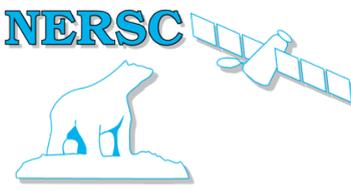




sea state
cci

System Specification Document (SSD)

version 3.0, 10 February 2022

Contents

List of Acronyms	3
1. Introduction	4
2. Requirements analysis	5
2.1 Trade-off analysis	5
2.2 Engineering methodology	5
2.2.1 Processing target	5
2.2.2 External support	6
2.3 Cost effectiveness	6
2.4. Other aspects	7
2.4.1 Security measures	7
2.4.2 Conformance to EU GDPR	7
2.5 System Architecture	7
3. System specifications	8
3.1 Processing workflow	8
3.2 Processing toolboxes	11
3.3 source code control	13
3.4 Processing platforms	13
3.4.1 Ifremer “Datarmor” platform	15
3.4.2 CNES HAL	16
3.5 Product distribution	17

Author	Approved	Signature	Date
JF Piolle	Fabrice Arduin Ellis Ash		10 February 2022
ESA Acceptance			

Issue	Date	Comments
1.0	14/11/2019	First version submitted for ESA approval
2.0	22/09/2021	Updates for dataset v2
3.0	10/02/2022	Updates for dataset v3

List of Acronyms

CCI	Climate Change Initiative
CMEMS	Copernicus Marine Environment Monitoring Service
CPU	Central Processing Unit
ECV	Essential Climate Variable
ESA	European Space Agency
GDR	Geophysical Data Record
HPC	High Performance Computing
IPF	Input Processor Function?
L1A	Level 1A
L1B	Level 1B
L4	Level 4
LRM	Low Rate Measurement
RA	Radar Altimeters
RR	Round Robin
S3A	Sentinel-3A
S3B	Sentinel-3B
SAR	Synthetic Aperture Radar
SSH	Secure Shell / Sea Surface Height
SWH	Significant Wave Height
WV	Wave (mode for SAR)

1. Introduction

This document presents the System Specification Document (SSD) for **Sea_State_cci**, deliverable 3.2 of the project. This third version is prepared at the end of the project in order to give updates on the processing system used for the version 3 dataset.

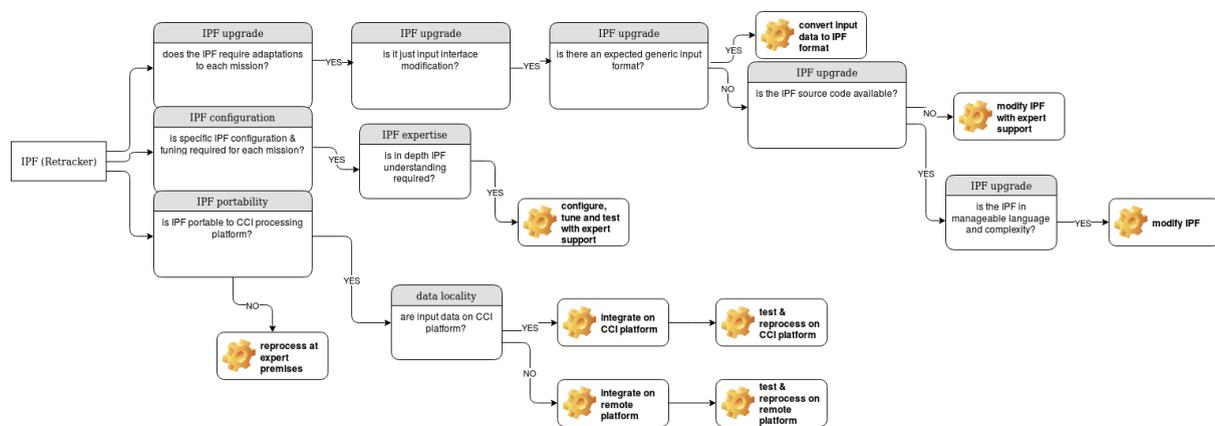
The remainder of this System Specification Document contains sections describing:

- Trade-off analysis
- Engineering methodology
- Cost effectiveness
- Other aspects
- System architecture

2. Requirements analysis

2.1 Trade-off analysis

At the time of system design, little is known on the final choice of IPF (altimeter retracker or SAR Wave Mode processor) that will come out of each round-robin activities. A wide range of different cases is expected, depending on the proposed IPF, and the processing system will therefore be strongly dependent on a trade-off analysis of the selected IPF. The following diagram sketches the decision tree anticipated for this trade-off analysis. We detail in the next section the possible strategies that may be put in place with respect to this analysis for the reprocessing of each mission considered in the CCI Sea State project.



