Copernicus Cal/Val Solution

D3.2 - Recommendations for R&D on Cal/Val Methods

Ref: CCVS.ARG.D3.2

Version: 1.1

Date: 29/01/2023

GA 101004242

Document Reference

	Company
Work Package	3
WP leader	S. Clerc (ACRI-ST)
Deliverable Title	
Delivery Number	3.2
Delivery Type	Report
Dissemination Level	Public
Version	1.1
Date of issue	29/01/2023
Lead beneficiary	ARGANS
Lead authors	B. Alhammoud (ARGANS), S. Clerc (ACRI-ST), T. Verhoelst, J-C Lambert (BIRA-IASB), F. Nencioli, G. Hajduch (CLS), A. Meygret, C. Pierangelo, M. Raynal (CNES), B. Pflug, M. Bachmann (DLR), N. Gobron (JRC), F. Boersma (KNMI)
Contributors	M. Hallow, C. Smithers (ARGANS), L. Bourg (ACRI-ST), S. Compernolle, M. De Mazière, I. De Smedt, D. Hubert, A. Keppens, B. Langerock, G. Pinardi, M.K. Sha, N. Theys, M. Van Roozendael, C. Vigourou (BIRA-IASB), S. Labroue (CLS), T. Guinle (CNES), G. de Leeuw (FMI), A. Ludewig, G. Tilstra (KNMI), B. Mota (NPL), T. Trent (U. Leicester)
Reviewed by	S. Clerc, L. Bourg (ACRI-ST), P. Strobl (JRC), B. Mota, S. Hunt (NPL), Y. Govaerts (Rayference)

Changes Log

Version	Date	Changes
Draft 1	02/12/2021	Initial Version for Half-Project Review
Draft 1.1	02/02/2022	Updated Version after Half-Project Review
Draft 1.2	12/05/2022	Updated Version after task-progress meeting
V1.0	02/06/2022	Initial Delivery
1.1	29/01/2023	Revised version (editorial)

This project has received funding from the European Union's Horizon 2020 programme under grant agreement No 101004242.

Table of Content

List of Figures

Aucune entrée de table d'illustration n'a été trouvée.

List of Tables

1. Introduction

1.1. Scope of the document

This document presents a gap analysis of the methods used in the calibration and validation of Earth Observation satellites relevant to the Copernicus programme and suggests recommendations for the research and developments required to fulfil this gap when/where possible.

The document identifies the gaps and limitations of the CalVal methods, used for calibration and validation (CalVal) activities for the current Copernicus missions. It will also address the development needs for future Copernicus missions. Four types of missions are covered based on the division used in the rest of the CCVS project: optical, altimetry, radar and microwave and atmospheric composition.

Finally, it will give a prioritized list of recommendations for R&D activities on the CalVal methods.

The information included is mainly collected from the deliverables of work packages 1 and 2 in the CCVS project and from the consortium experts in CalVal activities.

1.2. Differences between Gap and Limitations

General distinction between limitations and gaps:

- Limitations mentioned in RD 2.8 and RD 2.9 are draw-backs of methods that cannot be corrected.
- Gaps are shortcomings of methods that could be mitigated with increased effort or alternative solutions; or no method currently existing.

1.3. Definition of CalVal methods

In the context of remote sensing, the CEOS defines the **calibration** as “the process of quantitatively defining a system’s response to known and controlled signal inputs” and the **validation** as “the process of assessing, by independent means, the quality of the data products derived from those system outputs”

1.4. Structure of the document

The document is divided into seven sections.

Section 1 describes the scope of the document, outlines the structure of this document, identifies the reference documents, and explains the used acronyms and abbreviations.

Sections 2, 3, 4 and 5 are devoted to the gap analysis in CalVal methods for Optical missions (Section 2), Altimetry missions (Section 3), Radar and Microwave missions (Section 4) and Atmospheric composition missions (Section 5).

Section 6 is dedicated to the synergy between the different sensors and missions

Section 7 gives the conclusions and the prioritized list of recommendations for R&D activities needed on the CalVal methods.

1.5. Reference documents

1.5.1. CCVS Deliverables

RD1.5 - Optical Missions Cal/Val requirements, ref: CCVS.ACR.D1.5

RD1.6 - Altimetry Missions Cal-Val Requirement, ref: CCVS.CLS.D1.6

RD1.7 - Radar and Microwave Imaging Missions Cal/Val requirements, ref: CCVS.CLS.D1.7

RD1.8 – Atmospheric Composition Missions Cal/Val Requirements, ref: CCVS.BIRA.D1.8

RD2.8 - Vicarious methods on natural targets, ref: CCVS.ARG.D2.8 v1.0

RD2.9 - Inter satellite comparisons, ref: CCVS.CLS.D2.9 v1.0

RD3.1- Recommendations for R&D on Instrumentation Technologies, ref : CCVS.TAR.D3.1

1.5.2. Other Reference documents

Alonso, K.; Bachmann, M.; Burch, K.; Carmona, E.; Cerra, D.; de los Reyes, R.; Dietrich, D.; Heiden, U.; Hölderlin, A.; Ickes, J.; Knodt, U.; Krutz, D.; Lester, H.; Müller, R.; Pagnutti, M.; Reinartz, P.; Richter, R.; Ryan, R.; Sebastian, I.; Tegler, M. Data Products, Quality and Validation of the DLR Earth Sensing Imaging Spectrometer (DESIS). *Sensors* **2019**, *19*, 4471. https://doi.org/10.3390/s19204471

Bachmann, Martin und Alonso, Kevin und Carmona, Emiliano und Heiden, Uta und Marshall, David und Müller, Rupert und de los Reyes, Raquel (2021) *The spectral and radiometric quality of the DESIS data products and the influences on higher-level processing.* 1st DESIS User Workshop, 28. September - 01. OKtober 2021, Online. https://desis2021.welcome-manager.de/archiv/web/userfiles/desis2021/DESIS_Bachmann_etal_Tue1615.pdf

Bhatt, R., Doelling, D., Scarino, B., Haney, C., and Gopalan, A., 2017, "Development of Seasonal BRDF Models to Extend the Use of Deep Convective Clouds as Invariant Targets for Satellite SWIR-Band Calibration" Remote Sensing Vol. 9, No. 10, pp 1061, 2072-4292

Boynard, A., Hurtmans, D., Garane, K., Goutail, F., Hadji-Lazaro, J., Koukouli, M. E., Wespes, C., Vigouroux, C., Keppens, A., Pommereau, J.-P., Pazmino, A., Balis, D., Loyola, D., Valks, P., Sussmann, R., Smale, D., Coheur, P.-F., and Clerbaux, C.: Validation of the IASI FORLI/EUMETSAT ozone products using satellite (GOME-2), ground-based (Brewer–Dobson, SAOZ, FTIR) and ozonesonde measurements, Atmos. Meas. Tech., 11, 5125–5152, https://doi.org/10.5194/amt-11-5125-2018, 2018.

EUFAR DJ2.1.2 - Beekhuizen, J., M. Bachmann, E. Ben-Dor, J. Biesemans, M. Grant, G.B.M. Heuvelink, A. Hueni, M. Kneubuehler, E. de Miguel, A. Pimstein, E. Prado, I. Reusen, T. Ruhtz, M. Schaale 2009: Report on full error propagation concept. EUFAR FP7 JRA2 deliverable DJ2.1.2.)

Franquesa, M., Vanderhoof, M.K., Stavrakoudis, D., Gitas, I.Z., Roteta, E., Padilla, M. and Chuvieco, E., 2020. Development of a standard database of reference sites for validating global burned area products. *Earth System Science Data*, *12*(4), pp.3229-3246.

