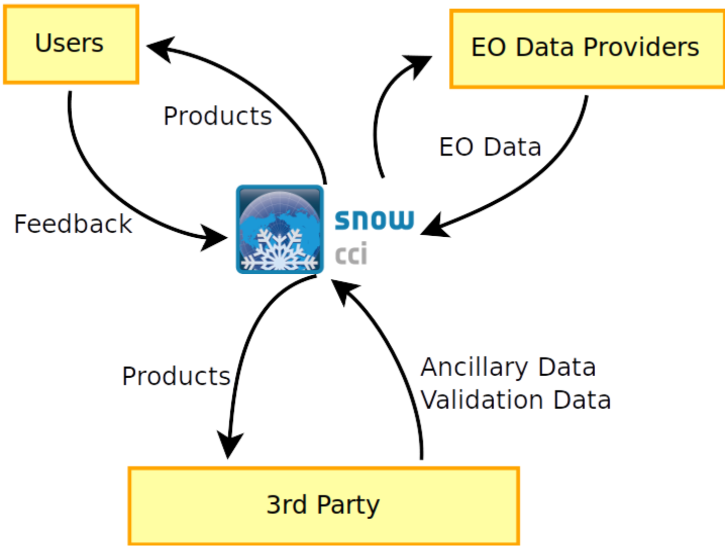

Figure 2.2: High level context of the *snow_cci* processing system.

2.2. User Requirements

The User Requirements are documented in the User Requirements Document [AD-1]. It outlines the requirements for *snow_cci* ECVs (snow cover fraction – SCF; snow water equivalent – SWE) obtained through engagement with users from across climate applications, including the detection of variability and trends, climate modelling, and aspects of hydrology and meteorology. The primary parameters to be evaluated are requirements for the GCOS parameters snow extent and snow water equivalent, from which snow depth can be inferred using different approaches to estimate snow density.

A synthesis of input from GCOS, IGOS, and OSCAR illustrates the full range of user needs, within which *snow_cci* requirements and product specification will occupy a specific niche. For both SCE and SWE, there are three general categories of users each with unique requirements: (1) climate, (2) terrestrial applications such as hydrology and agriculture, and (3) numerical weather prediction (NWP). In a general sense, the requirements for climate are relaxed with respect to characteristics such as spatial resolution and latency; NWP needs are most demanding, particularly for regional and/or nowcasting applications. Common requirements across applications are the need for daily data across extensive spatial domains.

The general baseline requirements for snow cover are daily data at spatial resolutions of at least 1 km with uncertainty of 5 to 10%. Because of the different variables related to snow cover (snow extent; snow covered area; snow cover fraction) it is not always clear from the requirements tables exactly what the uncertainty values correspond to, or how they can be derived. A distinction in spatial resolution requirements is made between alpine and non-alpine regions due increased heterogeneity in complex terrain. The update frequency, latency, and repeat interval vary between applications. Requirements for attributes specific to the derivation of snow cover parameters from optical remote sensing are not provided.

The general baseline requirements for snow water equivalent are similar to snow extent: daily data with spatial resolution of 1 km with uncertainty of approximately 10%. In some cases, the heritage of SWE retrieval from satellite passive microwave measurements is reflected in a more realistically achievable spatial resolution (5 to 25 km), and a distinction between shallow and deep snow. This is like the simple versus complex terrain requirements for SCE and reflects the tendency for passive microwave measurements to saturate under deep snow conditions.

2.3. Main System Requirements

System requirements are compiled in the SRD [RD-4]. The document lists system requirements grouped into functional, operational and performance requirements many with impact on the system design. Section 7 provides the complete matrix of forward tracing from requirements to sections and also indicates evolution of the requirements with time.

High-level requirements are to generate the *snow_cci* products (SW-FUN-0010, SW-FUN-0020, GL-FUN-0030). The processing line shall be well defined and flexible for future updates and adaptations (better algorithms, new input data) (SW-SIZ-0040). The available data shall be frequently reported and properly disseminated to the interested user communities (SW-INT-0050).

The functional scope of the system is not restricted to the processing, reprocessing, validation, and improvement cycle, although this is its main purpose. Also functions to make output products and documentation available to users included are in the scope of the system.

2.4. Main Functions

Requirements in this section are:

- SW-FUN-0010 Long-term storage
- SW-OPE-0050 Self-standing documentation
- SW-FUN-1050 Reprocess products
- SW-FUN-1060 Reprocess with new or so far unprocessed data

The PS hosts input data, performs pre-processing, classification processing, supports validation, and serves the output to users ('Dissemination' in Figure 2.2), the part that is often missing in other scientific studies (see section 2.1). It supports the interaction between the development team and users by information services. Processor interfaces, configurable data management, and version control with easy transfer to operations supports testing and development of new algorithms and continuous improvement.

