Climate Change Initiative+ (CCI+) Phase 2

Sea Surface Salinity

[D2.5]Product Validation Protocol (PVP)

Customer: ESA

Ref.: ESA-EOP-SC-AMT-2021-26

Version: v2.0

Ref. internal: 4000123663/18/I-NB

Revision Date: 19/06/2023

Filename: SSS_cci_PHASE#02_D2.5-PVP-v2r0.docx

Deliverable code: D2.5

Signatures

	Name	Signature	Date
AUTHOR	Fabrice Bonjean LOCEAN		22/05/2023
REVIEWED BY	Jacqueline Boutin LOCEAN		
	Rafael Catany (Project Manager) ARGANS		
	Roberto Sabia (Technical Officer) ESA		
APPROVED BY	Jacqueline Boutin (Science Leader) LOCEAN		
	Nicolas Reul (Science Leader) IFREMER		
	Rafael Catany (Project Manager) ARGANS		
ACCEPTED BY	Roberto Sabia (Technical Officer) ESA		

DIFFUSION LIST
Sea Surface Salinity Team Members
ESA (Roberto Sabia, Susanne Mecklenburg)

Amendment Record Sheet

	Document Change Record	
Date / Issue	Description	Section / Page
15-07-2019 / v1.0	Deliver to ESA	New document
19-07-2019/v1.0	Corrected some typos	Whole document
01-12-2019/v1.1	Addressed several issues	Whole document
05-05-2023/2.0	New Document. Start CCI+SSS Phase#02	Whole document

Table of Contents

List of Figures

List of tables

Acronyms

ATBD	Algorithm Theoretical Basis Document
CCI	The ESA Climate Change Initiative (CCI) is formally known as the Global Monitoring for Essential Climate Variables (GMECV) element of the European Earth Watch programme
CCI+	Climate Change Initiative Extension (CCI+) is an extension of the CCI over the period 2017–2024
CEOS	Committee on Earth Observation Satellites
CMEMS	Copernicus Marine Environmental Monitoring Service
DOI	Digital Object Identifier
ECV	Essential Climate Variable
EO	Earth Observation
FCDR	Fundamental Climate Data Record
FIDUCEO	Fidelity and uncertainty in climate data records from Earth Observations
FRM	Fiducial Reference Measurements
ISAS	In-situ Analysis System
ISDB	in-situ database (of Fiducial Reference Measurements and satellite measurements)
NASA	National Aeronautics and Space Administration
NOAA	National Oceanic and Atmospheric Administration
OA	Objective Analysis
OI	Optimal Interpolation
PI MEP	Pilot Mission Exploitation Platform
PUG	Product User Guide
PVASR	Product Validation and Algorithm Selection Report
PVIR	Product Validation and Intercomparison Report
PVP	Product Validation Plan
QA4EO	Quality Assurance Framework for Earth Observation
SMAP	Soil Moisture Active Passive [mission of NASA)
SMOS	Soil Moisture and Ocean Salinity [satellite of ESA]

SSS	Sea Surface Salinity
WOA	World Ocean Atlas

1 Introduction

1.1 Purpose and scope

The Product Validation Plan (PVP) encompasses a comprehensive list of reference datasets dedicated to validating each type of Sea Surface Salinity (SSS) product. As per the requirements stated in the Statement of Work (Task 2 SOW ref. ESA-EOP-SC-AMT-2021-23), the PVP is an integral part of the CCI+SSS phase 2 project, outlined in the document version v1.1.

This document outlines the validation protocol for assessing the accuracy of the SSS products derived from the ESA CCI+SSS phase 2 project, particularly when compared with alternative SSS sources. It builds upon the foundation laid by the PVP from CCI+SSS phase 1. Moreover, it provides explicit guidelines on utilising in-situ data to validate remote sensing products under the GEO/CEOS Quality Assurance framework for Earth Observation (QA4EO).

The primary objective of this PVP is to establish the appropriate selection of Fiducial Reference Measurements (FRM) for the comprehensive validation of satellite-derived SSS. FRMs are independent, fully characterised, and traceable ground measurements that play a crucial role in delivering the necessary confidence in data products to users. By adhering to the FRM definition provided by the European Space Agency (ESA) at https://earth.esa.int/web/sppa/activities/frm, the PVP ensures that the chosen FRMs yield maximum return on investment by providing independent validation results and estimating satellite measurement uncertainty throughout the entirety of a satellite mission's end-to-end duration.

To be eligible for a Fiducial Reference Measurement (FRM), an in-situ measurement must meet the following criteria:

- It should have documented evidence demonstrating its traceability to the International System of Measurements (ISM).
- It should be independent of the satellite geophysical retrieval process.
- It should have a comprehensive uncertainty budget for the instrumentation and measurement process, accounting for various usage conditions
- It should adhere to community-approved measurement protocols and management practices.

When writing this document, there was no established guideline explicitly addressing the use of FRM-compliant in-situ measurements for validating SSS satellite retrievals. Hence, one of the primary objectives of this initial PVP is to define the SSS FRM. This document will primarily focus on points 2 and 3 mentioned above, with point 1 considered a given due to its inherent nature. Achieving consensus through documents like this PVP will aid in accomplishing point 4, involving accepted practices and metrics for comparisons, such as in-situ salinity measurements and other types of measurements like structural and correlation measurements (to be addressed in future PVPs).

Validating remote sensing products using in-situ reference measurements is inherently challenging due to significant differences between satellite-based instrument measurements and other measurement types. For example, apart from distinct instrumental error characteristics, remote sensing measurements vary in spatial resolution, temporal scope, and representativeness (e.g., spatial integration). Consequently, comparing remote sensing products with other data types necessitates quality control and accounting for diverse sources of uncertainty in ground truth measurements. Moreover, it requires considering natural SSS variability sampled differently by various instruments to provide a meaningful assessment of the quality of remote sensing products.

The proposed approach for validating SSS products, including their uncertainties, during the initial phase of the CCI phase 2 project is outlined in the current version of PVP. This PVP builds upon the foundation established in the CCI+SSS phase 1 PVP V1 and expands upon the validation methodologies employed during that phase. The enhancements and considerations included in this PVP are as follows:

- PI-MEP facilities are extensively utilised to access in-situ databases and facilitate CCI collocation.
- In-situ data used for validation and algorithm adjustments are differentiated.
- The uncertainty arising from the "mismatch" between in-situ and CCI SSS sampling is further explored and considered.
- The spatiotemporal resolution capabilities of the CCI SSS products are evaluated more deeply.
- Polar products are validated additionally.
- First time validation of C-band derived SSS.

By incorporating these enhancements and considerations into the validation process, the PVP aims to strengthen the validation methodologies and improve the overall quality and performance of the SSS products within the CCI+SSS phase 2 project.

1.2 Structure of the document

This document is structured into six sections, each addressing a specific aspect of PVP:

Section 1 provides an overview of the document's purpose and scope.