2.2 Engineering methodology

The CCI Sea State approach on the reprocessing of data climate series is based, in the first phase, on a round-robin exercise that will support the selection of the best algorithm (implemented through a processor or suite of processors) to produce a consistent and stable multi-mission time series of the chosen sea state parameters.

Based on the above analysis and the selected processor(s), different steps and implementation paths are anticipated, with respect to two main aspects:

2.2.1 Processing target

It was originally expected to have all reprocessing activities running on the same CCI platform (at Ifremer) but the preliminary trade-off analysis leads to possibly different scenarios, in particular using other remote processing platforms:

- some input datasets may be too large to be transferred to the CCI processing platform: in particular the L1A/L1B data from Sentinel-3 proved to be very complex, time consuming and anyway too voluminous to restore from their original archive (at ESA or Eumetsat) and transfer electronically to Ifremer. This is a case where other processing platforms are to be sought and used, close to the original data.
- the processing time of a single orbit is highly variable depending on the selected processor for the reprocessing: from the requirement analysis, ratio of 1 to 30 are

expected for altimeter LRM retrackers, possibly x50 for SAR mode. Overall CPU time estimation ranges from hundreds of thousands to millions of hours... Completing the reprocessing of all expected missions within the time frame of the CCI project may therefore require to split the processing on different platforms too

- there is no clear commitment for Round-Robin participants to deliver an executable processor (only the algorithm is to be provided). Re-implementation of a processor would be too time consuming and would require an important validation/intercomparison effort. There may therefore be legal or practical blocking issues to transfer and integrate a processor on CCI (or another) platform: this is another possible case where the processing may have to be moved to a platform where the processor is already available and integrated within an IPF. The reprocessing responsibility may then have to be transferred to another entity/partner within the project.

2.2.2 External support

The outcome of the Round-Robin step will be an algorithm, implemented within an executable processor. At this stage it is not foreseen to re-implement a dedicated processor from an algorithm, as it is estimated as too costly and time consuming. It is therefore expected that in the engineering methodology for the CCI reprocessing activity that support from the processor expert will have to be sought to address possibly different aspects:

- adaptation of processor to other format and content: each satellite mission comes with a different format and content for the instrumental level observation data (L1). This is especially true on the considered CCI time frame that spans over 30 years. It is therefore very likely that either of the following approaches will have to be performed:
 - reformatting a mission's L1 data to the format and content expected by the processor, which can be done without expert support as long as this information is accurately specified and documented
 - adapting the processor to a mission format and content: depending on the language used in the processor implementation, the availability of source code,... this may not be doable by anybody else than the author (person or group) having implemented the processor: this is a case where support from this author will be mandatory.
- configuration, adaptation of specific processing steps and fine tuning of the processing to each mission: all instruments have specifics that will have to be addressed when applying the selected processor (sensing band, mispointing, sensor drift,...). This will require support from the experts behind the specification and design of the selected processor.

2.3 Cost effectiveness

As demonstrated in the above analysis, there are at the project start many uncertainties regarding the reprocessing activities, having no prior knowledge of the choice of processor(s) to generate the CCI climate data records.

In terms of processing and infrastructure cost, effectiveness will be achieved through the following drivers:

- relying on existing institutional or institution funded processing resources: internal supercomputers at involved agencies (Ifremer, possibly CNES) or new generation cloud computing platforms funded by Copernicus, ESA, etc... will be sought after in order to support the financial cost without any additional charge for the CCI project
- exploit the locality of the input data: process the data as close as possible to where they are stored whenever possible, instead of doing costly and time consuming data transfer. In particular, a remote platform such as the ESA Sarvator may be used for the reprocessing of at least some ESA missions (such as Sentinel-3). CNES supercomputer may also be an option for some missions, in addition to the Ifremer platform
- exploit the processing time availability: we demonstrated that the overall CPU time required to process all mission may be overwhelming for a single processing facility, depending on the choice of the processor; as for data locality, we may use different platforms to distribute the processing in order to take advantage of available institutional (free of charge) processing resources and being also more effective in the usage of the existing facilities resources
- leverage on existing parallel initiatives, such as other reprocessing activities with the same or similar retracker performances, to avoid unnecessary waste of processing resources and manpower

2.4. Other aspects

2.4.1 Security measures

The CCI Sea State project does not hold any sensitive data and no specific action is undertaken to address such aspects.

Each considered processing platform for the project is however operated in fully secured environments with access restrictions and controls for physical persons, identification and authentication for remote network connection.