To fulfil its purpose in such a context the PS provides three high level functions:

- Production
- Dissemination
- Life Cycle Management

In the following we will discuss the fundamental operations of the PS with regard to these three functions.

For production the focus is on processing and repeated reprocessing of complete products. Necessary functions are:

- Storage to gather and store inputs, intermediate products, output products and auxiliary data;
- Processors to produce output products from the input data;
- Processing Control;
- Quality Control of the intermediate and output products;
- Comparison with reference data;
- Documentation of the processing using meta-data;
- Ingestion of new input data and auxiliary data.

In general, we distinguish between the pre-processing, the main processing and the post-processing functionality covering the preparation of the input data, the processor itself, and the product generation steps, respectively.

For dissemination, the focus is on the service for the snow and hydrological and climate community. The products are distributed through existing and in development platforms of the ESA CCI data Portal and the Snow CCI Website. Functions are:

- Project Information/Introduction
- Processing Information
- Product Description (incl. meta-data)
- Online Product Access
- Validation Support
- Feedback Handling
- Long-term Preservation

Good Life Cycle Management helps improve service quality and reliability, crucial elements for the attractiveness of the provided service. A small effort should be necessary to implement an improved processor handling, improved algorithms and data of new sensors, given that basic characteristics of the data and the processing do not change (e.g. atmospheric windows for optical sensors, wavelength of LIDAR sensors). Consequently, fundamental necessary operations are:

- Test environment (for new processors)
- Access to test or benchmark input data (for tests and comparison)
- Version Management (this is linked to "Documentation of the Processing in Meta-data")

2.5. The modules of the PS

Like some of the other CCI projects, *snow_cci* produces products that rely on completely different processing chains in terms of input and ancillary data and the processor. The processing chains for each product are organized in modules that are part of the PS. In [RD-4] the following modules are part of the PS:

1. Snow Cover Fraction (SCF)
2. Snow Water Equivalent (SWE)

Further these modules will be developed for different input data, resulting separate modules such as e.g. the MODIS_SCF processor.

Within the framework of the *snow_cci* project we aim at developing the most efficient, consistent, and sustainable system addressing the needs of the corresponding community. We also aim at investigating synergies among the modules especially for the interfaces but potentially also with other CCI-projects if the system benefits. Hence, in the following we address both levels in the SSD, the module level and the system level. The distributed approach of the PS requires a community of producers and consequently a coordination mechanism to ensure continuity and management of multiple outputs.

2.6. High Level Decomposition

The functions and modules listed in the previous section are implemented as functional components. In this section we outline the high-level architecture of *snow_cci* on this subsystem level. Data-uptake is either through the mostly web-based services of satellite data providers (e.g. EOLI-SA, GLOVIS/Earthexplorer, Sentinel Science Hub) and those listed in [RD-3]. A key requirement is that all EO data are open and freely available. Data are then processed in the product generation modules and distributed through the CCI Common Basic Services (CCI data portal).

On a high-level we can distinguish the three subsystems: production, user services, and data stewardship based on the way EO data is encapsulated. This can also be viewed from a functional perspective as: production, dissemination, and improvement. Processing storage of the production system is accessed by the module processors and needs not to be openly accessible (i.e. with write access to the storage medium). The user services and archive are taken care by the CCI data portal and is not discussed here.

2.6.1. *Production Subsystem*

The production subsystem contains the production and development. Production control, processing storage and the processors provide the basic infrastructure for processing. A test environment with read access on input data serves the development needs. Where applicable, main functions of the processing system and processing storage shall be available for development. Due to the distributed processing system, development of new versions will always be performed in parallel to running an older version.

2.7. Hardware Infrastructure

The processing within the *snow_cci* project is done in a distributed way. Consequently, the different PS modules will be installed on different processing hardware infrastructure in Europe. The aim is to benefit from existing HPC clusters that can be accessed by the consortium.

2.7.1. *MODIS SCF Processor (ENVEO)*

The processing within the MODIS SCF Processor module is done within the supercomputing infrastructure in Norway, which is managed nationally by the company UNINETT Sigma2 AS (https://www.sigma2.no/). Processing time is granted on demand on a biannual basis, using the available supercomputing clusters within Norway, which is continuously updated and extended. The HPC runs in total about 1006 compute nodes featuring about 32256 CPU cores. The detailed specifications of the FRAM cluster are provided in Table 2.1.