Section 2 presents a comprehensive overview of the theoretical framework underlying this PVP.

Section 3 defines a valid scheme for implementing FRM of SSS within the validation process.

Finally, section 4 offers an overview of the reference in-situ data set utilised in the validation procedures.

Section 5 focuses on the specific validations for regions where particular CCI+SSS products will be delivered.

Finally, section 6 explores the integration of the previous sections within the SMOS PI-MEP.

Section 7 describes the approach and methodology for implementing this PVP, specifically the validation with in-situ measurements, throughout the entire duration of the CCI+SSS project.

1.3 Applicable Document

DSTD	CCI Data Standards, CCI-PRGM-EOPS-TN-13-0009	V2.1, 25/03/2019
SRD	System Requirement Document	SSS_cci-D3.1-SRD-i1r5
SSD	System Specification Document	SSS_cci-D3.2-SSD-i1r0
URD	User Requirement Document	SSS_cci-D1.1-URD-i1r0
DARD	Data Access Requirement Document	SSS_cci-D1.3-DARD-v1r3
PSD	Product Specification Document	SSS_cci-D1.2-PSD-v1r4
SoW	CCI+ Statement of Work	
ATBD	Algorithm Theoretical Baseline Document	SSS_cci-D2.3-ATBD_L3_L4-i1r0_v1.1

Table 1 – Applicable documents

1.4 Reference Document

ID	Document	Reference
RD01	L.R. Centurioni, V. Hormann, Y. Chao, G. Reverdin, J. Font & D.K. Lee (2015). Sea surface salinity observations with Lagrangian drifters in the tropical North Atlantic during SPURS: Circulation, fluxes, and comparisons with remotely sensed salinity from Aquarius**.** *Oceanography***, 28 (1): 96-105**	
RD02	G. Reverdin, S. Morisset, L. Marié, D. Bourras, G. Sutherland, B. Ward, J. Salvador, J. Font, Y. Cuypers, L.R. Centurioni, V. Hormann, N. Koldziejczyk, J. Boutin, F. D'Ovidio, F. Nencioli, N. Martin, D. Diverres, G. Alory & R. Lumpkin (2015). Surface salinity in the North Atlantic subtropical gyre during the STRASSE/SPURS summer 2012 cruise. *Oceanography* **28** (1): 114-123	
RD03	N. Hoareau, A. Turiel, M. Portabella, J. Ballabrera-Poy & J. Vogelzang (2018). Singularity Power Spectra: A Method to Assess Geophysical Consistency of Gridded Products - Application to Sea-Surface Salinity Remote Sensing Maps. IEEE Transactions on Geosciences and Remote Sensing 56, 5525-5536	
RD04	Boutin, J., Y. Chao, W.E. Asher, T. Delcroix, R. Drucker, K. Drushka, N. Kolodziejczyk, T. Lee, N. Reul, G. Reverdin, J. Schanze, A. Soloviev, L. Yu, J. Anderson, L. Brucker, E. Dinnat, A.S. Garcia, W.L. Jones, C. Maes, T. Meissner, W. Tang, N. Vinogradova, B. Ward (2016b), Satellite and In-situ Salinity: Understanding Near-surface Stratification and Sub-footprint Variability, Bulletin of American Meteorological Society, 97(10), doi: 10.1175/BAMS-D-15-00032.1.	
RD05	In-situ database Analyses Report. PI-MEP Consortium. March 15, 2019.	https://pimep.ifremer.fr/diffusion/analyses/insitu-database/report/pimep-insitu-report_20190315.pdf
RD06	Houndegnonto, O. J., Kolodziejczyk, N., Maes, C., Bourlès, B., Da-Allada, C. Y., & Reul, N. (2021). Seasonal variability of freshwater plumes in the eastern Gulf of Guinea as inferred from satellite measurements. Journal of Geophysical Research: Oceans, 126, e2020JC017041. https://doi.org/10.1029/2020JC017041	
RD07	Merchant, C. J., Paul, F., Popp, T., Ablain, M., Bontemps, S., Defourny, P., Hollmann, R., Lavergne, T., Laeng, A., de Leeuw, G., Mittal, J., Poulsen, C., Povey, A. C., Reuter, M., Sathyendranath, S., Sandven, S., Sofieva, V. F., and Wagner, W.: Uncertainty information in climate data records from Earth observation, Earth Syst. Sci. Data, 9, 511–527, https://doi.org/10.5194/essd-9-511-2017, 2017.	

ID	Document	Reference
RD08	Guimbard, S.; Reul, N.; Sabia, R.; Herlédan, S.; Khoury Hanna, Z.E.; Piollé, J.-F.; Paul, F.; Lee, T.; Schanze, J.J.; Bingham, F.M.; Le Vine, D.; Vinogradova-Shiffer, N.; Mecklenburg, S.; Scipal, K.; Laur, H. The Salinity Pilot-Mission Exploitation Platform (Pi-MEP): A Hub for Validation and Exploitation of Satellite Sea Surface Salinity Data. *Remote Sens.* **2021**, *13*, 4600. https://doi.org/10.3390/rs13224600	
RD09	Bingham, F.M.; Brodnitz, S.; Fournier, S.; Ulfsax, K.; Hayashi, A.; Zhang, H. Sea Surface Salinity Subfootprint Variability from a Global High-Resolution Model. *Remote Sens.* **2021**, *13*, 4410. https://doi.org/10.3390/rs13214410	
RD10	Boutin, J., N. Reul, J. Koehler, A. Martin, R. Catany, S. Guimbard, F. Rouffi, et al. (2021), Satellite-Based Sea Surface Salinity Designed for Ocean and Climate Studies, Journal of Geophysical Research: Oceans, 126(11), e2021JC017676, doi:https://doi.org/10.1029/2021JC017676.	

Table 2 – Reference documents

2 Overview

Validation of an oceanic variable product necessitates comparing it to an external, independent reference known as a ground truth. As outlined in condition 2 of the FRM definition, this comparison typically involves utilising different data sources that can be assumed to represent the actual value closely. However, ensuring the independence of these datasets can be challenging. In some instances, the derivation of the product may have utilised ancillary information that closely aligns with the ground truth data. For example, in the case of SSS, there is a well-established functional relationship with SST at appropriate scales.

In some cases, the precision of the data may be influenced by the source material used to produce the product. For example, this can occur when the data is obtained from a numerical model that depends on climatology for calibration, and the same climatology is employed to commence or restrict product extraction. Therefore, to ensure a statistically meaningful and unbiased comparison between the product and ground truth, it is crucial to establish complete independence in the data used for both types. This requires thorough documentation of the data generation processes for product and ground truth, as specified by the traceability condition in quality assurance for Earth observation (QA4EO). However, achieving this complete independence is only sometimes straightforward or well-documented.