2.4.2 Conformance to EU GDPR

The CCI Sea State project does not hold any personal data. As such, no action is needed to comply with EU GDPR.

2.5 System Architecture

Keeping long and massive mission archives alive by raising the level of data revisiting through multiple applications, demonstration products or services, or extensive data reprocessing are major requirements for projects such as the CCI Sea State in order to combine numerous sources of data, test new ideas and perform reprocessing over a significant amount of data.

The system architecture for the project therefore relies on existing infrastructures providing these capacities.

Among the key technical drivers identified to provide an efficient and cost-effective access to historical massive multi-mission archives are :

- fast and online access to massive collections of data
- avoiding data duplication and transfer to users
- minimizing time from algorithm development to processing
- allowing fast and easy to manage large scale reprocessing
- improving data storage and management
- transitioning towards pre-operational phase

There may be several target infrastructures fitting these criteria and the CCI Sea State project considered several of them, wrt their respective strengths and limitations for a particular mission retracking. Possible infrastructures that were initially envisaged:

- Partners local storage and processing infrastructures, such as Ifremer “Datarmor” platform, originally intended for most of the processing
- ESA “SARvatore”, which may prove more adapted for ESA missions (thanks to the availability of the input data) such as Sentinel-3A & B, CryoSat-2 or Envisat
- ESA CNES, which hosts all Jason mission data and has already integrated retrackers

Eventually local infrastructures are preferred and used throughout phase 1, in order to minimize the technical effort while focusing on the scientific improvement of the data processors. Different platforms are therefore used for the CCI Sea State processing, as detailed in section 3.4.

In phase 2, this architecture will be revised to allow transition to a pre-operational service. Planned evolutions will include:

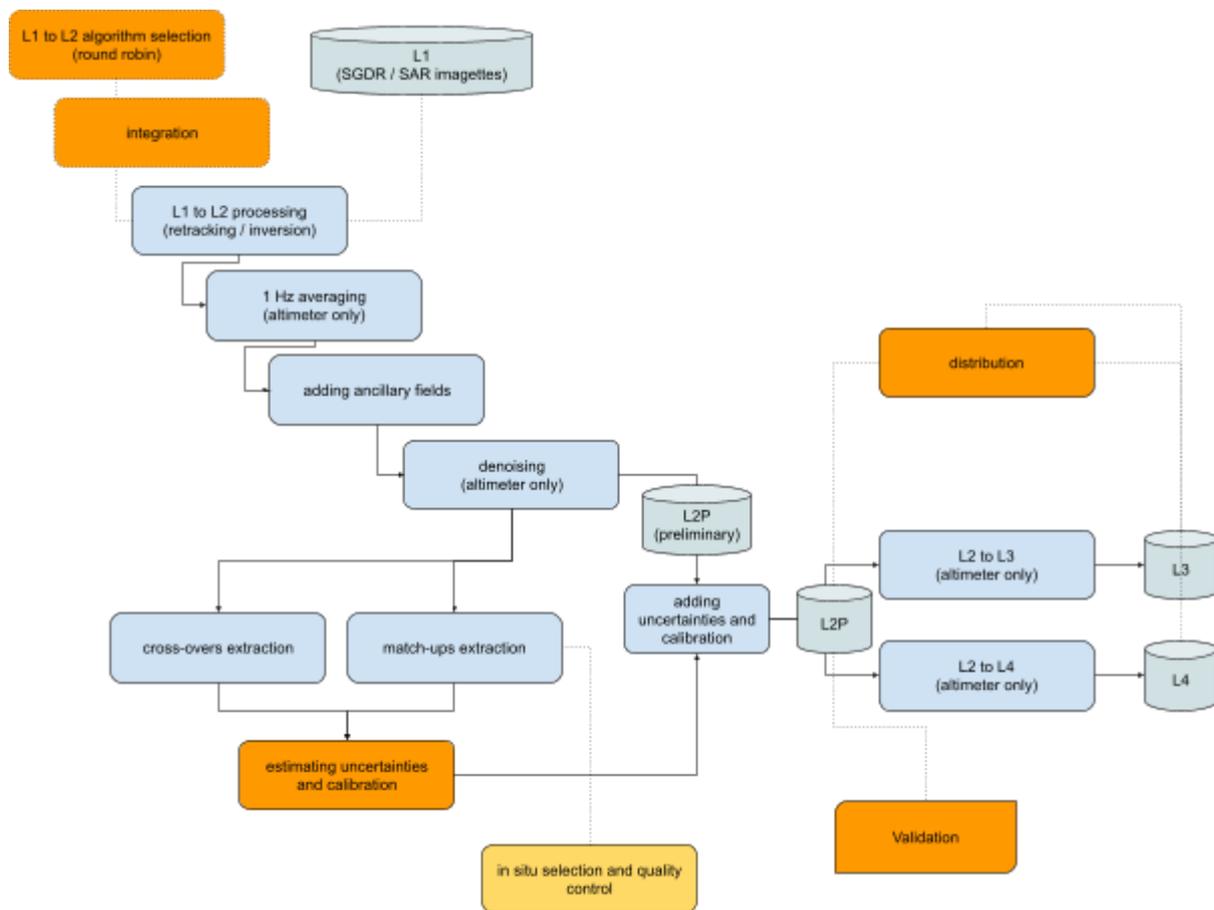
- dockerization of all processors and orchestration tools for portability to any processing platform
- access to data lakes and processing resources such as WEkEO, and demonstration of remotely operated processing