Gascon, F.; Bouzinac, C.; Thépaut, O.; Jung, M.; Francesconi, B.; Louis, J.; Lonjou, V.; Lafrance, B.; Massera, S.; Gaudel-Vacaresse, A.; Languille, F.; Alhammoud, B.; Viallefont, F.; Pflug, B.; Bieniarz, J.; Clerc, S.; Pessiot, L.; Trémas, T.; Cadau, E.; De Bonis, R.; Isola, C.; Martimort, P.; Fernandez, V. Copernicus Sentinel-2A Calibration and Products Validation Status. Remote Sens. 2017, 9, 584. https://doi.org/10.3390/rs9060584

Gobron N., Morgan O., Adams J., Brown L.A., Cappucci F., Dash J., Lanconelli C., Marioni M., and Robustelli M. Evaluation of Sentinel-3A and Sentinel-3B Ocean Land Colour Instrument Green Instantaneous Fraction of Absorbed Photosynthetically Active Radiation.Remote Sensing of Environment, 2022, 270 (112850). DOI: 10.1016/j.rse.2021.112850

Guanter, L.; Kaufmann, H.; Segl, K.; Foerster, S.; Rogass, C.; Chabrillat, S.; Kuester, T.; Hollstein, A.; Rossner, G.; Chlebek, C.; Straif, C.; Fischer, S.; Schrader, S.; Storch, T.; Heiden, U.; Mueller, A.; Bachmann, M.; Mühle, H.; Müller, R.; Habermeyer, M.; Ohndorf, A.; Hill, J.; Buddenbaum, H.; Hostert, P.; Van der Linden, S.; Leitão, P.J.; Rabe, A.; Doerffer, R.; Krasemann, H.; Xi, H.; Mauser, W.; Hank, T.; Locherer, M.; Rast, M.; Staenz, K.; Sang, B. The EnMAP Spaceborne Imaging Spectroscopy Mission for Earth Observation. *Remote Sens.* **2015**, *7*, 8830-8857. https://doi.org/10.3390/rs70708830

Guanter, L., Richter, R., Moreno, J. (2006): Spectral calibration of hyperspectral imagery using atmospheric absorption features. - Applied Optics, 45, 10, 2360-2370. https://doi.org/10.1364/AO.45.002360I

Granger, S. L., Leroy, S. S., Manning, E. M., Fetzer, E. J., Oliphant, R. B., Braverman, A., ... & Lambrigtsen, B. H. (2004, September). Development of level 3 (gridded) products for the Atmospheric Infrared Sounder (AIRS). In IGARSS 2004. 2004 IEEE International Geoscience and Remote Sensing Symposium (Vol. 4, pp. 2506-2509). IEEE.

Kleipool, Q., Ludewig, A., Babić, L., Bartstra, R., Braak, R., Dierssen, W., Dewitte, P.-J., Kenter, P., Landzaat, R., Leloux, J., Loots, E., Meijering, P., van der Plas, E., Rozemeijer, N., Schepers, D., Schiavini, D., Smeets, J., Vacanti, G., Vonk, F., and Veefkind, P.: Pre-launch calibration results of the TROPOMI payload on-board the Sentinel-5 Precursor satellite, Atmos. Meas. Tech., 11, 6439–6479, https://doi.org/10.5194/amt-11-6439-2018, 2018.

Khakurel, P.; Leigh, L.; Kaewmanee, M.; Pinto, C.T. Extended Pseudo Invariant Calibration Site-Based Trend-to-Trend Cross-Calibration of Optical Satellite Sensors. *Remote Sens.* **2021**, *13*, 1545. https://doi.org/10.3390/rs13081545

Loew A., W. Bell, L. Brocca, C.E. Bulgin, J. Burdanowitz, X. Calbet, R.V. Donner, D. Ghent, A. Gruber, T. Kaminski, J. Kinzel, C. Klepp, J.C. Lambert, G. Schaepman-Strub, M. Schröder, T. Verhoelst Validation practices for satellite-based earth observation data across communities Rev. Geophys., 55 (2017), pp. 779-817, 10.1002/2017RG000562

Ludewig, A., Kleipool, Q., Bartstra, R., Landzaat, R., Leloux, J., Loots, E., Meijering, P., van der Plas, E., Rozemeijer, N., Vonk, F., and Veefkind, P.: In-flight calibration results of the TROPOMI payload on board the Sentinel-5 Precursor satellite, Atmos. Meas. Tech., 13, 3561–3580, https://doi.org/10.5194/amt-13-3561-2020, 2020.

Mota, B., Gobron, N., and Wooster, M., Inter-comparison of four operational satellite Fire Radiative Power (FRP) products: A spatial and temporal consistency assessment., 2020. doi:10.5194/egusphere-egu2020-8385.

Tilstra, L.G., van Soest, G., and Stammes, P.: Method for in-flight satellite calibration in the ultraviolet using radiative transfer calculations, with application to Scanning Imaging Absorption Spectrometer for Atmospheric Chartography (SCIAMACHY), J. Geophys. Res., 110, D18311, https://doi.org/10.1029/2005JD005853 , 2005.

Tilstra, L.G., de Graaf, M., Aben, I., and Stammes, P.: In-flight degradation correction of SCIAMACHY UV reflectances and Absorbing Aerosol Index, J. Geophys. Res., 117, D06209, https://doi.org/10.1029/2011JD016957, 2012.

Tilstra, L.G., de Graaf, M., Wang, P., and Stammes, P.: In-orbit Earth reflectance validation of TROPOMI on board the Sentinel-5 Precursor satellite, Atmos. Meas. Tech., 13, 4479–4497, https://doi.org/10.5194/amt-13-4479-2020, 2020.Gaps in atmosphere Cal/Val vicarious methods (BIRA/JRC/KNMI/CNES)

CEOS AC/VC white paper, A constellation architecture to monitor carbon dioxide and methane from space. 2018.

Kataoka, Fumie; Crisp, David; Taylor, Thomas; O'Dell, Christopher; Kuze, Akihiko; Shiomi, Kei ; Suto, Hiroshi; Bruegge, Carol; Schwandner, Florian; Rosenberg, Robert; Chapsky, Lars; Lee, Richard. (2017). The Cross-Calibration of Spectral Radiances and Cross-Validation of CO2 Estimates from GOSAT and OCO-2. Remote Sensing. 9. 1158. 10.3390/rs9111158.

Chevallier, F.: On the statistical optimality of CO2 atmospheric inversions assimilating CO2 column retrievals, Atmos. Chem. Phys., 15, 11133–11145, https://doi.org/10.5194/acp-15-11133-2015, 2015.

Tu, Q., Hase, F., Blumenstock, T., Kivi, R., Heikkinen, P., Sha, M. K., Raffalski, U., Landgraf, J., Lorente, A., Borsdorff, T., Chen, H., Dietrich, F., and Chen, J.: Intercomparison of atmospheric CO2 and CH4 abundances on regional scales in boreal areas using Copernicus Atmosphere Monitoring Service (CAMS) analysis, COllaborative Carbon Column Observing Network (COCCON) spectrometers, and Sentinel-5 Precursor satellite observations, Atmos. Meas. Tech., 13, 4751–4771, https://doi.org/10.5194/amt-13-4751-2020, 2020.

Inness, A., Flemming, J., Heue, K.-P., Lerot, C., Loyola, D., Ribas, R., Valks, P., van Roozendael, M., Xu, J., and Zimmer, W.: Monitoring and assimilation tests with TROPOMI data in the CAMS system: near-real-time total column ozone, Atmos. Chem. Phys., 19, 3939–3962, https://doi.org/10.5194/acp-19-3939-2019, 2019.

Keppens, A., Compernolle, S., Verhoelst, T., Hubert, D., and Lambert, J.-C.: Harmonization and comparison of vertically resolved atmospheric state observations: methods, effects, and uncertainty budget, Atmos. Meas. Tech., 12, 4379–4391, https://doi.org/10.5194/amt-12-4379-2019, 2019.

O'Dell, C. W., Eldering, A., Wennberg, et al: Improved retrievals of carbon dioxide from Orbiting Carbon Observatory-2 with the version 8 ACOS algorithm, Atmos. Meas. Tech., 11, 6539–6576, https://doi.org/10.5194/amt-11-6539-2018, 2018.

Wizenberg, T., Strong, K., Walker, K., Lutsch, E., Borsdorff, T., and Landgraf, J.: Inter-comparison of CO measurements from TROPOMI, ACE-FTS, and a high-Arctic ground-based FTS, Atmos. Meas. Tech. Discuss. [preprint], https://doi.org/10.5194/amt-2021-190, in review, 2021.