When comparing a product to a ground truth reference, a significant challenge arises from the assumption that the ground truth is a perfect reference, which is rarely the case. Such comparisons are inherently subject to two types of uncertainties:

- Class 1 Uncertainty: Accuracy and precision errors
- Class 2 Uncertainty: Representativeness errors

The first class of uncertainty (accuracy and precision errors) relates to the quality of acquiring or generating the ground truth and is typically well-documented. It encompasses the following aspects:

- For instrument-based measurements, it involves considering the instrument specifications provided by the manufacturer, such as absolute accuracy and granted lifetime.

- For interpolated products, such as those obtained through Objective Analysis (OA) or Optimal Interpolation (OI), it involves constructing an error matrix that accounts for the propagation of errors from the source data through the interpolation scheme.

- Numerical models involve estimating model errors using appropriate error propagation schemes.

However, these estimates of Class 1 Uncertainty are theoretical and represent an ideal situation. In addition, real-world factors contribute additional sources of uncertainty, including:

- Fouling, drifts, poor quality control, and other issues affecting in-situ instrumental data.

- When dealing with interpolated data, there may be issues with sampling inhomogeneities, sampling biases, poor estimation of correlation radii and matrices, and the inability to accurately describe exceptional events.

- Numerical instabilities, spin-up effects, poor representation of physical processes, and other factors affecting numerical models.

To accurately determine the magnitude of Class 1 Uncertainties, it is crucial to properly preprocess the ground truth data by performing self-consistency checks and error assessments. However, standardised procedures are only available for making these types of estimates. We will present some recommendations to evaluate Class 1 Uncertainties for SSS in-situ measurements in Section 3.3.1.

The second class of uncertainty (representativeness errors) pertains to the discrepancy between product values and the ground truth due to differences in spatial and temporal scales represented by each data type. For example, remote sensing SSS products typically provide average values over relatively large spatial areas sampled at regular intervals of several days (except for level 2 products with acquisition times of a few seconds). On the other hand, in-situ data are always associated with very small spatial areas (typically a few centimetres horizontally and vertically) and represent instantaneous measurements obtained within seconds or less.

Given the geophysical variability of SSS, it is expected that the difference between an in-situ value and the corresponding remote sensing SSS value, considering their typical spatial and temporal resolution, exhibits a range of variability of 0.2 practical salinity units (pss) or greater [RD01, RD02]. In highly dynamic regions like river plumes, this difference can exceed 0.5 pss [RD06]. Therefore, the proper characterisation of Class 2 Uncertainty, as demonstrated by RD07, is crucial for determining the significance of observed differences between in-situ and remote sensing SSS measurements.

3 Definition of the validation protocol

3.1 Introduction to SSS product validation within the QA4EO Guidelines

In line with the QA4EO guidelines, the validation of SSS products using ground truth data necessitates the following measures:

1. Standardisation of Reference Data: To ensure consistency and comparability, it is vital to establish standardised reference data. This involves defining a reference (measurement) standard, ideally including a stated uncertainty. The reference can be individually or collectively defined within the scientific community.
2. Ensured Traceability of Products and Validation Datasets: Traceability is critical to SSS product validation. It refers to establishing a documented and unbroken chain of calibrations, enabling the measurement results to be linked back to a reference. This traceability instils confidence in the products' and validation datasets' reliability and accuracy.
3. Well-Characterised Uncertainty of Reference Data: A comprehensive understanding of the uncertainty associated with the reference data is crucial. Uncertainty quantifies the dispersion of quantity values attributed to a measurement. Ideally, it should be derived from experimental evaluations, but it can also be estimated based on other relevant information or experience.
4. Meaningful Quality Indicators: Quality indicators play a vital role in assessing the suitability of SSS products for specific applications. They provide users with quantitative information about the traceability of the data to an agreed reference or measurement standard, such as ISM. Quality indicators can be presented as numerical values or descriptive text if they link quantitatively to the defined reference.

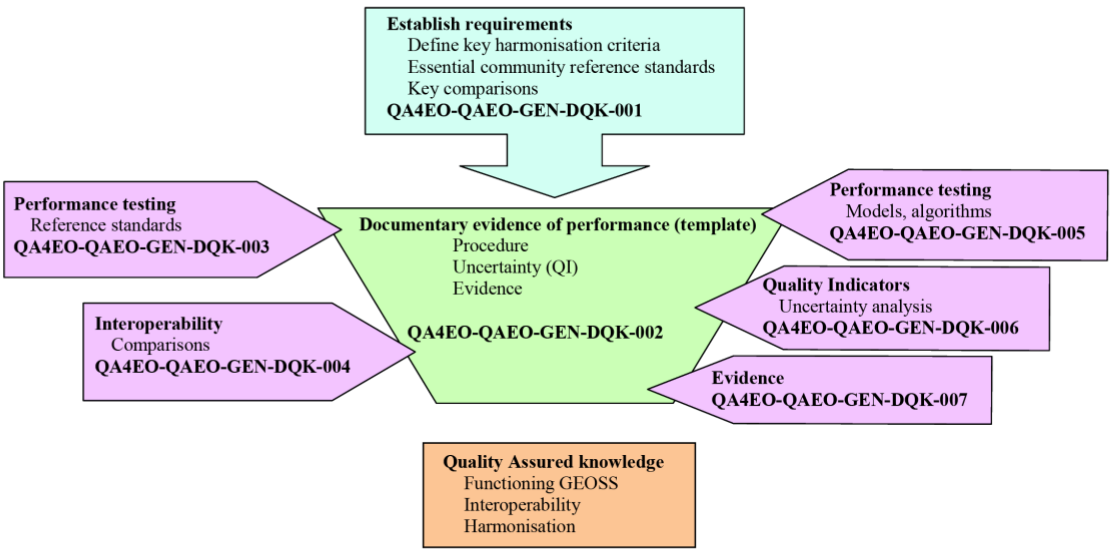

Figure 1 Summary of data quality aspects in QA4EO.

To fulfil these four criteria, we will establish a validation strategy that encompasses standardised reference data, ensures traceability, characterises uncertainty in the reference data, and incorporates meaningful quality indicators.

3.2 Validation strategy overview

- **Adhere to Community Standards**: To ensure data reliability and consistency, it is recommended to adhere to community-endorsed standards for data selection. Given the role of PI-MEP, we suggest using the in-situ datasets recorded by PI-MEP, following their quality control procedures. If necessary, recommendations for including new datasets in PI-MEP or revising their quality control protocols can be made (refer to "Interaction with PI-MEP"). A case-by-case discussion is required for other data sources to determine the most appropriate datasets. These datasets should have recorded traceability and uncertainty estimates as a minimum requirement (see Section 3.3).
- **Ensure Traceability**: Validation datasets, including in-situ and other data sources, should have well-documented traceability records detailing how the data was generated. It is important to avoid including data directly or indirectly used in generating the remote sensing SSS product to the extent possible. If there is a need to include such data, it should be explicitly noted in the validation metrics and justified. For example, while specific statistics of ISAS SSS are used to construct the CCI+SSS product, they are not expected to influence the interannual or longitudinal variability of the product.
- **Characterization of Uncertainties**: Prior to computing quality metrics, assessing both Class 1 and Class 2 uncertainties is essential. This characterization provides the necessary information to determine the significance level of the metrics.
- **Compute Quality Metrics**: Appropriate methods will be used to compare ground truth data and remote sensing products. Normalizing the salinities by their uncertainties (Class 1 and Class 2) is necessary to validate both the SSS values and their associated uncertainties.