3. System specifications

This chapter describes the system developed to meet the requirements analysed in the previous chapter.

3.1 Processing workflow

The following diagram describes the overall workflow for the generation of CCI Sea State Datasets. The processing steps in light blue are purely computational steps whereas the steps in orange also require expertise and interaction among the partners.



The different steps of this workflow are detailed in the following table:

<p>L1 to L2 algorithm selection</p>	<p>Evaluation and selection round of the best algorithm for SWH retrieval from altimeter (retracker) or SAR. The different algorithms are evaluated over a selection of tracks with respect to the statistical and spectral properties of the data and comparisons with model and in situ data.</p> <p>The methodology and results described are in Product Validation and Algorithm Selection Report (PVASR) and the Round-Robin Final Report (https://drive.google.com/file/d/1vTRResptNcvecwirNO4DF-X3kawTPF9r/view?usp=sharing)</p>
<p>Integration</p>	<p>Integration step consists in implementing the selected algorithm onto the appropriate processing platform. This includes the following tasks:</p> <ul style="list-style-type: none"> • code modification and testing for portability on targeted processing platform(s) • putting the code under source control management (gitlab) • building conda environments for easy integration and deployment, and keeping the processing context of each dataset version
<p>L1 to L2 processing</p>	<p>Massively distributed reprocessing of the altimeter or SAR</p>