Zhu Sifeng, Xingfeng Chen, Li Liu, Lili Qie, Zhengqiang Li, Jinji Ma, Shule Ge, Jin Hong, Xin Li, and Hailiang Gao "Evaluation of radiometric performance of MODIS visible bands using the Rayleigh scattering method," Journal of Applied Remote Sensing 13(1), 018503 (22 March 2019). https://doi.org/10.1117/1.JRS.13.018503

Padilla, M., Stehman, S.V., Hantson, S., Oliva, P., Alonso-Canas, I., Bradley, A., Tansey, K., Mota, B., Pereira, J.M., Chuvieco, E. (2015) Comparing the Accuracies of Remote Sensing Global Burned Area Products using Stratified Random Sampling and Estimation. Remote Sensing of Environment 160, 114-121, https://doi.org/10.1016/j.rse.2015.01.005

Padilla, M., Olofsson, P., Stehman, S.V., Tansey, K., Chuvieco, E. (2017) Stratification and sample allocation for reference burned area data. Remote Sensing of Environment 203, 240-255, https://doi.org/10.1016/j.rse.2017.06.041

Boschetti, L., Stehman, S.V. and Roy, D.P., 2016. A stratified random sampling design in space and time for regional to global scale burned area product validation. Remote sensing of environment, 186, pp.465-478.

Rodgers, C. D., & Connor, B. J. (2003). Intercomparison of remote sounding instruments. Journal of Geophysical Research: Atmospheres, 108(D3).

Xu, H., Chen, Y., & Wang, L. (2018). Cross-track infrared sounder spectral gap filling toward improving intercalibration uncertainties. IEEE Transactions on Geoscience and Remote Sensing, 57(1), 509-519.

1.6. Acronyms and abbreviations

AOCS: Attitude and Orbit Control System

BOA: Bottom of Atmosphere

BRDF: Bidirectional Reflectance Distribution Function

CARD4L: CEOS Analysis Ready Data for Land. Initiative of the Committee on Earth Observation Satellites

DESIS: DLR Earth Sensing Imaging Spectrometer

DCC: Deep convictive clouds

DH: Dual Polarisation, H transmitting

DIMITRI: Database for Imaging Multi-spectral Instruments and Tools for Radiometric Intercomparison

DV: Dual Polarisation, V transmitting

EW: Extra Wide Swath (One of the operating modes of Sentinel-1 instrument)

FRP: Fire Radiative Power

FOV: Field of view

FRM: Fiducial Reference Measurement

GCP: Ground Control Points

GEO: geosynchronous satellite

GMF: Geophysical Model Function

GRI: Global Reference Image

GSLC: Geocoded Single Look Complex

HH: H transmit / H receive polarisation configuration (type of operating mode for SAR instrument)

HV: H transmit / V receive polarisation configuration (type of operating mode for SAR instrument)

IFOV: Instantaneous Field of view

ISRF: Instrument Spectral Response Function

IW: Interferometric Wide Swath (One of the operating modes of Sentinel-1 instrument)

LEO: Low-Earth Orbit Satellite

MICMICS: Mission Integrated Calibration Monitoring Inter-Calibration System

MTF: Modulation Transfer Function

MUSCLE: Multi Sensors Calibration Environment

NIR: Near Infrared

OSCAR: Optical Sensor Calibration with simulated Radiance

PICS: Pseudo Invariant Calibration Sites

POLDER: POLarization and Directionality of the Earth's Reflectances

PROBA-V: Project for On-Board Autonomy - Vegetation

R&D: Research and Development

ROSE-L: Radar Observing System for Europe-L Band. Copernicus Expansion mission / SAR mission in L Band

RTM: Radiative Transfer Model

SI: International System of Units

SPOT-1 HRV : Satellite pour l'Observation de la Terre (1) High Resolution Visible

SWIR: Short Wave Infrared

SADE: Structure d’Accueil de Données d’Etalonnage – Satellite cAlibration Data basE

TOA: Top of Atmosphere

VCC: Vicarious Cold Calibration

VNIR: Visible-Near Infrared

VH: V transmit / H receive polarisation configuration (operating mode of SAR instrument)

VV: V transmit / V receive polarisation configuration (operating mode of SAR instrument)

2. General approach

The recommendations on the gaps are ranked by their importance (need/criticality) and required efforts (time/funding) as low, medium or high (table 1 and 2).:

Table 2-1: Effort assessment for the R&D recommended activities.

Effort	Definition
Low	Corresponds to less than 2 person year activity, and not requiring any field campaigns or acquisition of data.
Medium	Corresponds to 2 to 5 persons year activity, and not requiring any field campaign. May require acquisition of limited set of specific data for the activity.
High	Corresponds to more than 5 persons year activity and requiring specific field campaigns or acquisition of large set of specific data for the activity.

Table 2-2: Criticality assessment for the R&D recommended activities.

Criticality	Definition
Low	Improvement of current Cal/Val activity Not performing the R&D activity do no prevent to realise future Cal/Val activity
Medium	Need of Cal/Val activity for specific subset of the products Not performing the R&D activity do not prevent to realise Cal/Val activity for the major components of the products.
High	Not performing the R&D activity prevents to realise the Cal/Val activity for the major components of the products.

3. Optical component

3.1. Gaps in Radiometry Cal/Val Vicarious methods

As mentioned in D2.8, even if the pre-flight characterisation of the calibration system has been performed with the best precision, it suffers from stresses (launch vibrations, ageing and contamination of the Spectralon surface related to UV intense irradiation during the in-orbit life …). As a consequence, for example, the radiance reflected by the diffuser will progressively be overestimated as the diffuser darkens, and absolute radiometric gains underestimated. While the expected uncertainty of solar diffusers is a priori compatible with targeted radiometric performances (3% typically), it was not met for SLSTR A and B and OLCI-A (see D1.5).

In this context, calibration results obtained from the on-board calibration device need to be validated and monitored by indirect methods in order to improve the reliability of the calibration procedure, assessing the multi-temporal decrease of the radiometry sensor sensitivity. It must be stressed however that indirect methods can provide a corrective term to apply to calibration coefficients, assumed valid at a given time, to correct for a time drift. **Vicarious calibration technique consists on equivalent TOA radiance/BOA reflectance computation for a known surface reflectance and a known atmosphere.**

A variety of vicarious techniques using natural Earth targets have been developed and applied successfully for the Copernicus Sentinel missions, for example, Rayleigh Scattering over Ocean, Deep convective clouds, PICS method, Sunglint Scattering, Lunar calibration, Bright and stable stars etc (see D2.8 & D2.9 for more details). The advantages and limitations of the aforementioned sources are presented in the D2.8. Here below, we recall the gaps overall the methods and the needs for R&D activities.

- Rayleigh scattering calibration methodologies utilise open ocean observations, thus it is very sensitive to clouds contamination. In addition, cloud filtering is difficult over PICS method without a dedicated cloud band. Although there is several cloud detection algorithms (e.g CMIX initiative), an accurate cloud-mask and cloud shadow are needed. **Criticality – High, effort needed – Medium**
- PICS method rely on the knowledge of the target surface characteristics (e.g. reflectivity, homogeneity, temporal variability) to simulate the TOA signal. In this context, Better characterization of the sites are needed (e.g. CEOS-PICSCAR initiative). **Criticality – Medium, effort needed – High**
- The DCC methods are used to monitor the inter band calibration over VIS and NIR bands. Since then efforts have been made to characterize and model the variability of DCC in order to monitor the trends of calibration gain over VIS and NIR bands. However, SWIR domain requires more development (Bhatt et al, 2017; Lamquin et al, 2017). **Criticality – Low, effort needed – High**
- Most vicarious CalVal methods rely on RTMs, and their contribution is important. Improvements on this aspect would be desirable. For example, better representation of

the microphysical/macrophysical state of the cloud properties in the RTMs . **Criticality – Low, effort needed – Medium**