3.3 Metric for validation diagnosis

3.3.1 Assessing Class 1 Uncertainties in the ground truth

Ground data sources used for validation, adhering to QA4EO guidelines, provide a traceable record of uncertainties encompassing biases and random errors. These errors are referred to as identified errors and are accounted for in the validation process. However, as mentioned earlier, additional errors may stem from instrument degradation, incomplete statistical descriptions, or limitations in the models employed. These errors, known as unidentified errors, cannot be predetermined or quantified in advance. To estimate these errors, the intercomparison of comparable ground truth datasets becomes crucial, provided enough such datasets are available. This intercomparison allows for quantifying these unidentified errors, contributing to a more comprehensive understanding of the overall error budget in validating SSS products.

There is no definitive or universally correct method for intercomparing different datasets representing the same variable. For this PVP, we propose the following approach for intercomparing ground truth datasets based on the type of data:

- *In-situ data:* Since different sets of in-situ data will never be perfectly equivalent, there will always be spatial and temporal differences between them. Considering Class 2 Uncertainties (explained in the next section), this additional source of error should be considered. For example, when computing the standard deviation of errors, the intercomparison error between two datasets, Dataset 1 and Dataset 2, each with identified standard deviations of errors σ1 and σ2, can be calculated as:

$$\varepsilon_{12}^2 = \sigma_{12}^2 - \sigma_1^2 - \sigma_2^2 - r_{12}^2$$

  Here, ε_{12} represents the intercomparison error of datasets 1 and 2, σ_{12} is the error of the difference (specifically, the standard deviation of the difference between datasets 1 and 2), and r_{12} is the standard deviation of the representativeness error (explained in the next section). Again, these errors are assumed to be independent of each other.
- *Interpolated data:* The considerations mentioned above apply when comparing datasets representing different spatial or temporal scales. If the intercompared data are assumed to represent the same scales, the representativeness error r_{12} can be considered zero.
- *Outputs from numerical models:* The same principles as for interpolated data apply.

Notice that the intercomparison error ε_{12}^2 represents the sum of the unidentified errors in Dataset 1 and Dataset 2, $\varepsilon_{12}^2 = x_1^2 + x_2^2$. The precise contribution of each dataset to the intercomparison error cannot be determined, so it is proposed that this error is attributed proportionally to the identified error. For instance, in the example mentioned above, the unidentified errors for each dataset can be calculated as follows:

$$x_1^2 = \frac{\sigma_1^2 \varepsilon_{12}^2}{\sigma_1^2 + \sigma_2^2} \frac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2} \quad ; \quad x_2^2 = \frac{\sigma_2^2 \varepsilon_{12}^2}{\sigma_1^2 + \sigma_2^2} \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2}$$

If multiple unidentified errors are estimated for the same datasets, the arithmetic mean of all the errors will be taken.

The final total error for a given dataset will be the sum of the identified and unidentified errors, $\varepsilon^2=\sigma^2+x^2$.

In the context of CCI+SSS, PIMEP and the validation team conduct rigorous quality checks on in-situ datasets, which are vital for validating satellite products.

3.3.2 Assessing Class 2 Uncertainties in the ground truth

The magnitude of geophysical variability in in-situ SSS data within the spatial and temporal scales of remote sensing products is influenced by several factors, including the specific spatial resolution and time window of the remote sensing products. It also depends on the region being studied, as inter-regional variability can be significant.

To address representativeness uncertainty, Pi-MEP colocations, led by Nicolas Reul and Sebastien Guimbard, will employ two methodologies:

- RD09 provide high-resolution simulations projected onto SMOS and SMAP Level 2 grids. These simulations will be utilised to estimate Class 2 Uncertainties.
- Clovis Thouvenin Masson's work on Level 3/4, using updated GLORYS reanalysis from 2010 to now, will also contribute to estimating uncertainties.

The estimates derived from these two methodologies will be compared.

While vertical representativeness errors can be significant in some cases, estimating them poses considerable challenges. For example, vertical stratification on the scale of a few centimetres, which differentiates satellite-derived SSS from near-surface salinity measured by buoys, can occur due to factors such as persistent weak winds or freshwater lenses. However, characterising this stratification accurately requires detailed information on surface wind stress and ocean currents, which remains complex without a dedicated product.

On the other hand, the issue of rain lenses can be quickly addressed. The most significant impact is the freshening effect immediately after rainfall, typically within one hour. To mitigate this effect, satellite-derived SSS data obtained within 0.5 hours after a significant rainfall event (detected by IMERG) exceeding 1mm/hr can be discarded. Similarly, in-situ SSS data can be disregarded if IMERG detects rainfall rates above 1mm/hr [RD04].

3.3.3 Quality metrics

The proposed quality metrics are as follows:

- Mean difference (bias): This metric is estimated by calculating the average difference between the value of the remote sensing product and the corresponding ground truth dataset. A minimum of 30 independent samples is required for statistical significance, following the criterion based on the Central Limit Theorem. The mean difference provides information about systematic biases in the remote sensing product.
- The median of difference (robust bias): The median of the difference between the remote sensing product and the ground truth dataset is considered a robust measure of bias.
- The standard deviation of difference (random error): This metric is calculated by determining the standard deviation of the difference between the remote sensing product and the ground truth dataset. Like the mean difference, a minimum of 30 independent samples is necessary for statistical significance. The standard deviation reflects random errors in the remote sensing product.
- The robust standard deviation of difference (STD*): STD*(x) is defined as $median(|x - median(x)|/_{0.67}$. This metric helps filter out outliers and focuses on the central part of the statistical distribution of differences (e.g., see Figure 7 bottom right in [RD10]).

- Root mean square difference (total error): The root mean square difference is obtained by taking the square root of the sum of squares of the mean difference and standard deviation. It provides total error information, including systematic and random components.
- Correlation coefficient (Pearson and Spearman): The correlation coefficient, which can be either the Pearson coefficient or the Spearman coefficient, provides information about the degree of linearity between the remote sensing and ground truth data. Pearson coefficient is determined by the ratio of the covariance of the two data types to the product of their standard deviations when using the values of each data type. On the other hand, the Spearman coefficient is obtained by using the rank of the values for each data type. Both correlation coefficients indicate the linearity between the variables being compared, with a perfect correlation coefficient being 1. Pearson coefficient is commonly used but can be influenced by clustered points on either side of the distribution, leading to a false perception of good predictability.