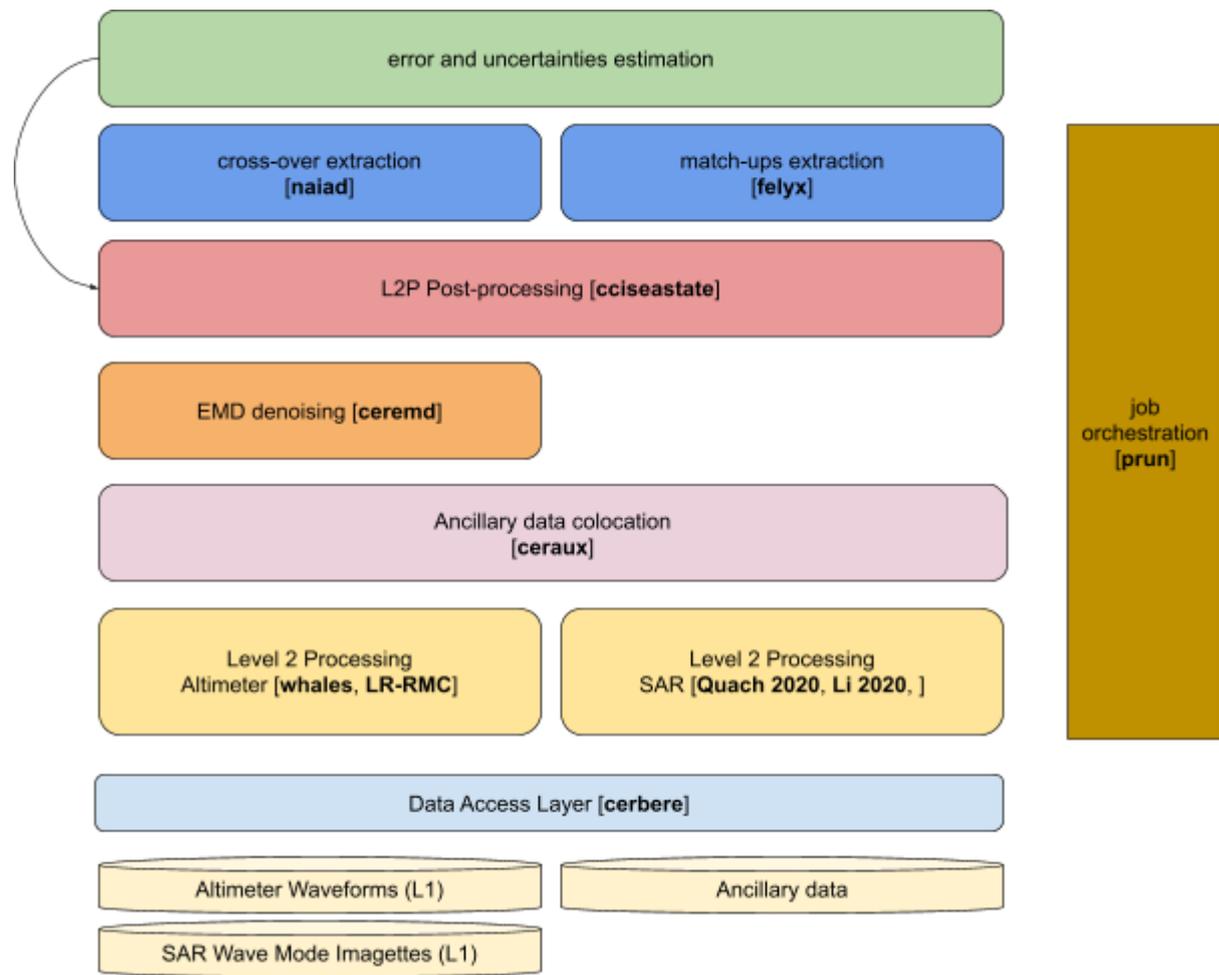
	<p>Level 1 data archive to produce a complete time series of Sea State parameters using the selected algorithms:</p> <ul style="list-style-type: none"> • WHALES for LRM altimeter missions • LR-RMC for SAR altimeter missions • Quach 2020, Li 2020 and for SAR Wave Mode <p>This step (and the following computational steps) involves the usage of dedicated tools for job array multiprocessing and monitoring of reprocessing progress and status.</p>
1 Hz Averaging	Averaging of the full resolution altimeter measurements to 1 Hz values.
Adding ancillary data	<p>colocation and addition of complementary fields including:</p> <ul style="list-style-type: none"> • SGDR/GDR variables (wind speed, radiometer water vapour content) • land mask • distance to coast • bathymetry • sea ice concentration • sea level • weather model fields (wind speed, pressure, air temperature) • wave mode fields (SWH)
denoising	<p>EMD-filter based denoising processing for along-track altimeter data. Provides EMD-filtered significant wave height, an adjusted and denoised significant wave height estimated by CCI Sea State project and based on Quilfen and Chapron, 2021.</p> <p>[Quilfen Y., Chapron B. (2020). On denoising satellite altimeter measurements for high-resolution geophysical signal analysis. <i>Advances in Space Research</i>, 68. https://doi.org/10.1016/j.asr.2020.01.005]</p>
In Situ data selection and QC	<p>selection of buoys measuring SWH from Copernicus CMEMS In Situ TAC, applying additional quality control procedures to detect:</p> <ul style="list-style-type: none"> • wrong positions • stationary measurements • low resolution measurements (≥ 0.5 meter for SWH) <p>The data are saved into a format compatible with the match-up extraction system.</p>
Cross-over extraction	<p>cross-overs between altimeter missions, against the reference mission (jason-2) for intercalibration or other pairs for verification.</p> <p>cross-over extraction system uses Naiad open-source software: https://gitlab.ifremer.fr/naiad</p>
Match-up extraction	match-ups extraction between altimeter L2P products and

	qualified reference in situ buoys from Copernicus CMEMS In Situ TAC. match-up extraction system based on <i>felyx</i> open-source software: https://gitlab.ifremer.fr/felyx
Uncertainties and calibration estimation	estimation of the cross-mission intercalibration from cross-overs/match-ups and of the SWH uncertainties from match-ups. This step requires expertise and manual analysis, and is performed by a group of experts from the team. It produces look-up tables and calibration formulation.
Adding uncertainties and calibration	implementation of the look-up tables and parametrization for the calculation of the uncertainties and SWH adjustment (cross-calibration), applied to each measurement from the L2P data.
L2 to L3 processing	Concatenation of along-track L2P into a single multi-mission daily files containing all the valid (quality = good) SWH measurements and selection of variables from the source L2P.
L2 to L4 processing	Production of monthly multi-mission statistics of SWH on a 1°x1° grid from the L2P.
Validation	Expert (manual) activity on the assessment and validation of the produced dataset before release. The methodology is described in the Product Validation Plan (PVP) and the results in the Product Validation and Intercomparison Report (PVIR)
Distribution	Push of the produced datasets to Ifremer distribution server (HTTPS and FTP) and to the ESA CCI central repository.