- Most the above mentioned methods has a comprehensive uncertainty budget ,but not full-traceability to SI standard. **Criticality – Medium, effort needed – High**
- An accurate instrument characterization (ISRF, FOV variations, nonlinearity, noise, …) is needed for all vicarious methods (see D3.1 for more details). **Criticality – High, effort needed – High**
- It is commonly known that the simulation of the TOA radiance/reflectance signal relies on the atmosphere conditions and the BRDF of the target-site. Although several BRDF models exist, there is a lack in the input dataset with large range of sensor/solar geometry required for better accuracy of the BRDF measurements. **Criticality – Medium, effort needed – Medium**
- Better wind speed estimation is needed, which is often still uses Cox and Munk (1954) method, the resolution of which is much coarser than the high resolution sensors such as Sentinel-2. Improving resolution of wind speed data from the current 0.33 x 1 degree with uncertainty of ±2ms^{-1} would yield better results (Zhu et al, 2019). **Criticality – Medium, effort needed – Medium**
- Both Rayleigh Scattering and Sunglint methods uses the marine reflectance measurements in order to compute the TOA signal over ocean. Hence, an accurate estimation for the surface marine reflectance would significantly contribute to the TOA-signal estimation. Thus, the improvement of marine reflectance retrieval is desired. **Criticality – Medium, effort needed – High**
- Regarding the Lunar calibration and Bright Stars methods, both techniques require that the instrument must view the Moon/Star, involving a minimum level of agility for the platform such as a manoeuvre of the satellite to observe moon/star (See D3.1 for more details) . **Criticality – Medium, effort needed – High**
- Better characterization of the moon albedo is desirable, for example models intercomparison exercise. **Criticality – Medium, effort needed – Medium**
- The calibration of pixel relative response (i.e. equalisation) and the validation of spatial and temporal radiometric noises are performed on uniform scenes, both at low and high radiances. Uniform scenes are widely used for instruments which are not equipped with on-board calibration devices (e.g. SPOT-5 Pascal et al. 2003). However, the accuracy of this method rely on the accuracy of the model for the detector’s response (linearity). Surely the improvement of the model’s accuracy would be of interest for Copernicus missions . **Criticality – Medium, effort needed – Medium**

3.1.1. Recommendations

This section provides a list of recommendations for R&D on Cal/Val method based on the gaps identified above, with the associated effort and criticality.

- Sensors and system performance may be quantified in terms of the accuracy of measurements for advanced sensor characterization, enabling a pre-flight validation of the calibration. In this context we recommend more efforts in R&D on the pre-flight/post-fligh instrument characterization (ISRF, FOV variations, nonlinearity, noise, …). **Criticality – High, effort needed – High**
- In fact, cloud masking in remote sensing prepares imagery for processing and improves product generation. Most CalVal methodologies are sensitive to clouds contamination. Already efforts have been done on the comparison of the current cloud masking (e.g CMIX initiative), nonetheless, an accurate cloud-mask and cloud shadow are recommended for R&D in the remote sensing imagery. **Criticality – High, effort needed – Medium**
- Most Copernicus-missions requirement on radiometric accuracy are better than 3% and the up to better than 2% for future one. In order to enable the assessment of the sensors performance, a full uncertainty budget and full-traceability to SI standard of the vicarious CalVal are required. **Criticality – Medium, effort needed – High**

3.2. Gaps in spectral characterisation methods

Even though the topic of an appropriate in-orbit spectral characterization and calibration is sometimes neglected, it is of increasing importance already for broad-band multi-spectral missions. For example, the relatively broad spectral bands of Sentinel 2 MSI and the Landsat series need to be known when both products are combined, as done within the ESA Sen2Like and the NASA HLS (Harmonized Landsat Sentinel2) products. Currently empirical spectral bandpass adjustment factors (SBAFs) are used due to the lack of in-orbit spectral characterizations.

In the following, the different approaches are outlined, and evaluated regarding gaps; note that a closer description including the limitations can be found within D2.1 and D2.2.

3.2.1. Vicarious assessment based on external absorption features

In addition to the spectral characterization approaches based on laboratory and on-board calibration sources, also vicarious approaches for fine spectral resolution sensors exist.

One approach useful for fine spectral resolution sensors is based on Fraunhofer lines. It is used for Sentinel 3 OLCI, and is briefly described in D2.2 (Ch. 2.1.8). This approach has an absolute accuracy of ~0.2 nm with a relative uncertainty below 0.05 nm. The limitation of this approach is that Fraunhofer lines only occur up to ~898 nm, but is adequate for the given range.

Another approach is based on additional atmospheric absorption features, with the Oxygen A band at ~760 nm being the most prominent example, but useful absorption features are available in the full VIS-NIR-SWIR region. This approach allows for the spectral characterization of hyperspectral sensors having a spectral bandwidth smaller than ~15 nm (see Guanter et al., 2006). For airborne hyperspectral instruments, the uncertainty related to this approach was estimated in EUFAR (2009) and is within the range of 0.25 nm – 0.50 nm. Based on recent studies using the spaceborne DESIS spectrometer having an average spectral resolution of 2.5 nm (FWHM), the uncertainty is likely lower as spectral smile residuals of ~0.2 nm could be reliably estimated (Bachmann et al., 2021).

Both approaches are well understood, and have an additional limitation that in a strict sense only bands around the spectral features can be characterized, and all other wavelength regions must be interpolated. The gap is again that broad-band multi-spectral instruments can't adequately resolve these narrow absorption features, so no vicarious approach is available. Thus, currently it is important to ensure the spectral stability through pre-flight characterization and laboratory tests.

3.2.2. Summary and recommendations on gaps regarding spectral calibration

- Regarding the laboratory characterization and calibration of spectral sensor properties, no gaps were identified (see also CCVS-TAR-D3.1). For on-board calibration facilities, adequate means do exist for narrow-band sensors (e.g., spectral diffuser for OLCI), and the limitations of the various approaches (e.g., stability of LED arrays) are known. One minor gap is that currently used tunable laser diodes can not cover the short wavelength range (UV, blue). But for **broad-band sensors** (Sentinel 2 MSI, Landsat series), there is a **gap for on-board spectral calibration means**. For example, both Sentinel 2 MSI sensors were characterized on ground, and the assumption was made that the spectral response is still valid in orbit. This gap is less important -but still relevant- in case of the mentioned stable filter instruments (e.g., stripe filers of MSI, interference filters of Landsat). When having more robust and accurate in-orbit knowledge of the spectral bandpass, then the empirical SBAF determination can be improved, resulting in better harmonized multi-sensor products. **Criticality – medium, effort needed – medium**
- For the vicarious spectral calibration, the *limitation* identified within D2.2. is that it is only applicable for instruments having a spectral resolution and sampling better than ~2 nm (Fraunhofer lines) or better than ~15 nm (atmospheric absorption features, see references above). **This underpins the *gap* of having adequate in-orbit means for spectral calibration for broad-band sensors. Criticality – medium, effort needed – high**

3.3. Gaps in geometry CalVal methods

A variety of vicarious techniques using natural Earth targets have been developed and applied successfully for the geolocation CalVal. The following sources are commonly used for optical missions.

3.3.1. Geolocation

As seen in [ref doc D2.2], the calibration of the line-of-sight bias can either be performed offline or online using a database of Ground Control Points (GCPs). This data base needs to be global for online use. We then talk about a 'Global Reference Image' (GRI) even though, it might not be a single gap-free image. The GCPs planimetric and altimetric quality needs to be well mastered and consistent with the required geolocation performance of the sensor to be calibrated. This means that the errors in position and elevation of these GCPs have to be negligible with regards to the required geolocation performances.

The calibration method is commonly based on correlation between a reference image and the image to calibrate. Correlation algorithms have now reached very high performances (0.05 pixel uncertainty) and we do not see particular gaps there. They are also able to manage images with different spectral content. However, the calibration performance strongly depends on the quality of the input reference image and its associated DSM (Digital Surface Model). The spatial resolution of the reference image is generally better or equal to the resolution of the sensor to calibrate and may need to be resampled in the geometry of the sensor to calibrate. The resampling quality is very sensitive to the interpolation method but the latter have reached high performances today and once again we do not see particular gaps here. However, the workflow must ensure, that repeated resampling is avoided.