Table 2.1: FRAM cluster properties (from https://www.sigma2.no/Fram)

Details	*FRAM*
System	Lenovo NeXtScale nx360
Number of Cores	32256
Number of nodes	1006
CPU type	Intel E5-2683v4 2.1 GHz Intel E7-4850v4 2.1 GHz (hugemem)
Max Floating point performance, double:	1.1 Petaflop/s
Total memory	78 TB
Total disc capacity	2.5 PB

The Norwegian Computing Centre (NR) has a project contract (NN5005K: SNOW-CCI - Global snow products from satellite data) with the UNINETT Sigma2, assuring access to the Linux cluster FRAM in Tromsø (https://www.sigma2.no/Fram) for the period 2019-08-01 to 2022-04-30. The project proposal was compiled by ENVEO and submitted together NR. For this period, ENVEO has an account on the FRAM for the production of SCF climate research data packages from MODIS data for the *snow_cci* project. For the *snow_cci* project, in the first year, a quota of 400'000 CPU-hours was available for the period 1 Apr – 30 Sep 2019. In the second year, a quota of 500'000 CPU-hours for the period 1 Apr - 30 Sep 2020, and for the following half year periods until April 2022 between 500 000 and 1 000 000 CPU-hours per period have been approved on FRAM. All input MODIS data are stored within the premises at NR in Oslo and are transferred over the UNINETT national backbone network to the FRAM in Tromsø. Processing software developed at ENVEO and all required libraries for compilation at of the system were installed on the FRAM cluster.

2.7.2. ATSR-2/AATSR SCF Processor (NR)

The ATSR-2/AATSR SCF Processor module is developed for two processing alternatives, either for in house processing in the Norwegian Computing Centre's (NR's) high-performance processing facility or by using the FRAM supercomputer that is also used with the MODIS SCF Processor module (described above). NR's HPC runs in total 12 compute nodes featuring 192 CPU cores. The detailed specifications of the NR HPC are provided in Table 2 .2.

Table 2.2: NR HPC properties

Details	*NR HPC*
Number of Cores	192 cores / 384 threads
Number of nodes	12
CPU type	Intel Xenon Silver 4110 CPU@2.1 GHz
Max Floating point performance, double:	Per CPU: Spec rate 2017 FP base: 88 (24 CPUs)
Total memory	3 TB
Total disc capacity	200 TB (+ 12 TB and 24 TB NVME)

The Norwegian Computing Centre (NR) has a similar contract as for the MODIS processing with the UNINETT Sigma2, assuring access to the Linux cluster FRAM in Tromsø for the period 2021-05-01 to 2021-12-31. All input ATSR-2/AATSR data are stored within the premises at NR in Oslo and are transferred over the UNINETT national backbone network to the FRAM in Tromsø.

2.7.3. AVHRR GAC SCF Processor (UniBE)

The AVHRR data are stored on RSGBstore a dedicated data storage of the Remote Sensing Research Group at the University of Bern. For processing the data needs to be transferred from RSGBstore to the processing hardware. The AVHRR GAC processor is running on an internal server ('rsgb-r2d2') of the Remote Sensing Research Group of the University of Bern. The server comprises 10 CPU kernels with 64 GB RAM, facilitating a smooth and fast processing. The 'rsgb-r2d2' is directly linked to the RSGBstore enabling a direct transfer of input data and output results.

2.7.4. SLSTR SCF Processor (ENVEO)

The SLSTR processor is running on ENVEO’s computing system in Innsbruck. ENVEO’s computing system has more than 700 TB disk space and is connected via a 10Gbit network with a workstation with about 20 nodes. The Sentinel-3A/B SLSTR Level 1B data for one year were downloaded from the Copernicus Service Hub and stored at ENVEO’s premises.

2.7.5. SSMI/I SWE Processor

FMI utilizes parallel computing cluster that resides in FMI premises in Sodankylä. In connection to the computing cluster there is an object storage system with a capacity of 1 PB. Both systems are easily scalable. In average, one year of SSMI SWE products per CPU can be processed in a day.

2.7.6. PROBA-V SCF Processor (ENVEO)

The PROBA-V processor is running on ENVEO’s computing system in Innsbruck. ENVEO’s computing system has more than 700 TB disk space and is connected via a 10Gbit network with a workstation with about 20 nodes. The PROBA-V Level 3 Top Of Atmosphere Reflectance daily synthesis data at 1000m resolution for one year (2019) were downloaded from VITO and stored at ENVEO’s premises.

3. *SNOW_CCI* PROCESSING SYSTEM WORKFLOW AND OPERATIONAL SCENARIOS

3.1. Roles

The development team consists of scientists and operators that manage the production and continuous development. Actors in the operational scenarios are users with different roles depending on how they use the system:

- *snow_cci* Users (client)
    - are interested in best existing snow products
    - are skilled in snow applications
    - provide feedback and proposals
    - request data format compatible with their communities
- Development team
    - mandated to run snow_cci
    - in dialogue with users
    - develops the PS further
    - issues product versions
- Validation experts
    - are part of the international community
    - support development team
    - provide local expertise
    - feedback on the products
- Auditor/Project Manager
    - project supervision

3.2. User Information and Data Access

Users access the *snow_cci* data products using the ESA CCI Open Data Portal. Updates to the system are disseminated via the *snow_cci* website.