  In contrast, the Spearman coefficient is more robust but generally yields lower values than the Pearson coefficient. In addition, it does not provide information about the magnitude of the error. Therefore, it is recommended to utilise both correlation coefficients. To determine significant linearity, a Pearson coefficient above 0.8 (corresponding to an error variance below 36%) and a Spearman coefficient above 0.5 (equivalent to an error rank variance below 75%) can be considered. This approach ensures that a satisfactory level of linearity is achieved while considering each coefficient's limitations and characteristics.

- Linear regression of the difference vs signal: This metric examines the correlation between the value difference between the remote sensing product and the ground truth data and the value of the remote sensing product itself. This regression's slope, intercept, and Pearson correlation coefficient provide information about the relationship between the error and the signal. A good slope with a correlation coefficient above 0.8 indicates a well-characterised linear relationship.
- Skewness and kurtosis: These metrics quantify the deviation of the difference distribution from a normal distribution, providing insights into the distribution's shape.
- Reduced centred difference: Differences normalised by their Class 1 and Class 2 uncertainties variances, along with the corresponding standard deviation, are used for this metric. In addition, it includes the sampling/mismatch uncertainty.

These metrics will be derived from the following sets of differential data:

- The match-up of Level 2 CCI+SSS with direct in-situ SSS data
- The match-up of Level 3 and Level 4 CCI+SSS with well-sampled direct in-situ measurements within their observation domain
- The match-up of Level 3 and Level 4 CCI+SSS with in-situ interpolated fields

It is recommended to analyse the statistics as a function of the distance to the coast. Additionally, assessing the spatial coverage of the satellite products and comparing the ice filtering with independent ice edge maps should be considered.

3.3.4 Choice of ground truth

The selection of ground truth data for validating remote sensing SSS products depends on several factors, including the specific product being validated, its intended purpose, and the region of interest. In addition, when utilising in-situ datasets, it is essential to consider errors arising from the spatial and temporal mismatch and differences in the spatial and temporal integration or undersampling between in-situ and satellite products.

Interpolated ground truth sources may be used for validation, but it is crucial to ensure that all their errors are well-characterised before their inclusion. For example, in the case of L4 validation, operational numerical models with data assimilation can be considered, provided that the L4 product is not generated using the same numerical model and the absolute errors of the models are well understood.

The choice of ground truth matching depends on the specific application and objectives, such as characterising seasonality or assessing anomalies. It is important to note that each application requires an appropriate type of remote sensing data (e.g., L3 products with approximately 1-month time resolution for seasonality assessment and L2 products for anomaly detection). For example, suppose a remote sensing product unsuitable for a particular application is used. In that case, additional post-processing steps, such as temporal low-pass filtering or accounting for detailed representativeness errors, may be necessary. However, in some cases, the desired brand of remote sensing product may not be available for the given application, requiring using the closest alternative while considering expected deviations.

3.3.5 Validation Issues: Inhomogeneous Sampling of in-situ data

There are several approaches to consider when evaluating a CCI+SSS L4 product with an inhomogeneous sampling of in-situ data. We propose the following methods, along with their advantages and drawbacks:

1. Subsampling: A possible strategy is to choose a subset of in-situ data that corresponds to the grid cells of the L4 product. This can be accomplished by selecting and aggregating or averaging the in-situ measurements within each grid cell, ensuring adequate data points for reliable estimates. One straightforward approach is randomly selecting a single measurement from each grid cell.

   Subsampling in this manner can create a comparable dataset for validation.

   Drawbacks:

   a) Loss of Spatial Information: Aggregating or averaging in-situ measurements within grid cells removes spatial variability and fine-scale information, limiting the capture of localized variations and patterns.
   b) Impact on Extreme Values: Aggregation may dilute or mask extreme values, affecting the representation of variability in the gridded data.
   c) Potential Bias: Spatial aggregation may introduce bias if in-situ measurements are clustered or exhibit spatial trends, leading to biased validation results.

Next, we propose using a Monte Carlo method coupled with subsampling to address these drawbacks.

2. By combining Monte Carlo simulation with the subsampling approach mentioned earlier (1.), it is possible to address the challenges associated with inhomogeneous sampling and overcome some of the drawbacks of the subsampling method.

   Here is a step-by-step process for incorporating Monte Carlo simulation into the validation:

   A. Random Sampling: Generate multiple random subsamples from the available in-situ dataset, representing different spatial and temporal distributions of data points.
   B. Metric Calculation: Calculate the desired validation metrics for each subsample by comparing the gridded data product to the corresponding subset of in-situ data.
   C. Aggregation: Collect the metric values obtained from each subsample, creating a distribution of metric values.
   D. Statistical Analysis: Perform a comprehensive statistical distribution analysis to estimate the overall performance and uncertainty. This analysis can involve calculating summary statistics, such as mean or standard deviation, or constructing confidence intervals to quantify the variability and reliability of the validation metrics.

   Repeating these steps allows a Monte Carlo estimate of validation metrics to account for bias and uncertainty resulting from inhomogeneous sampling.

   Figure 2 depicts a visual representation of the approach. In each subsampling iteration, a chosen metric (e.g., the standard deviation of the difference) is computed, and by conducting a sufficiently large number of trials, a distribution of that metric is obtained. This distribution can offer valuable information, such as a 95% confidence interval, mean, standard deviation, and other relevant statistics. This process resembles bootstrapping, wherein the aim is to obtain a reliable estimate of the metric by repeatedly sampling from the available data.

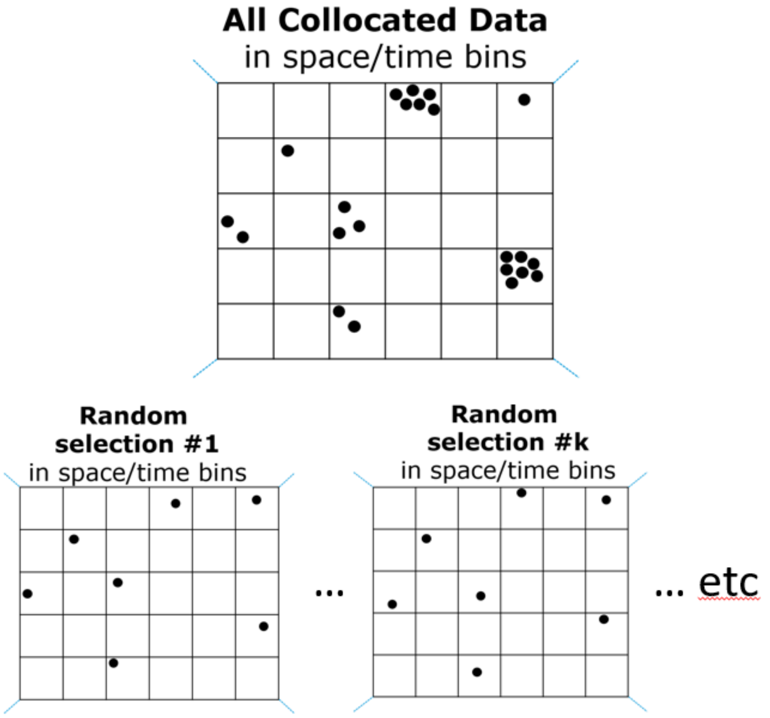

Figure 2: Visual representation of the Monte Carlo subsampling approach. The top panel illustrates the initial sampling, while the bottom panel showcases random subsampling realisations. In each subsampling realisation, only one data point is selected from each populated bin.