The gaps to work on is related to the reference image so as to build:

- An accurate reference image having a geolocation performance at least twice better than the sensors to calibrate
- A reference image to be global answering different needs:
    - for online calibration as the increased spatial resolution of sensors makes the line-of-sight motion more sensitive to high frequency perturbations
    - to ensure the multi-temporal registration of the images acquired by the same or different sensors
- A reference image depending on time so as to manage landscape evolution or seasonal variations
- An associated accurate and co-registered DSM so as to manage large viewing angles combined with high altitude targets
- A public reference image and DSM to foster the interoperability of space systems

SENTINEL-2 GRI (Global Reference Image) is an example of what has to be done and improved. This GRI is made of L1B products which geometric model has been refined through a global spatio-triangulation using reference data very well geolocated such as PLEIADES or IGNE GCPs data base (DB-Ortho). The L1B product which is expressed in the focal plane geometry was suited to SENTINEL-2 needs but is not usable for other missions, including COPERNICUS ones without further processing. The next steps could be in terms of priority:

- Extraction of a dense set of GCPs from the S2-GRI and projection into a common ground-based reference system (e.g. EPSG:326xx) using a very accurate DSM documenting uncertainty in x, y, and z. This activity is planned by ESA and should be encouraged. **Criticality – medium, effort needed – high**
- Production of a projected and gap-free global reference (projection to be defined) mosaic (GRM) (based on S2) co-registered with a DSM (Copernicus DEM) and

documented per-pixel x, y, and z uncertainty estimates. **Criticality – medium, effort needed – high**

- To assess and manage time evolution. **Criticality – medium, effort needed – medium**
- To continue to improve the geolocation of GRI and GRM **Criticality – medium, effort needed – high**
- To make it public so that satellite operators can anticipate its use in their ground segment. **Criticality – medium, effort needed – medium**
- The 30 m COPERNICUS DEM is available today but there will be a need for higher resolution. **Criticality – medium, effort needed – high**
- The geometric refinement methods (on-line calibration) could be made more robust to cope with difficult cases (isolated islands, cloudy scenes). **Criticality – medium, effort needed – high**

3.3.2. Focal Plane Calibration

The focal plane calibration is generally performed once in the satellite lifetime as the focal plane is assumed to be rigid. As described in [RD2.8] there are currently two technics:

- The most frequent is based on a very accurate (planimetry and altimetry) and dense reference covering the complete field of view in the ideal case but the calibration can also be managed by piece of the focal plane. The method is based on the correlation of images and the previous comments (see section 2.4.1) are still valid.
- By steering the line of sight so as to acquire the same target in a cross mode (yaw steering). It requires satellite agility but relaxes the needs in terms of targets which only have to provide good correlation performances.

We have two recommendations here:

- To foster the interoperability of systems, it is important that the viewing directions of the different operational sensors (COPERNICUS and beyond) are well calibrated; so sharing some public dense and accurate references data between satellite operators would benefit this objective. **Criticality – medium, effort needed – low**
- To recommend space agencies and satellite manufacturer to put requirements on the agility of the platform as it would benefit several geometric or radiometric calibration needs. **Criticality – low, effort needed – high**

3.3.3. Multi-spectral and Multi-temporal layering

The quality of the multi-spectral and multi-temporal registration depends:

- on the capacity to accurately manage through the geometric model some variations of the line of sight motion which are not provided by the AOCS (Attitude and Orbit Control System)

- on the performance and flexibility of the correlation algorithm which has to work with landscapes varying spectrally or temporarily
- on the quality of the DSM which affects the comparison of images acquired with different viewing angles

We would then recommend:

- to put requirements, at satellite level, on the perturbations of the line-of-sight motion so that the residue can be easily modelled (linear with time). **Criticality – medium, effort needed – medium**
- to build and make available an accurate global DSM, consistent with the highest Copernicus spatial resolution. **Criticality – medium, effort needed – medium**
- When a GRI is used to foster the multi-spectral or multi-temporal registration, to manage the time evolution of this GRI to deal with landscape evolution or seasonal variations. **Criticality – medium, effort needed – medium**

3.3.4. Line of sight stability

We have seen in task 2.2 [RD2.8] that we had 4 possible methods to assess the line-of-sight stability:

- By correlation of images acquired in different spectral bands when there is a time delay between these acquisitions
- By steering the line of sight so as to fix a ground target
- Line by line correlation
- By steering the line of sight so as to fix a star

For the two first methods, an accurate DSM is required to manage the different viewing angles between successive images. The last one does not need any DSM but requires an agile satellite with the capacity to point stars.

We have seen that correlation algorithms have now reach good performances. These performances are emphasized by the ability of the landscape to well correlate.

We would then recommend here:

- to identify the best correlation landscapes at a global scale. **Criticality – low, effort needed – low**
- to put requirements on the agility of the platform. **Criticality – medium, effort needed – low**

3.3.5. Summary and recommendations on gaps on geometry CalVal methods

From the gaps identified above, we would recommend the next steps to be considered for R&D in terms of priority up on its High-medium criticality:

- We would like to emphasize on ESA's planning of the extraction of a dense set of GCPs from the S2-GRI and projection into a common ground-based reference system (e.g. EPSG:326xx) using a very accurate DSM documenting uncertainty in x, y, and z. In addition, the improvement of the geolocation of GRI and GRM and the assessment/management of the temporal evolution are required. **Criticality – medium, effort needed – high**
- The improvement of the geolocation calibration of satellite instruments to meet the challenges of high accuracy measurements requires high resolution DEMs better than the 30m COPERNICUS DEM available today. **Criticality – medium, effort needed – high**
- To foster the interoperability of systems, it is important that the viewing directions of the different operational sensors (COPERNICUS and beyond) are well calibrated; so sharing some public dense and accurate references data between satellite operators would benefit this objective. **Criticality – medium, effort needed – low**
- To recommend space agencies and satellite manufacturer to put requirements on the agility of the platform as it would benefit several geometric or radiometric calibration needs. **Criticality – medium, effort needed – low**

3.4. Gaps in MTF Cal/Val methods

We recall that the MTF (Modulation Transfer Function) expresses the spatial frequency response of the imaging system. It gives the contrast as a function of the spatial frequency. The MTF is the Fourier Transform of the Point Spread Function (PSF), which is defined as the image of a point source of light.

We have seen in task 2.2 [ref Doc D2.2] that different techniques can be used to calibrate or validate the system MTF based on:

- Bright stars
- Edges and Bridges
- A least square method consisting in comparing images acquired over the same target by two sensors with different spatial resolution

The star-based method depends on the IFOV of the instrument to calibrate. It is suited to high resolution sensors which are more sensitive to focus change and require the MTF knowledge when deconvolution is part of the level 1 processing. This method requires several stars observations when no hypothesis is made on the MTF but using a physical MTF model strongly reduces the number of observations. The main constraint of this method is the platform agility.

The step edge method based on natural targets depends on the quality of the edge which is assumed to be perfect in the calibration method. It also depends on the orientation of the edge with regards to the satellite along and across track directions.

The step edge method can also take advantage of moon edges.

The least square method is a relative method which requires the knowledge of the reference sensor MTF or requires that the MTF of the reference sensor is perfect (close to one) at the Nyquist frequency of the sensor to calibrate. Moreover it requires quasi-simultaneous observations in close spectral bands which may be seen as a limitation.

3.4.1. Recommendations on gaps in MTF CalVal methods

From the gaps identified above and in RD2.8, we would recommend the following points to be considered for R&D:

- To develop the star-based method which is new and promising for high resolution sensors. This recommendation also calls for an agile platform and/or the capacity to view stars. **Criticality – medium, effort needed – medium** (in terms of method)
- To develop the moon-based method which is new and promising but also calling for an agile platform and/or the capacity to view stars. **Criticality – medium, effort needed – medium** (in terms of method)
- To develop artificial targets within super-sites: step edges or point sources like projectors or mirrors. The step edges targets are preferable as they have less constraints than point sources which need to be placed on a very uniform background so as to oversample the point spread function of the sensor to calibrate. Point sources also have to be oriented towards the sensor. **Criticality – medium, effort needed – high** (in terms of implementation)

3.5. Gaps in Inter-satellite comparison methods

Inter-satellite comparisons for optical missions often rely on simultaneous nadir observations data (SNO) or near concomitant observations (NCO). Although it is a common practice to compare measurements obtained by different satellites, the comparisons are limited by:

- Temporal variations of the geophysical signal;
- Variations of the observation geometry;
- And variations of atmospheric conditions.

3.5.1. Intercalibration methods

In Task2.3 [D2.9] both methods - Angular Match-ups (so-called Near coincident Observation) and the simultaneous nadir observations data (SNO)- have been described

The gaps to work on are:

- Stable and well characterized targets such as homogeneity, temporal stability etc. **Criticality – medium, effort needed – medium**
- Multiple sources of uncertainty to be investigated such as adjacent effects, spectral response etc. **Criticality – medium, effort needed – high**

3.5.2. Tandem analysis

At the beginning of its mission, the Sentinel-3B satellite has been placed on a tandem formation at a 30 s interval with Sentinel-3A. Data acquired during the tandem phase (which lasted 4 months) has proved highly valuable for optical instruments (see e.g. Clerc et al., 2020).