Typical functionality performed in the context of user information and data access:

- The development team submits new products to the ESA CCI Open Data Portal. New versions or major updates will be announced on the snow_cci website.
- A user accesses the snow_cci web site as an entry point and gets general information on snow_cci. It also provides links to software, services, and resources.
- The user uses the CCI Search catalogue to find snow_cci products. The catalogue also provides metadata and quick looks.
- The user downloads product documentation (e.g. the Product User Guide).

3.3. Processing System Workflow

The basic processing system production workflow is given in Figure 3.1. It has the main parts pre-processing, retrieval, product generation and verification/validation. The input data are EO data, station data (point data), and auxiliary data such as DEM and masks. The output of the production are the SCF and SWE products. In the following the two *snow_cci* modules will be presented in further detail.

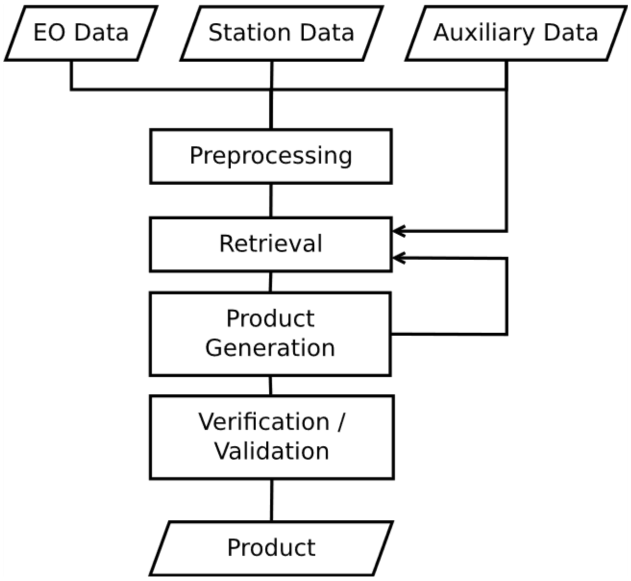

Figure 3.1: *snow_cci* module high level diagram.

3.3.1. *Pre-processing*

The pre-processing includes all steps necessary that the retrieval can be applied. This covers input and auxiliary data selection and input data preparation (conversion, reprojection, sampling, conversion).

3.3.2. *Retrieval*

In the retrieval process the algorithms from the ATBD [RD-6], updated according to the ADP [RD-7] are applied.

3.3.3. *Product Generation*

In the product generation step the resulting maps are put in the delivery product format as requested and specified in the Product Specification Document [RD-2].

3.3.4. *Verification/Validation*

The products undergo defined procedures to ensure product integrity and quality. This step must not be confused with the product validation process that is done in WP42x following the Product Validation Plan [RD-7].

The system operators are in charge of system (module) verification. The verification is done separately for the different PS modules. It covers:

- System (module) validation (after an upgrade or new installation)

The PS provides tools that facilitate these tasks such as

- Benchmark test data (SW-OPE-4050)
- Test tools (SW-OPE-4050)
- Verification Report (SW-RAM-5040)

3.4. Algorithm Improvement

The development team decides about features or processes to be improved in order to meet user requirements. The development team implements the improvements as new versions of processors and if necessary, as a modified workflow definition. The development team tests and validates the new products. The development team also decides about new versions to be released.

An improvement cycle is defined as:

1. New requirements are identified and analysed
2. Modified processor implementation
3. Test production
4. Validation
5. Decision to a) go to 1. and iterate again, b) implement go on, c) stop here
6. Release a new version (code freeze) while retaining older ones for cross check
7. Start production

After each validation, a decision is taken if the improved algorithm is accepted, further developed or the development is stopped. Only in the case of acceptance, the development leads to a new version of the processor and full reprocessing of the archive or implementation for new data. This cycle of innovation and improvement typically takes 5-10 years. With an effective control mechanism in place, the update cycle can be reduced to a few years.

4. FUNCTIONAL DESIGN

Here we will discuss and present the major functional blocks of the *snow_cci* PS.

4.1. Services

Most of the necessary services are provided through the CCI environment such as the data exchange storage, the document management system and user interaction. For the PS development software repositories and an issue tracker are hosted at the different partners involved in the PS module development.

4.1.1. Processor software repository

An important element of the modern software development process is source control (or version control). Cooperating developers commit their changes incrementally to a common source repository, which allows them to collaborate on code without resorting to crude file-sharing techniques (shared drives, email). Source control tools track all prior versions of all files, allowing developers to "time travel" backward and forward in their software to determine when and where bugs are introduced. These tools also identify conflicting simultaneous modifications made by two (poorly communicating) team members, forcing them to work out the correct solution (rather than blindly overwriting one or the other original submission). The software repository contains the actual processing code and all prior versions. The write access to the processor repository is restricted to the development team. As all software changes are updated directly in the repository, the software changes are published almost immediately and are made available for review.