3.2 Processing toolboxes

The following figure shows the software layers used in CCI SeaState datasets processing, covering the different processing functions:

- accessing and reading the source data
- processing the altimeter and SAR data from L1 to L2
- adding the ancillary fields
- post-processing to L2P, L3 and L4 level (and adding quality, error and uncertainty information)
- producing cross-overs and match-ups for validation and estimation of the errors and uncertainties
- estimating the error and uncertainties
- running the different processing steps in parallel



The different packages used in each layer are referenced in the following table, with the corresponding source control repository:

cerbere	https://gitlab.ifremer.fr/cerbere/cerbere	a python unified data access API to read L1 and ancillary products in various formats
ceraux	https://gitlab.ifremer.fr/cerbere/ceraux	a python package to colocate with ancillary data such as sea-ice masks, bathymetry, land mask and distance to coast
naiad	https://gitlab.ifremer.fr/naiad/	a python framework to extract cross-overs between different satellite missions
felyx	https://gitlab.ifremer.fr/felyx/	a python framework to extract match-ups between satellite and in situ data
whales	https://gitlab.ifremer.fr/cciseastate/whales	the selected retracker for LRM altimetry missions, written in python (used for all altimeter -

		except Sentinel-3 A & B, L2P production)
ceremd	https://gitlab.ifremer.fr/cerbere/ceremd	a python package to denoise data using EMD filter
quach2020	https://github.com/agrouaze/sar_hs_nn	a python processor to produce the Sentinel-1 SAR SWH L2P
DLR2020	https://gitlab.com/dlr-earth-observation-center/cci-sea-state	a processor to produce the Sentinel-1 SAR ISSP L2P
Li2021	not distributed openly	a processor to produce the Envisat SAR ISSP L2P
LR-RMC	not distributed openly	a processor by CNES/CLS for Sentinel-3 SRAL altimeter in PLRM and SAR mode (used for S3A L2P production)
cciseastate	https://gitlab.ifremer.fr/cciseastate/cciseastate	the python post processing layer to generate full L2P, L3 and L4 products
prun		a python tool to run distributed jobs on a HPC cluster in job array - used for parallel reprocessing.

3.3 source code control

The processing software used for CCI Sea State production is versioned under source control on gitlab or github, and openly accessible whenever it is possible. When restrictions apply, they are mentioned in above table.

3.4 Processing platforms

The processing of CCI Sea State Dataset is distributed over multiple platforms, depending on the availability of the input data or how easy it is to migrate the processing software, though most of the processing was completed on Ifremer / Datarmor infrastructure. The used platform for each dataset is detailed in the following table:

CCI Sea State product	Production Platform	Processing step	Motivation
Dataset v1.1			
Altimeter	Ifremer /	All processing	

L2P	Datarmor		
Altimeter L3	Ifremer / Datarmor	All processing	
Altimeter L4	Ifremer / Datarmor	All processing	
Dataset v2			
Altimeter L2P	Ifremer / Datarmor		
	TUM		data were already preprocessed at TUM. Next extensions will be processed on Ifremer / Datarmor.
Altimeter L3	Ifremer / Datarmor	All processing	
Altimeter L4	Ifremer / Datarmor	All processing	
SAR L2P S1A &B	DLR	L1 to L2 for ISSW product	
	Ifremer / Datarmor	L1 to L2 for SWH product Post-processing	
SAR L2P Envisat	CAS / AIRI	L1 to L2 for ISSW product	
Dataset v3			
Altimeter L2P	Ifremer / Datarmor	All processing except S3A	
	CNES / HAL	L1 to L2 S3A	high cost to extract the LR-RMC retracker from the whole altimeter processing framework at CNES/CLS.
Altimeter L3	Ifremer / Datarmor	All processing	
Altimeter L4	Ifremer / Datarmor	All processing	

3.4.1 Ifremer “Datarmor” platform

Physically, the CCI production platform is mainly based on the *Datarmor* platform operated by Ifremer IT department (refer to the facility section in the management proposal), though a few reprocessing tasks were delegated to other platforms (refer to above table). The current

storage capability is about 20 PB and the available capacity largely exceeds the need for CCI products.

The reprocessing framework for CCI therefore makes use of direct access to the complete CCI input data archive on disk, physically located within the cluster, and distribution of the processing over multiple nodes of the cluster. The task of managing and distributing the processing jobs is alleviated by the use of batch tools implemented by CERSAT (and used for all its reprocessing works) such as *prun*.