The instruments on board of the two satellites are observing the same scenes under nearly exactly identical conditions, which allows to assess non-systematic uncertainties in a detailed way. As documented in Lamquin et al. 2020a, geometric co-registration, conversion to reflectances and spectral adjustments need to be performed before radiometric comparisons can be made. This is true for all inter-satellite comparison activities but in the context of the tandem these error sources are dominant over other terms (differences in observation and atmospheric conditions). One of the strong advantages of the tandem configuration is that it allows a direct validation of prognostics uncertainties (see e.g. Hunt et al. 2020).

The Sentinel-3 tandem study has developed relevant methodologies to perform inter-comparisons of optical missions which can be applied to future tandem phases (Sentinel-2, CHIME, LSTM). The future FLEX tandem with Sentinel-3 could also provide useful intercomparisons opportunities. Although there is no major gap in terms of methodology for these intercomparisons, some aspects could be further improved:

- Geometric uncertainty assessment through co-registration analysis could be further developed, especially for Sentinel-2. Dense inter-comparisons could provide interesting insights in short-term geometric errors and assess the relative contributions of deterministic thermo-elastic effects and other random error terms.
- The drift phase at the end of the tandem provides opportunities for quasi-simultaneous measurements from slightly different viewing directions (a few degrees). Data acquired in these conditions is therefore very useful to assess the impact of viewing conditions on L2 retrieval algorithms (atmospheric and BRDF effects). This aspect has been addressed in a preliminary way in Lamquin et al. 2022c but could be further developed. It could also provide a dataset to validate BRDF kernels, or to derive digital surface models.

The following recommendations are provided:

- Tandem phase opportunities should be implemented whenever practical (during commissioning phase or at end of mission). Similarly, dedicated cal/val activities should be planned to benefit from tandem configurations with a Copernicus satellite (e.g. FLEX with Sentinel-3). **Criticality – high, effort needed – high**.
- Methodologies to analyze short term geometric effects for Sentinel-2 should be developed. **Criticality – low, effort needed – low**.

- Methodologies to analyze the impact of viewing conditions on geophysical retrievals should be developed. **Criticality – medium, effort needed – low**.

3.5.3. Inter-comparison with models

As mentioned in [ref 2.3] comparisons against radiative transfer model simulations request the evaluation of the coupled surface-atmosphere RTM used to assess their own uncertainties. There is also a gap of actual description of state variables of the surfaces that represent the ensemble of pixels. Examples of Level 1 comparisons are provided in [ref D3.5]. If we want to compare the surface reflectances or biophysical products with RT simulations or state variables, they are often a gap of the knowledge of spectral characterization of scattering elements as well as the structure of the surface. In term of methodologies per se, uncertainties traceability (i.e. from state variables and RT models) should be further develop.

Recommendations:

- Develop 3D Radiative Transfer Models for reference validation sites, as well as tools to facilitate such efforts (e.g. tree identification tools, lidar data processing) and spectral characterization data. **Criticality – medium, effort needed – high**.
- Develop uncertainties traceability for RTM models. **Criticality – medium, effort needed – high**

3.5.4. Inter-comparison of products

Uncertainty is often not considered in products inter-comparison, such as in burned area (BA) products and Landcover products. Classification maps result from Machine Learning based algorithms that take into account temporal spectral changes of the surface. Currently, L2 reflectance uncertainties are not taken into account in the process, but algorithm developers provide information on the degree of confidence of the classification of each pixel. This confidence can incorporate many aspects (ML based probability, cloudiness prior to detected date, and rate of spectral change and consistency). However, these are product/algorithm dependent.

As such, recommendations could be:

- Definition of a standard method for estimating thematic classification (BA or LC) confidence, to allow for full product compatibility and comparability. **Criticality – high, effort needed – medium**
- Development of method/models to allow for conversion/transform between classification confidence and uncertainty. **Criticality – high, effort needed – medium**

For Fire Radiative Power (FRP) products, as fire can be an ephemeral and rapid changing event, the lack of simultaneous observations between different satellite-based sensors is a serious limitation to allow for product comparison. In addition, small differences in angular effects, spatial resolution, and atmospheric contamination (cloud and smoke) highly impact the fire related radiance reaching the sensor. Current inter-comparison exercises are only made at the level of fire clusters using the classical linear approach. Furthermore, only a small number of satellite-based FRP products (MSG and S3) provide estimates on the uncertainty that is based on modelling exercises - no per-pixel radiance uncertainty is taken into account. This significantly limits the ability to evaluate any comparison between products. Finally, the developed higher-level products (L3/L4) FRP products results in products that are, in terms of completeness, spatially and temporally non-consistent, because each product is a snapshot of a fire event at a particular time.

As such, the recommendations are:

- Development of robust statistical methods for non-simultaneous products that allow for the identification of the limitations and the impact of each product assumptions and uncertainties. The use of probability density functions to define each product baseline “signature” shows promising results (Mota et al., 2020). The same can be transferred to other highly variable product. **Criticality – high, effort needed – medium**
- Assimilation of per-pixel radiance uncertainty (at L1) in order to propagate the uncertainty to the final product. As a minimum, assess the error of using model-based uncertainties. **Criticality – high, effort needed – high**

Definition and guidelines on spatial and temporal consistent higher level FRP products (L3/L4). **Criticality – high, effort needed – medium**

More generally, for other biophysical products like FAPAR, LAI and Albedo, no name a few, inter-comparisons are often done without considering the various definitions and assumptions made in retrieval algorithms. The main gaps are the assessment of potential discrepancies linked to the difference in viewing geometries and the original spatial resolution as optical sensors do not have the same technical characteristics. To benchmark land products, remapping them over a geographical grid is mandatory and should include some geo-spatial uncertainties, especially for vegetated areas that are not homogenous (Loew et al., 2017, Gobron et al., 2022).

The recommendation is:

- Development of a community accepted standard for geo-spatial uncertainties for regrided/reprojected products. **Criticality – medium, effort needed – medium**.
- Development of product uncertainties considering the assumptions made in retrieval algorithms and uncertainty propagation from input products. **Criticality – medium, effort needed – medium**.

3.5.5. Recommendations on gaps in the Inter-satellite comparisons CalVal methods and products

From the gaps identified above and in RD2.9, we would recommend the following aspects for improvement and R&D:

- Tandem phase opportunities should be implemented whenever possible with a Copernicus satellites (e.g. FLEX with Sentinel-3). **Criticality – high, effort needed – high**.
- For the intersatellite inter-comparison, further investigations are required over multiple sources of uncertainty. **Criticality – medium, effort needed – high**
- RTMs are used for CalVal activities, as well as tools to facilitate data inter-comparison, in this context, development of 3D RTMs and uncertainties traceability are needed. **Criticality – medium, effort needed – high**.

Regarding the products inter-comparison, the recommendations are:

- Development of robust statistical methods for non-simultaneous products. **Criticality – high, effort needed – medium**
- Assimilation of per-pixel radiance uncertainty (at L1) in order to propagate the uncertainty to the final product. **Criticality – high, effort needed – high**
- Development of product uncertainties considering the assumptions made in retrieval algorithms and uncertainty propagation from input products. **Criticality – medium, effort needed – medium**.

3.6. Gaps in ground-based Cal/Val methods

Gaps in ground-based or airborne measurements available for validation of products provided from satellite data are summarized in “D3.8 - Copernicus operational FRM network and supersites”. This document concentrates on considering gaps in methodologies for using these ground-based data for validation.

We start with identification of gaps in surface reflectance validation as surface reflectance (SR) is a key parameter for most derived parameters characterizing land and water products. Then we come to gaps in derived parameters following the split in land- and water products equivalent to D3.8. Last point addressed is cloud masking.

3.6.1. Surface reflectance

As a preliminary remark, we note that surface reflectance (SR) refers to a “family” of reflectance quantities. As a first recommendation, there is a need to have clear and commonly agreed definitions on the different type of reflectance products, which should be used by product providers in a consistent way (see Nicodemus et al., 1977 and Schaepman-Strub et al., 2006).