4.1.2. Issue Tracker

During software development a Redmine issue tracker is used (http://www.redmine.org/). Redmine is a flexible project management web application written using Ruby on Rails framework. Redmine integrates the version control system into its user interface and manages the access control to the version control system resulting in a state-of-the-art FOSS software development environment.

4.2. Processors

The processors cover the necessary tools to produce the different *snow_cci* products. The *snow_cci* is organized in modules covering the production of the different products (cf. Table 4.1). In general, a distributed processing approach is followed. Consequently, the modules are portable.

Requirements:

- SW-FUN-0040 Flexible production
- SW-FUN-5010 Data overwrite

Table 4.1: Overview on processors used for the product generation in snow_cci.

Component	Purpose	Content	Implementation
MODIS processor	Generate L3 snow_cci SCF Product	Input Data	New processor, adapted from CP2
AVHRR_GAC processor	Generate L3 snow_cci SCF Product	Input Data	New processor, adapted from CP2
ATSR-2/AATSR processor	Generate L3 snow_cci SCF Product	Input Data	New processor
SLSTR processor	Generate L3 snow_cci SCF Product	Input Data	New processor
SMMR/SSMI/S processor	Generate L3 snow_cci SWE Product	Input Data	New processor, adapted from CP2, Heritage from GlobSnow
PROB-V processor	Test generation of L3 snow_cci SCF prototype product	Input Data	New processor

4.3. Concept for continuous Improvement

Continuous improvement is an important aspect in the CCI projects. To ensure a transparent process Software Modularity, Software Version Concept, Version Control, and Version Numbering are important issues. This section defines the structures and functions that extend the production environment for continuous improvement. Focus is on flexibility, rapid testing, and prototyping. The concepts described are processors, versioning, and a test environment. The concept of processors and versioning contribute to the modularity of the system.

Requirements:

- SW-OPE-4010 Development under version control
- SW-OPE-4020 Decoupled from own research

The software of the PS and the processing algorithms code are under version control. The software repository contains the actual processing code and all prior versions. All software changes are documented in the repository. Version numbering of the processor is reflected in the repository by revisions and tags. Revisions are usually linked to commits and indicate the sequential order of documented changes. Tags are set to indicate software releases of frozen software states. Subversion is a good candidate for version control. Together with Redmine it is a complete FOSS version control and issue tracking system.

Data processors help to organize the data processing in modules. Due to the differing input and output datasets/formats, the modules are normally not shared among products (even if the functionality is the same). A processor is a software component that can be parametrised and that generates a (higher level) output product of a certain type from one or several input products of certain types. A PS module consists of the sequential call of processors. Each processor has its own version information. Parameters are usually provided as command line arguments, environment variables or as information in a parameter file. Feedback is received by a return code, messages on stdout/stderr and in log files.

Table 4.2: Overview on the software repository to be used for version control.

Component	Purpose	Content	Implementation
Software Repository	Stores all versions of the processor code in a transparent way, with branches, authorship	Code	Subversion Redmine Tools

4.4. System Documentation

The documentation contains the PS documentation consisting of manuals, specifications, and reports, as well as the product documentation consisting of product specifications, manual and validation reports. At this stage no advanced functionality such as collaborative editing etc. seems to be necessary so that the basic functionality of any FOSS CMS might be sufficient for this task.

Requirements addressed by this section are:

- SW-OPE-4030 Product Description
- SW-RAM-5020 Self-standing documentation

The PS documentation includes requirement documents, design and interface control documents, test documents, manuals, and maintenance information. *snow_cci* deliverables to name here are ATBD and the PUG. The SRD and SSD define requirements and design of the system.

Table 4.3: Overview on the system documentation.

Component	Purpose	Content	Implementation
Documentation Management	Stores documentation in a structured and transparent way	Documentation	Project Website at ENVEO

5. DEVELOPMENT, LIFE CYCLE, COST AND PERFORMANCE

This section discusses the system development in the future, potential development strategies, and efforts. The development is driven by several factors such as the availability of new technology, faster algorithms, new scientific findings, and improved product algorithms, new available EO data, and user needs.

5.1. Re-use and Development

Development is needed to bring the existing prototypes of the PS modules to a higher operational level satisfying the requirements listed in the previous sections and to add the missing components such as those for user services, data handling, life cycle management, archiving etc.

Requirements addressed by this section are:

- PF-OPE-4030 Development based on User Requirements
- PF-OPE-4020 Development decoupled from research
- PF-OPE-4060 Freeze prototype
- PF-FUN-5030 Verification of implementation

Table 5.1 summarises the tools that were used, adapted, configured, and integrated during development of the PS within the CCI and beyond.

Table 5.1: Tools for the development of the processing systems in *snow_cci*.