Prun is a tool which aims to ease the execution and monitoring of [embarrassingly parallel](#) processings. It is a wrapper which submits jobarrays to batch schedulers (torque/maui, [oar](#), [pbspro](#)...) and manages the output logs and progress reports, with an easy monitoring.

Jobarray submission to batch schedulers is generally easy, but monitoring the job progress and accessing to the error logs of a few tasks among thousands is the same as finding a needle in a haystack... This wrapper was created to avoid spending more time in manual "logfile-mining" (grep, tail...) than the processing time itself. Some other needs it addresses:

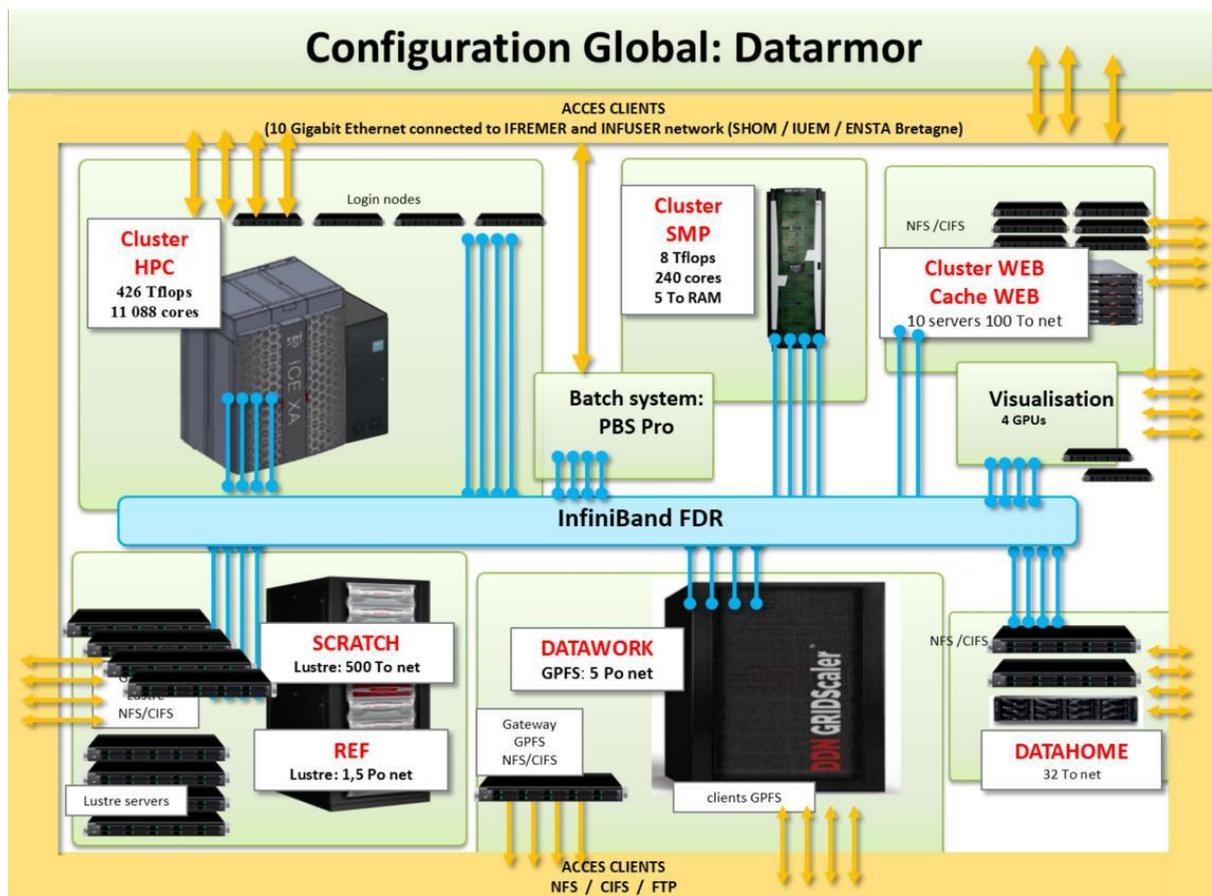
- listing management avoids too big jobarrays (which crashes some batch schedulers)
- list the tasks in errors to ease their reprocessing
- allowing several execution modes : sequential, streaming (pipe)
- allowing multiple batch scheduler as backend
- having tasks status reporting and real-time monitoring
- having job history, meta-data & log files organization

The Ifremer supercomputer, *Datarmor*, provides scalable capabilities for:

- CPU intensive or memory demanding applications and processing
- massive data storage

It consists of:

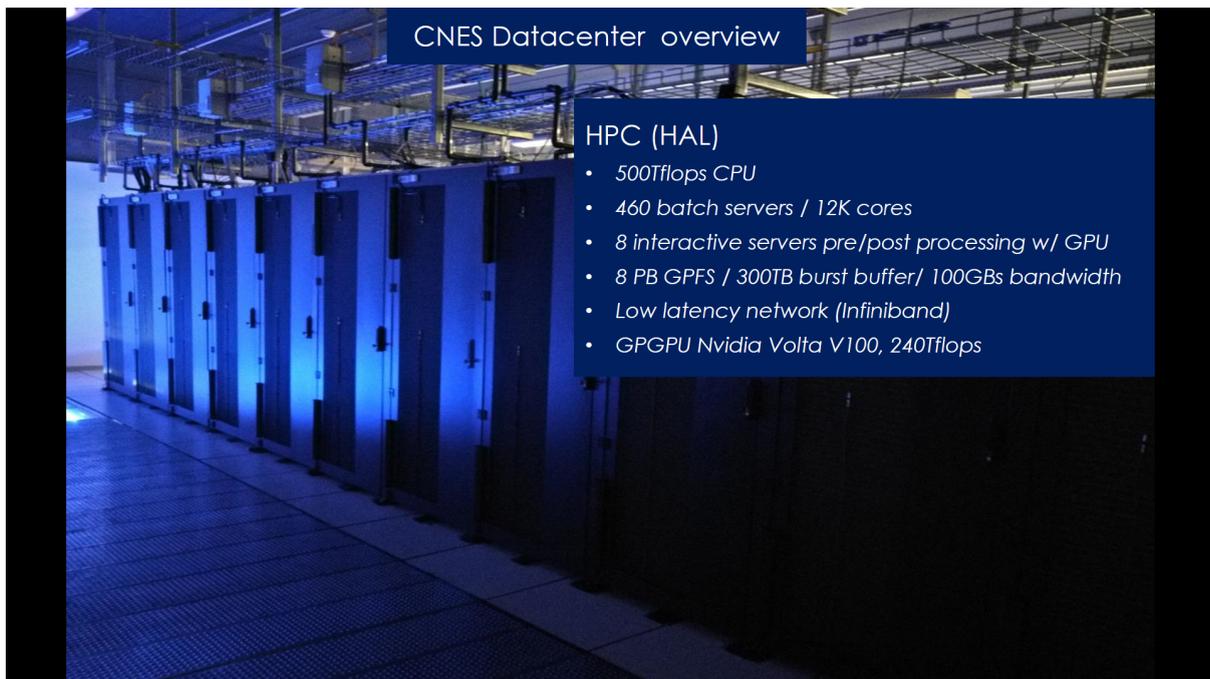
- 11088 cores - 426 Tflops (128 GB memory and 28 cores per node) for HPC applications
- 240 cores, 5 TB RAM for non MPI processing
- 10 servers for web services
- large data storage capacity
 - 500 TB Lustre for HPC
 - 1.5 PB Luster storage for reference data
 - 5 PB GPFS storage for project and work data
 - 100 TB for services and web applications
 - 32 TB per home directory



1.5 PB storage is currently reserved for CERSAT processing in this infrastructure which is deemed to be sufficient for the CCI Sea State requirements.

3.4.2 CNES HAL

The CNES HPC center is already used for some altimetry reprocessing campaigns for AVISO project. It is used in CCI Sea State for the reprocessing of Sentinel-3 altimeter data.



3.5 Product distribution

The CCI Sea State Datasets are available on three different servers for different stages of the dataset life cycle, as described in the following table:

Server	Access	Usage
Partner Access	FTP access: ftp://eftp.ifremer.fr login: e0321bf password: incitent-professeraït-culbutees	Shared access for project partners before public release for assessment and validation
User Preliminary Access	FTP access: ftp://eftp.ifremer.fr login: pe31b4c password: yellowsubrolling	Preliminary access to data for early releases and updates before they are integrated into CCI Portal
CCI Portal	HTTP Access: https://climate.esa.int/en/projects/sea-state/data/	Permanent public user access

Specific DOIs are associated to each dataset version and product level by the CCI Data Portal's Technical Team (e.. for CCI Sea State Dataset version 2, specific DOIs are minted independently for L2P, L3 and L4 products).

The data organization follows the recommendation of *CCI Data Standards version 2.0* and are described in the *Product Specification Document*.