3. Interpolation: Another option is to use interpolation techniques to estimate values at grid cell locations based on available in-situ measurements. This generates a gridded dataset for direct comparison with the L4 data product.

   Drawbacks:

a) Mismatched Scales: Interpolation assumes that in-situ data variability aligns with the grid cell scale, leading to inaccuracies if scales do not match.
b) Spatial Sampling Density: The interpolation accuracy depends on the density and distribution of in-situ data, potentially resulting in less accurate results in sparsely sampled regions.
c) Uncertainty Assessment: Interpolation methods often need more explicit certainty quantification, limiting robust conclusions about agreement/disagreement.

4. Area-based validation: Area-based validation can be used when a point-to-grid comparison is challenging due to significant sampling differences. Aggregating L4 and in-situ data to a standard (larger) spatial unit allows meaningful comparison.

   Drawbacks:

   a) Loss of Spatial Detail: Aggregation may smooth or average out spatial variability, resulting in local information loss.

b) Scale Mismatch: The choice of spatial units for aggregation affects results, with potential mismatches impacting validity.
c) Sensitivity to Aggregation Method: Different aggregation techniques may yield varied results, requiring careful selection.
d) Sensitivity to Grid Cell Size: The choice of grid cell size for aggregation can impact the validation outcomes, requiring careful consideration.

It is important to note that the choice of method depends on factors like data nature and spatial distribution of in-situ measurements.

4 Reference data sets

Defining the reference datasets for validation is a crucial aspect of the validation protocol for new datasets. The reference datasets must undergo quality control and meet the QA4EO guidelines to be considered acceptable as FRMs (Reference Measurement Systems). As we continue establishing the criteria for SSS FRMs, we suggest following an established quality control facility like Pi-MEP for in-situ data. In addition, we recommend utilising the resources available in Copernicus Marine Core Services for interpolated maps and outputs from numerical models.

4.1 In-situ measurements, including SSS FRMs

The primary in-situ datasets can be classified as follows:

- Close-to-surface Argo
- Surface measurements from SSS drifters.
- Quality-controlled TSG transects
- Long series of SSS from mooring observation systems.
- The close-to-surface measurement from mounted instruments on marine mammals.
- Surface measurements from sail drones and glider-mounted instruments.

To ensure the quality of the products, we suggest obtaining the data from Pi-MEP (https://www.salinity-pimep.org/). In addition, a comprehensive report detailing their quality and limitations can also be accessed at [RD05] and [RD08].

4.2 Interpolated data sets

We recommend using the following interpolated sets of in-situ SSS for validation purposes, listed in order of preference:

1. Delayed mode ISAS products created by LOPS laboratory: These products offer comprehensive quality-controlled Argo profiles and other in-situ measurements from various sources such as ships of opportunity, research ships, sailing ships, surface drifters, and marine mammals. They provide reliable and thoroughly validated SSS data at a 5m depth.
2. Delayed mode ISAS product available on Copernicus Marine Environment Service: If the LOPS product is unavailable for a particular period, we suggest using the delayed mode ISAS product accessible through the Copernicus Marine Environment Service. This product, CORA OA SSS, provides surface-level SSS derived from objective analysis using a range of in-situ data sources, predominantly Argo floats. They can be accessed at: http://marine.copernicus.eu/services-portfolio/access-to-products/?option=com_csw&view=details&product_id=INSITU_GLO_TS_OA_REP_OBSERVATIONS_013_002_b
3. Near Real-Time (NRT) products available on Copernicus Marine Environment Service: In cases where neither the LOPS product nor the delayed mode ISAS product is available, the NRT products provided by the Copernicus Marine Environment Service can be used. These NRT products offer timely SSS data derived from various in-situ sources. They can be

accessed at: http://marine.copernicus.eu/services-portfolio/access-to-products/?option=com_csw&view=details&product_id=INSITU_GLO_TS_OA_NRT_OBSERVATIONS_013_002_a

4.3 Outputs from numerical models

Regarding interpolated datasets, we suggest accessing the Copernicus Marine Environment Service to obtain relevant data. However, it is worth noting that the choice of products may vary depending on the region and specific application requirements. While several products are available with different resolutions, we consider the 0.25˚ daily GLORYS reanalysis the most suitable reference dataset. It can be accessed at: http://marine.copernicus.eu/services-portfolio/access-to-products/?option=com_csw&view=details&product_id=GLOBAL_REANALYSIS_PHY_001_025. This dataset has been extensively evaluated and demonstrated to accurately represent the structural and spectral characteristics of Sea Surface Salinity (SSS) [RD03].

4.4 Sea Ice Edge

For evaluating the filtering of sea ice, we recommend using either the OSI SAF sea ice concentration dataset (SIC) based on SMMR/SSMI/SSMIS or CCI SIC dataset based on AMSR E and AMSR 2, both available at https://cds.climate.copernicus.eu/cdsapp#!/dataset/satellite-sea-ice-concentration?tab=form. These datasets provide valuable information for assessing the distance to the sea ice edge. Sébastien Guimbard at Pi-MEP has incorporated the OSI SAF ice edge (with a resolution of 25km and 12km for the ice edge) into the evaluation process.

The appropriate value of satellite-derived SIC to characterise the ice edge between sea ice and the free ocean is typically around a 15% to 30% SIC threshold.

5 Regional Validations

In addition to validating the global products following the recommendations detailed in the previous sections, regional products will be delivered during CCI+SSS phase 2, which will require specific validations.

5.1 Arctic Ocean

As part of the CCI+SSS phase 2, polar products on the polar EASE 2 grid will be provided in addition to global products. It is essential to conduct a specific validation effort in the Arctic Ocean, considering the unique characteristics of this region. Therefore, the Arctic option requires dedicated attention.

The Arctic seas exhibit regional variability in SSS and have varying in-situ sampling patterns. For example, the Barents and Nordic Seas have a higher availability of Argo and ICES data for validation than other seas, where data collection primarily relies on TSG and sporadic cruises.