Name	*Usage*	*Remarks*
Subversion	Version control	FOSS
Python3	Scripting, netcdf4 reader, pyresample, rasterio	FOSS
SWE Processor	Produces the SWE product	Depends on Matlab
Matlab Runtime Environment	Needed by the SWE processor	Proprietary
AVHRR GAC SCF Processor	Produces the SCF product from AVHRR	UniBE
MODIS SCF Processor	Produces the SCF product from MODIS	UNINETT Sigma2 FRAM
ATSR-2/AATSR SCF Prototype	Running prototype in retrieval	NR HPC & Sigma2 FRAM
SLSTR SCF Prototype	Running prototype in retrieval	ENVEO
PROBA-V SCF Prototype	Running prototype in retrieval	ENVEO

Name	*Usage*	*Remarks*
ESP v2.1	Modules for reading, geolocating, analysing and writing EO data needed by the SCF Prototype processor for MODIS data	ENVEO's in-house developed Software Package
LIBRARIES	GDAL, HDF-EOS	UNINETT Sigma2 FRAM
AUX TOOLS	GDAL	

5.2. System Life Cycle Drivers and Considerations

The PS needs to be incrementally adapted to integrate new functional extensions, improved algorithms, and input datasets. New EO data make adaptations necessary and most likely also have an impact on the hardware infrastructure. The life cycle plan cannot be static as it is not foreseeable. Currently the following driving factors are identified:

- Availability of the existing processor module prototypes
- Functional extension of the system
- New workflows
- Improved algorithms
- New Sensors
- Hardware improvements
- Dependencies on 3rd parties (other ECVs, data providers, new users)

To answer the first two points the system is initially based on the prototype. Incrementally, additional components and functions are added and interfaces to data providers and users are extended. The third and fourth point of workflow and algorithm development requires the addition of tools for validation and user feedback.

The new sensors and the increased data volume are a qualitative change, too. The existing methods need to be adapted to make use of new sensors. For the longer perspective renewal of hardware and optional change of software layers must be taken into account. The PS design is prepared for this by the modularity of its functional components. The last item is not so relevant for *snow_cci* at the moment as the dependence on other CCI projects is minor.

Requirements:

- SW-FUN-1010 Long-term storage
- SW-FUN-1020 Unique identifier
- SW-FUN-1030 Structured storage
- SW-FUN-1050 Reprocess Products
- SW-OPE-4010 PS shall be under version control
- SW-OPE-4050 Test tools
- SW-RAM-5040 Verification

5.3. Sizing and Performance Analysis

In the SRD, no specific requirements are present concerning the processing time performance. At this stage it is mainly labour hours that drive the processing rather than CPU core hours. Full reprocessing of historical data requires a variable amount of work, depending on the product (see tables below). The data storage budget for inputs and outputs for historical data and for the yearly increase of acquired data is in the low TB range initially but might grow to more substantial TB values with new sensors becoming available.

There exist requirements on disk space that are modest:

- SW-SIZ-2060 Space for input data
- SW-SIZ-2050 Space for auxiliary data
- SW-SIZ-2040 Space for output data
- SW-SIZ-2030 Run on available hardware
- SW-SIZ-2010 Time series

Below the budgets for data storage and processing capabilities are estimated. The budget for data storage mainly depends on the amount of input data to be managed. This comprises historical data and data acquired continuously.

5.3.1. SCF

The satellite data used for the SCF product generation are from different sensors. The AVHRR instruments are available for the period from 1982 to 2020, while MODIS data are available from 2000 to 2020. ATSR-2 was available from 1996 – 2008 and AATSR from 2002 - 2012. SLSTR data are in general available since 2016, data pre-processed by ESA with the most recent processing baseline version and used for the demonstration production in the second year are only available since January 2020. PROBA-V data of the year 2019 were selected to test the SCF retrieval from this sensor. The sensors have different pixel spacings and numbers of spectral bands resulting in different amounts of disk space needed for storing and processing the data. All details about the EO data used for the SCF production from AVHRR, MODIS, and A/ASTR-2 and SCF prototype generation for each one year from SLSTR and PROBA-V in the third year are summarized in Table 5.2.

Auxiliary data used for the SCF processing are prepared with global coverage with the pixel spacing as used for the SCF product generation.

Three auxiliary data sets are used for the SCF prototype processing from MODIS, SLSTR and PROBA-V data with 0.01 degrees pixel spacing. These data are stored as 8-bit layers and require each about 620 MB disk space.

For the SCF prototype product from AVHRR data UBELIX the University of Bern Linux cluster is used. The production runs in the queuing system on several nodes in parallel. The data is stored on RSGBstor, a storage service of the Remote Sensing Research Group University of Bern and needs to be transferred to UBELIX for the processing. The standard procedure is therefore to process the archive in bunches of annual data. The processing of a single day takes minutes. The current implementation runs 10 processors with 8GB memory each in parallel. The number of processors is not critical and could be increased if necessary. Considering the short processing time in comparison with the data transfer time there is currently no need to allocate more resources.