Validation of spectral surface directional reflectance (or spectral and broadband albedo) is currently limited. This is mainly due to a lack of reference measurements on ground, a point which is addressed in D3.3 Copernicus measurement networks. However, new validation respectively quality assessment methodologies should be developed in parallel.

One common approach for quality assessment of surface reflectance is the use of computed reference data (Doxani et. al., 2018). Here SR reference is computed from top-of-atmosphere data using a radiation transport model with atmospheric parameters provided by AERONET as input. This has the advantage that SR reference data are provided for a huge number of pixels allowing representative statistical analysis. Also reference data can be obtained for all surface types. The drawback is that this method includes the use of a radiative transfer model which introduces an additional uncertainty in the computed reference data. Estimation of this uncertainty is still a gap in application of that method. Therefore, in the terminology should be strictly differentiated between quality assessment and (true) validation which must rely on real measurements.

Following aspects are to be considered for validation of SR based on ground-based or airborne measurements:

- Spectral adjustment between bands of satellite instruments and reference measurements. This can be easily done in case of reference measurements with a hyperspectral instrument by convolution of the sensor spectral response function with the measurements. Lowest uncertainty will be obtained if the sensor spectral response is known per pixel which is not the case for current satellite sensors. Empirical spectral bandpass adjustment factors (SBAFs) are used in case of multispectral reference measurements.
- Spatial adjustment (upscaling) from area covered by the ground-based measurement to the pixel size of sensors onboard of satellites. Recommendations and methodologies for upscaling pointwise surface reflectance measurements to satellite pixel resolution should be clarified and standardized. Best practices documents are available for arranging ground-based measurements to allow representative upscaling. (see Malthus et.al,. 2018, Wang et.a. 2019, SVC 2019).
- The higher the spatial resolution of sensors in orbit is, the more important becomes taking into account the point-spread function of the instrument. This influence can be minimized for validation by averaging over several pixels. However, for heterogeneous surfaces and if some products have to rely on single pixels only, then a convolution of point-spread function with signals from surrounding pixels may become important. This requires knowledge of the point-spread function which is currently not provided.

UAVs are a good option for campaign-based SR estimates (but still relatively new) and don't solve all the issues. Establishment of good practices for UAV measurements for Cal/Val is starting with SRIX4VEG (Surface Reflectance Intercomparison Exercise for Vegetation) and should be continued. Some aspects regarding measurement towers are discussed in D3.1 for instrumentation technologies.

Another aspect concerns the evaluation of BRDF effects, which is a critical aspect when comparing measurements performed with different viewing conditions as is generally the case for ground-based validations.

Field BRDF measurements could help to improve this difficulty. However, they are limited by:

- Impact of changes in illumination conditions during the multi-angular measurements (for small areas with goniometers)
- Impact of atmospheric conditions

The difficulty is higher for vegetated surfaces (especially for high canopy) and complex waters than for bare soil or clear water surfaces. In order to progress toward a better inter-operability and an open-data validation approach, there is a need to apply common protocols and approaches to model surface BRDF, ideally with a community processor.

We need a lot more focus on the importance of the calibration/characterization of instruments/artefacts (e.g. reference (white) panels). Even when using Atmospheric correction to convert top-of-atmosphere to bottom-of-atmosphere data is a very complex process far from being a standardized procedure. There are several special corrections included in some atmospheric correction processors which have to be validated but don't have a validation methodology so far. Most important is to develop a validation procedure for topographic corrections of SR retrieval in satellite data. Current validation methods are applicable only for flat terrain.

The recommendations are:

- Estimation of uncertainty in computed reference data for SR validation coming from radiative transfer model. **Criticality – high, effort needed – medium**
- Development of a community processor providing reference data from ground measurements. This processor should consider the calibration procedure, modelling of surface BRDF from field measurements, necessary QA-checks and yield measurement uncertainties in a clearly defined procedure. This includes permanent updating of best practices documents with necessary measurement protocols. **Criticality – medium, effort needed – high**
- Establishing best practices for applying UAV measurements for Cal/Val. **Criticality – medium, effort needed – medium**
- Developments on new validation methods for special atmospheric correction steps like terrain correction. **Criticality – high, effort needed – medium**

3.6.2. Water products

Commonly accepted guidelines for reference measurements over ocean and coastal area are well developed. They are also widely applicable to inland waters. One aspect more to consider is adjacency effect which is more important over water than over land surfaces, most of all for inland water and coastal waters. There is still no method for validation of adjacency correction. The use of 3D RTM simulation to support validation activities has already been mentioned in section 2.5.3. and may be also useable for generation of simulated validation data sets. Another option could be special experiments setting up reference measurements on and around lakes. An example of such setups was already realized in Belgium.

Ocean Colour processing relies critically on vicarious radiometric adjustment using the so-called System Vicarious Calibration (SVC) approach. Analysis of L2 ocean color products using Sentinel-3 tandem data showed that the SVC is not able to ensure a perfect harmonization of the products from the Sentinel-3A and B units (see Lamquin et al., 2020b). On the other hand, a preliminary harmonization of the satellites opens the way to a multi-sensor system vicarious calibration approach which could improve the overall performance.

The recommendation is:

- Development on a method for adjacency correction. **Criticality – high, effort needed – medium**

3.6.3. Land Products

Optical land products can be validated using in-situ reference measurements, but a methodology needs to be developed to assess the impact of non-homogeneity at the spatial scale of a satellite pixel. Several up-scaling methodologies have been developed, with two main approaches:

- Transfer function based empirical mathematical approaches
- Transfer function based on high resolution images processed with Radiative Transfer Models

The latter approach has gained more attention recently in the context of new DHP-based systematic measurement networks. Efforts to consolidate the methodologies and the associated uncertainty assessment should be continued, especially in view of higher resolution missions like Sentinel-2. Effort: low, Priority: low

The use of 3D RTM simulation to support validation activities has already been mentioned in section 2.5.3. Vegetation products (LAI and FAPAR) would probably benefit a lot from comparisons with simulation-based data. In particular, the sensibility to observation conditions (BRDF effects) could be assessed and taken into account e.g. when assessing comparisons with in-situ measurements. The model used for validation should be different (and ideally more accurate) than the one used (if any) for generating the retrieval algorithm. The development specific validation methodology for vegetation products should be performed as part of the R&D activity recommended in 2.5.3 (**effort: high, priority: medium**).

3.6.4. Fire products

There are currently no guidelines for ground-based cal/val methods for fire products (BA and FRP). In addition, no FRM standards and framework exists.

For FRP, the main challenge is the generation of reference data due to the ephemerality of the phenomenon to be mapped and the current revisiting times of moderate-high spatial resolution observing systems. Reference data samples are scarce and not coincident with product estimates, and field experiments are limited to controlled environments, meaning that they focus on low intensity fires consuming low amount of dry matter during the fire off-season. This leads to the issues of representativeness, in terms of landcover, intensity and scale. Very few opportunistic experiments of coincident measurements using airborne thermal imaging systems (High altitude UAV or airplanes) exist, as is the case of the FIDEX campaign. For FRP products validation efforts are mainly dependent on product intercomparison exercises (section 2.5.4).

Non simultaneous EO-based observations and the lack of uncertainty estimations are the main limitations in the validation of FRP products. This imply that FRP products will never reach CEOS LPV level-4 validation status. Some thinking needs to go into addressing how to adapt fire to this framework.

For Burned area products, Cal/Val methods mainly depend on finer resolution imagery that are classified in a semi-supervised way to produce burned area maps over sampled areas. Methods and frameworks have been developed to select these areas randomly but ensuring landcover representativeness and addressing temporal consistency (Padilha et al 2015, 2017, Boschetti et al, 2016). The generation of reference BA datasets is a major bottleneck as it requires large resources, but this is starting to be addressed (Franquesa et al., 2020).