A region-specific approach should be implemented to validate the SSS products in the Arctic, focusing on areas with different types of variabilities. Here are three suggested regions for validation (Figure 3):

1. Beaufort and Chukchi Seas: These regions experience substantial SSS variability. Validating SSS products in these areas will provide insights into accurately capturing and representing this variability.
2. Eurasian Shelf (East Siberian, Laptev, and Kara Seas): Like the Beaufort and Chukchi Seas, the Eurasian Shelf regions exhibit significant SSS variability. Including these areas in the validation, effort will enhance the understanding of regional SSS dynamics.
3. Nordic (Norwegian and Greenland) and Barents Seas: While the SSS variability in these seas is comparatively weaker than in the previous regions, validating the products in these areas is still important.

The accuracy and reliability of the SSS products in the Arctic Ocean can be assessed effectively by conducting validation in these specific regions characterized by varying types of variability.

Figure 3: Circled Arctic regions where specific validation can be carried out for the accuracy of the CCI+SSS and polar products. 1. Beaufort and Chukchi Seas. 2. Eurasian Shelf, including East Siberian, Laptev and Kara Seas. 3. Nordic and Barents Seas, including Norwegian and Greenland Seas.

An alternative approach could involve dividing the data based on SSS levels and categorising them into different salinity ranges, such as weak, medium, and large salinities. These salinity ranges can be associated with SSS variability regimes, with the most substantial SSS variability observed in areas with the weakest SSS. Parting the data in this manner allows for a more focused analysis of SSS variability within specific salinity ranges.

Data availability:

- Regarding the availability of in-situ data, there is a limited number of Argo floats in the Nordic Seas and Bay of Baffin. The ICES data can be utilised as an alternative source in these regions. We recommend integrating the ICES data into the Pi-MEP platform for comprehensive data coverage.
- Additionally, the CNSS TSG data in the Nordic Seas and data from the research vessel Polarstern can contribute to the validation efforts.
- It would be highly beneficial to gather additional data from drifters, gliders, and the Danish coastguards to enhance the characterisation of SSS in the highly variable region around Greenland. This additional data collection would provide valuable insights into the variability of SSS in this specific area and contribute to a more comprehensive understanding of the region's SSS dynamics.

Sea-ice filtering: See Section 4.4.

5.2 Antarctic Ocean

In addition to the Argo float data, there are various other types of in-situ data available in the Antarctic region, including data collected along TSG tracks as depicted in TSG tracks spanning from 2010 to 2022, with each type of TSG cruise represented by a different colour. These data were extracted from the Pi-MEP facility. Furthermore, a substantial amount of data from two ships equipped with SAMOS platforms must be incorporated into Pi-MEP.

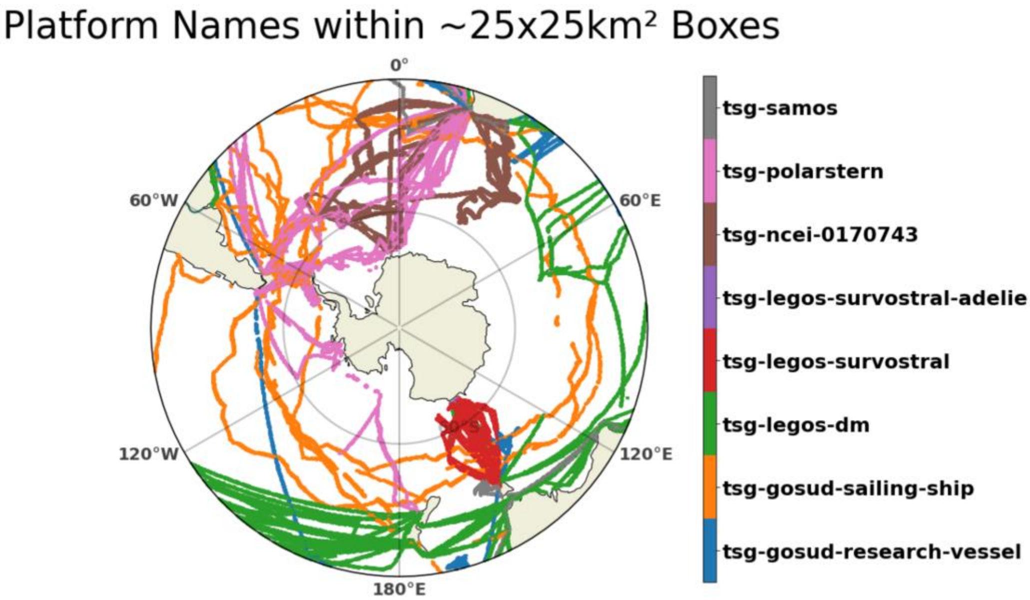

Figure 4: TSG tracks spanning from 2010 to 2022, with each type of TSG cruise represented by a different colour. These data were extracted from the Pi-MEP facility.

Sea-ice filtering: See Section 4.4.

5.3 River plumes (C/X band products)

Four areas will be considered for analysis:

- Gulf of Mexico,
- Gulf of Guinea,
- Amazon Mouth, and
- Bay of Bengal.

It is recommended to perform colocation with Argo data from EN4, as the CORA dataset needs more quality checking.

Similar issues as encountered with L band data are expected, including:

1) Uncertainty in representativeness (as discussed in Section 3.3.2),
2) RFI pollution, and
3) Coastal contamination.

It is advised to calculate statistics based on the distance to the coast.

Due to limited coverage of in-situ data, conducting intercomparison with various models such as Glorys (1/12, 2v4), ORAS5, UKMO, and CMMC may provide additional valuable insights into SSS variability.

Figure 5 and Figure 6 (courtesy N. Reul) give an example of diagnostics obtained in the Bay of Bengal.

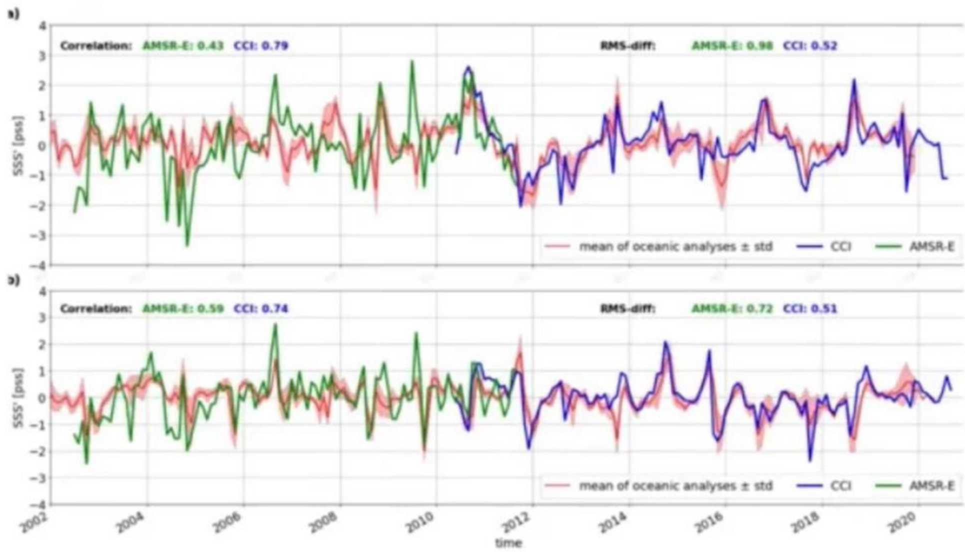

Figure 5: Time series of SSS from (Green) the AMSR-E-derived SSS, (Blue) CCIV3.2, and (Red) the ensemble mean of several ocean reanalyses in the Bay of Bengal.