Table 5.2: Data used for SCF prototype product generation in the first and second *snow_cci* year.

Data	Product	Time Span	Historical Data	New Data
AVHRR	GAC Level 1c	1982 – 2018	7.5 TB	~0.2 TB / yr.
MODIS	MOD021KM (Calibrated Radiances 5-Min L1B Swath 1km), MOD03 (Geolocation Fields 5-Min L1A Swath 1km), only land areas	2000 - 2020	~200 TB ~50 TB	~10 TB / yr. ~2 TB / yr.
ATSR-2	Level 1B 1km resolution, 500km swath	1995 – 2002	60 TB	
AATSR	Level 1B 1 km resolution, 500 km swath	2002 – 2012	80 TB	
SLSTR	Level 1B 1km resolution, 1270 km swath	2020 - present	~20 TB	~20 TB / yr.
PROBA-V	S1_TOA (Level 3 TOA reflectance, 1km, daily synthesis)	2019	~426 GB	426 GB / yr.
Total			**850 TB**	**~460 Tb / yr.**

ATSR-2/AATSR processing on FRAM supercomputer will follow the same approach as described in detail below for MODIS. The dataset is smaller than for MODIS, and experience from the first project year indicates that processing of the whole dataset might take less than one week. Data will be stored at NR in a new storage system (like the system for MODIS data), which has already taken place. ESA is currently reprocessing the AATSR archive, and the new dataset (v4) will be stored in SLSTR format. For ATSR-2 we will use data in the old format (v3) as reprocessing will probably not be completed before 2022.

For the SCF product generation from MODIS data the infrastructure on the HPC FRAM is used, which runs applications as jobs in a queuing system scheduling the tasks and process to run on compute nodes. The terminology of jobs and nodes follows the definitions of the UNINETT Sigma2 (https://documentation.sigma2.no/jobs/framqueuesystem.html).

One job on the FRAM must use between four and 32 nodes. One node has 32 CPU-hours and 64 GB RAM, so with each job, 128 CPU-hours of the totally approved 500’000 CPU-hours per six months are consumed. For optimized usage of the nodes the full CPU-hours per job should be used by each processing step. All normal jobs get exclusive access to whole nodes, so all CPUs and memory reserved for a node are charged by the system.

The download of MODIS tiles from NR to FRAM is started manually and is thus executed immediately. This processing step uses between four and six parallel bash scripts. One download script can download the raw MODIS tiles for one day in about 15 minutes. Considering the disk space available for processing on the FRAM, raw MODIS tile data of only 400 to 500 days will be stored on the FRAM at the same time.

Two jobs are used for the SCF processing on the MODIS tile level. Given that one job has four nodes, and one node is processing one day of MODIS tiles, one job processes four days of MODIS tiles. Per job, maximum 300 days can be considered per day for the processing. So, two jobs should run in parallel and continuously for the tile processing to handle the downloaded data. A limiting factor for this processing step is that the starting time of these jobs depend on the queue on the FRAM and cannot be controlled manually. The waiting time per job from submission to execution is currently about one day.

One job is used for generating the daily composite of the MODIS based SCF product. For this processing step, one node can handle the daily SCF composite of six days in parallel for optimized usage of the CPU-hours.

The estimated processing times on the FRAM assuming a single node and the required disk space for the temporary files for the generation of SCF products from MODIS data are summarized in Table 5.3.

Based on experience of the first production round, the generation of SCF products for the full time series of MODIS data is estimated to be completed in less than a month.

Table 5.3: Estimated processing time and disk space for MODIS required for temporary files using one node.

Processing step	*Duration of processing step*	*Required disk space for temporary files*
Download MODIS tiles from NR to FRAM	15 min / day	34 GB / day
SCF processing at tile level	25 min / tile	60 GB / day
Composite of SCF tiles to global daily SCF product including uncertainty layer	18 min / day	1.4 GB / day

The SCF processing from SLSTR and PROBA-V data in the third year is planned for demonstration purposes. SLSTR data pre-processed with the most recent baseline processing version at ESA are only available since January 2020. PROBA-V data were downloaded for the year 2019, and will be tested for SCF retrieval. Considering the direct link of the raw data to ENVEO's processing system and taking into account the available RAM, the demonstration production from SLSTR and PROBA-V is estimated to be each completed in less than a month.

5.3.2. SWE

The SWE processing is based on radiometer data. In the case of NRT processing the satellites (DSMP-series) can cover the Northern Hemisphere within 24h. Once all the overpasses are available the pre-processing of the data takes about 15 minutes. However, when Climate Data Records (CDR) are considered the data can be acquired already pre-processed. The processor also uses synoptic Snow Depths (SD) as an input but due to the small size of the data the acquisition of SD data is very fast compared to radiometer data. The processing of the SWE product takes in average about 8h during the snow season. It can vary considerably, and worst case is about 10h. On the other hand, during autumn the processing can go through in 2h or less. In case of CDR processing multiple days (~half month) can be processed in parallel.