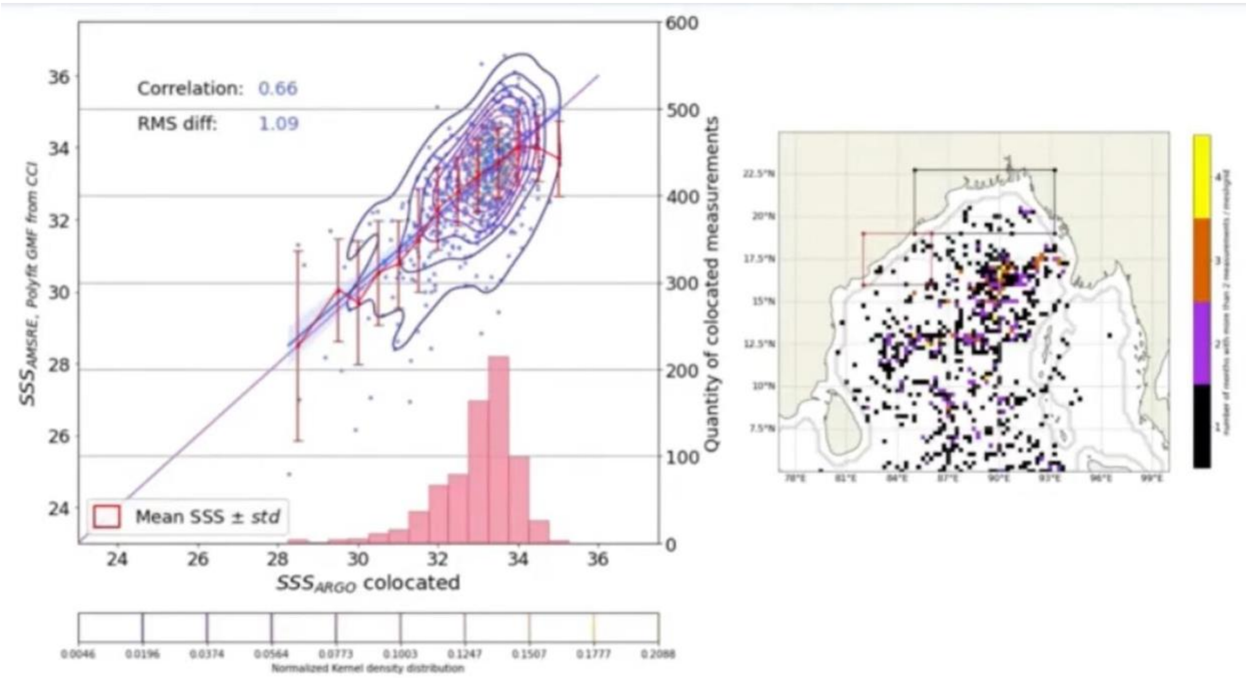

Figure 6: Scatterplot analysis between collocated SSS from Argo and AMSR-E-derived SSS in the Bay of Bengal.

6 Integration in the Pilot Mission Exploitation Platform (Pi-MEP)

6.1 Evaluating the Suitability and Effectiveness of SSS Validation Activities within PI-MEP

The metrics outlined in this PVP closely resemble those currently employed by Pi-MEP, with the primary distinction being the inclusion of representativeness and unidentified error considerations. Consequently, after incorporating these factors (which are crucial for assigning significance levels to statistical tests), the standard validation procedures utilised by PI-MEP can be applied.

6.2 Establishing Standardised SSS Validation Protocols to Accommodate PI-MEP Quality Control

The primary challenge in standardising validation protocols lies in conducting the necessary tests to verify the suitability of the existing protocols for estimating errors in the ground truth. Per Pi-MEP quality control guidelines, these errors can be utilised as a threshold for significance levels or to determine the confidence interval for the correspondence between the ground truth and remote sensing product.

6.3 Integration of Standardised Quality-Controlled SSS Validation Procedures into the PI-MEP Validation System

The integration process becomes straightforward once the Pi-MEP system validates the procedures for accurately estimating errors in ground data.

6.4 Evaluation by End-users

To ensure the effectiveness of the error and quality metrics provided, it is crucial to involve expert users in oceanography. Their assessment will help determine whether the metrics align with their expectations. In the event of significant deviations, conducting a survey to identify potential conflicts or deviations becomes necessary. The survey aims to quantify the problems in the products considered significant, allowing for a focused investigation into the root causes of these issues. The survey should include questions regarding expected absolute values, gradients, spatial and temporal positions of frontal zones, observed biases, and more. By gathering insights from end-users, the evaluation process can better address any observed problems and refine the system's overall performance.

7 PVP implementation

7.1 Timeframe for Implementation

The implementation of quality metrics for SSS validation follows specific temporal planning. Once the in-situ dataset is identified and compiled, implementing the quality metrics takes two weeks to one month. The metrics are tested and fine-tuned during this time to ensure their effectiveness and accuracy.

After the metrics are in place, producing validation reports for the entire period takes approximately two weeks. This includes analysing the data, generating the necessary visualisations, and summarising the findings in a comprehensive report.

Overall, the temporal planning for SSS validation involves approximately six to eight weeks, with specific durations allocated for different stages, such as metric implementation, testing, and report generation.

7.2 Resources

There is no need for additional resources for the original validation plan in this ITT.

7.3 Contingency Plan

No obstacles are anticipated regarding data access, as it has been consistently granted over the years. The only challenge identified pertains to obtaining negative estimates for the standard deviation of unidentified errors. In such cases, these estimates must be treated as zero, indicating a negligible or unquantifiable error. According to the principles of error analysis, the standard deviation of the difference between two independent measurements should be the sum of their standard deviations. However, if other sources of independent error, such as representativeness or intercomparison, are present, this standard deviation may increase.

Concerning the PVP itself, a review will be necessary if specific criteria for defining SSS FRM are provided in the future, such as through a dedicated white paper. This ensures that the PVP remains aligned with the latest guidelines and requirements for SSS validation.

End of document