Table 5.4: Data used for SWE prototype product generation in the first and second *snow_cci* year.

Data	Product	Time Span	Historical Data	New Data
SSMR	NSIDC EASE Gridded Tb	1978 – 1987	30 GB	NA
SSMI / SSMI/S	NSIDC EASE Gridded Tb	1987 – present	96 GB	-
SSMR	MEaSUREs Calibrated Enhanced-Resolution Passive Microwave Daily EASE-Grid 2.0 Brightness Temperature ESDR, Version 1.	1978 – 1987	238 GB	NA
SSMI / SSMI/S	MEaSUREs Calibrated Enhanced-Resolution Passive Microwave Daily EASE-Grid 2.0 Brightness Temperature ESDR, Version 1.	1987 – present	2TB	-

6. REQUIREMENTS TRACEABILITY

ID	Title	Reference
SW-FUN-0010	Develop and validate algorithms to approach the GCOS ECV and meet the wider requirements of the Climate Community (i.e. long term, consistent, stable, uncertainty-characterized) global satellite data products from multi- sensor data archives. (CR-1)	§2.3
SW-FUN-0020	Produce, validate and deliver consistent time series of multi-sensor global satellite ECV data products for climate science. (CR-2)	§2.3
SW-FUN-0030	The *snow_cci* system shall generate two different products: 1. SCF 2. SWE	§2.3
SW-FUN-0040	Generate and fully document a production system capable of processing and reprocessing the data, with the aim of supporting transfer to operational activities outside CCI (such as C3S). [CR-4]	§2.3
SW-OPE-0050	All project documentation shall be made publicly available via the CCI Open Data Portal: http://cci.esa.int.	§2.3
SW-OPE-0050	The PS shall capitalise on existing European assets through their reuse, particularly Open Source scientific tools and prototype ECV processing systems from prior projects. (heritage)	§2.3
SW-INT-0030	The global Snow community shall play an active role in its creation according to given guidelines and advice from a strategic operations team. They shall also give feedback from the implementation to the strategic team.	§2.3
SW-SIZ-0040	The system shall implement a data production line that is sufficiently flexible to continuously update and extend the database (e.g. with data from new sensors or better acquisitions).	§2.3
SW-INT-0050	The available data shall be frequently reported and properly disseminated to the interested user communities.	§2.3
SW-FUN-1020	The Products shall be uniquely identified.	§2.3
SW-FUN-1030	The PS shall store data in a structured way using type, revision, date.	§5.2
SW-FUN-1040	If input data is retrieved directly from a third-party ground segment, the PS has to ensure that links are maintained and functionality is regularly checked.	§4.3
SW_FUN-1050	The PS shall be able to reprocess also parts of the products.	§2.4 §5.2
SW-FUN-1060	The PS shall be able to do partial processing.	§2.4 §5.2
SW-SIZ-2010	The PS shall be able to do long time series processing in due time.	§5.3
SW-SIZ-2020	The PS shall be able to do reprocessing in due time.	§5.3
SW-SIZ-2030	The PS shall be able to run on the available hardware infrastructure	§2.7
SW-INT-3010	The PS shall have the capability and interfaces to extend for future adaptations.	§5.3

ID	Title	Reference
SW-OPE-4010	Development of the PS shall be under version control.	§5.1
SW-OPE-4020	The system should be decoupled from the research	§5.1
SW-OPE-4030	Development of the system shall be based on the user requirements, the selected algorithms and the validation protocols used to generate the baseline products.	§2.2
SW-OPE-4040	The PS development shall be overseen by a science team that drives the development process.	§3.1
SW-OPE-4050	Each PS installation includes a set of test tools, data and benchmark data to test PS integrity (end-to-end, interfaces)	§5.1
SW-OPE-4060	If a module is based on a prototype, the prototype state must be frozen until it is implemented.	§5.1
SW-OPE-4070	The verification is regarded as successful, when all tests agree within TBD limits. Hashes are to be preferred where applicable.	§3.3
SW-FUN-5010	The operational processor shall not overwrite existing data. Versioning shall be used instead.	§5.2
SW-RAM-5020	The system developed shall be detailed as a separate self-standing document providing an overview of the system and its components, functionality of the system and its subsystems, inputs, outputs, resource key interfaces, and resource requirements.	§5.1
SW-FUN-5030	Verification of the correct implementation of the prototype system against the algorithms developed in Task 2 is a fundamental part of the process.	§2.3
SW-RAM-5040	The verification shall be documented in a Verification Report. It shall contain the chosen approach and the justification, the selected verification data set and the verification results.	§2.3
SW-RAM-5050	The PS shall provide means against data loss of its input / output products.	§2.3
SW-FUN-5060	All data stored in the system shall be available for the long-term (at least 15 years).	§2.3