As such, the recommendations are:

- Develop a framework to allow for the retrieval and generation of FRM fire data by establishing the required protocols to ensure that data is fully traceable in-situ measurements – a good practice guide. **Criticality – high, effort needed – high**
- Define a community-based roadmap and all the requirements in order for FRP products to achieve CEOS Level-4 validation status. **Criticality – medium, effort needed – medium**
- Define a community agreed validation strategy for classification-based mapping that ensures representativeness, in terms of landscape, in time and space. **Criticality – medium, effort needed – medium**

A FRM style project (FRM4FIRE) should be promoted to address these issues

3.6.5. Cloud masking

The validation of cloud masking is a difficult but crucial aspect, as it impacts drastically the uncertainty of optical data products.

The current validation methodologies rely essentially on human image interpretation. This activity is therefore highly time consuming and not performed on a regular basis. There are very few open-source validation datasets. There is no consensus on the methodology to build the validation datasets, in particular criteria to determine if a pixel is cloudy. Things becoming still more difficult considering transparent clouds and including shadows. As human operators are not perfect references, the operator uncertainty needs to be assessed and taken into account in the validation process.

A new promising development is the use of Hemispherical camera for cloud cover assessment which had already been discussed in D3.1 - Recommendations for R&D on Instrumentation Technologies.

Current cloud masks are binary masks (cloud or no_cloud) which are supplemented with cloud probability masks by some processors. Transforming to cloudless confidence allows to include the cloud masking information in the per-pixel uncertainty estimation of satellite products. Therefore, provision of cloudless confidence should be stimulated. However, validation of cloudless confidence seems to be impossible and is a limitation.The following points could be recommended:

- Provide (an objective) definition of “cloud”, which (ideally) includes a numerical metric including definitions for transparent clouds and cloud boarder. **Criticality – high, effort needed – medium**
- Develop an open-source cloud masking validation reference archive for Sentinel sensors in order to foster the development of new algorithms. **Criticality – medium, effort needed – medium**
- Continue inter-comparison exercises CMIX. **Criticality – medium, effort needed – medium**
- Work toward commonly agreed guidelines for generation of cloud masking references and validation of cloud masks. **Criticality – high, effort needed – medium**

A FRM4cloud project should be promoted to address these issues.

3.6.6. Recommendations on gaps in ground-based CalVal methods

From the gaps identified above, we would recommend the following aspects for improvement and R&D:

- Estimation of uncertainty in computed reference data for SR validation coming from radiative transfer model. **Criticality – high, effort needed – medium**
- Developments on new validation methods for special atmospheric correction steps like terrain correction. **Criticality – high, effort needed – medium**
- Development on a method for adjacency correction. **Criticality – high, effort needed – medium**
- Develop a framework to allow for the retrieval and generation of FRM fire data by establishing the required protocols to ensure that data is fully traceable in-situ measurements – a good practice guide. **Criticality – high, effort needed – high**
- Define a community-based roadmap and all the requirements in order for FRP products to achieve CEOS Level-4 validation status. **Criticality – medium, effort needed – medium**

- Define a community agreed validation strategy for classification-based mapping that ensures representativeness, in terms of landscape, in time and space. **Criticality – medium, effort needed – medium**
- Provide (an objective) definition of “cloud”, which (ideally) includes a numerical metric including definitions for transparent clouds and cloud boarder. **Criticality – high, effort needed – medium**
- Develop an open-source cloud masking validation reference archive for Sentinel sensors in order to foster the development of new algorithms. **Criticality – medium, effort needed – medium**
- Continue inter-comparison exercises CMIX. **Criticality – medium, effort needed – medium**
- Work toward commonly agreed guidelines for generation of cloud masking references and validation of cloud masks. **Criticality – high, effort needed – medium**

4. Altimetry component

4.1. Introduction

In altimetry, the following CalVal technical activities are common to all different surfaces:

- Identification and characterization of missing measurements. Beyond the basic quality control activity, this analysis goes a step further, detecting missing values within the products and providing additional information about the sensors health and their sensitivity to ground surface characteristics.
- Editing of corrupted measurements. Removing such measurements is an important activity to ensure the good quality of the final product. Relevant statistics on edited measurements (e.g. final percentage of relevant measurements) are also computed and made available to end users.
- Estimation of relative biases and drift. This activity consists in comparing the different variables measured by a given mission against the ones from a reference mission already validated. Through such comparison, geographical biases and temporal drifts can be characterized and monitored.
- Estimation of absolute biases and drift. This activity aims at completing the comparisons performed with respect to the reference altimetry mission by using an absolute ground reference provided by in-situ measurements, instead.
- Definition of the mission error budget. This activity relies on the two previous estimations (i.e. relative and absolute biases and drift) and aims at providing a description of the errors associated with the mission main variables. Some of the metrics obtained are used to validate the mission requirements.
- Assessment of new algorithm and/or standard performance to validate future ground segment evolutions or identify promising R&D outcomes.

To perform these CalVal activities, different methods have been designed and developed. Overall, they all rely on one or more of the following general approaches:

- Self-comparisons or “mono-mission” comparisons. These comparisons provide metrics that are used to assess the products performances without external references. Thus, they have the advantage that the resulting error estimations do not include additional uncertainty from external datasets. These methods can provide a good overview of the overall global quality of the mission products and can be used to assess the coherence of the observations (e.g. between ascending and descending passes). However, they cannot detect errors that correlated between ascending and descending passes.
- Comparisons with other altimetry missions or “multi-mission” comparisons. These analyses provide information about the products quality and performances with respect to other flying missions already validated. They are accurate and allow the detection of

small residual errors. However, they assume the reference mission errors to be perfectly characterized and not correlated with the signal to be validated.

- Comparisons with in-situ measurements. In-situ measurements provide a reference that is often considered as the ground truth. However, the precision of the commonly used in-situ networks is currently not good enough to measure small residual errors over limited temporal scales. As discussed more in detail in D3.3, this limitation is mainly related to the intrinsic precision of the in-situ instruments as well as their spatial and temporal sampling.
- Comparisons with models. These comparisons are currently performed only over a specific subset of altimetry variables (i.e. troposphere, ionosphere corrections and wind, wave parameters). Models provide an external pseudo-independent reference (not totally independent as they assimilate altimetry data) useful to detect potential errors at long temporal and spatial scales.
- Comparisons with alternative ground processing algorithms. The development of alternative Level-2 products derived from independent (with respect to the mission ground processing) processing prototypes allow for direct comparison and detection of small residual errors with a high precision. Knowing the main processing differences between the two algorithms can often provide important insights to better understand the detected errors.
- Comparisons with observations from other satellite-based technologies (radar from Sentinel-1, optic from Sentinel-2 & -3, radiometer from Sentinel-3, lidar from ICEsat-2, SWIM from CFOsat). These analyses provide interesting results, which are independent from the specific sensor technology, for several common variables (waves, winds and troposphere correction and surface topography to a lesser extent with ICEsat-2). However, their interpretation requires the different instrument characteristics (resolution, footprint size, radar frequency …) and uncertainty level to be taken into account.

The main limitations of these CalVal activities and approaches are discussed in the following paragraphs.

4.2. Limitation in the detection of missing measurements

As mentioned in the introduction, the characterization of missing measurements provides reliable information about the health and performances of the onboard sensor and the ground stations network. The detection method relies on time differences measured between consecutive measurements. It is precise and reliable to detect instrument anomalies and assess its acquisition performances in close loop mode.

However, when the instrument operates in open loop, this simple method cannot be applied. That’s because the onboard OLTC (Altimeter Open Loop Tracking Command) returns an instrument record whether there is a returned signal or not. Thus, time differences between consecutive measurements cannot be used to identify gaps in the measurements. Over continental surfaces, additional methods based on the analysis of sigma0 have been developed to assess the OLTC quality and detect if the correct returned signal is record.

The main limitation identified in such methods is that they do not account for the specific tuning of the onboard OLTC. The tracking command is adjusted over a set of identified and documented water bodies (defined in OLTC (altimetry-hydro.eu)). Thus, it is important to regularly verify that the return signal is correctly register over those areas where the tracking command has been optimized and assessing the overall OLTC quality outside them.

To make it easier to implement such analysis, it has been recommended to add OLTC tuning information in the Level-2 altimetry products. **Criticality – low, effort needed – low**

4.3. Limitations in editing methods

The radar waveform signal can be contaminated by various effects (Land contamination in coastal areas, rain cells, blooms …) preventing a correct retrieval of the geophysical parameters. Before computing statistics and performing comparisons, these corrupted measurements must be identified and removed. Although editing methods differ according to the surface types and the product sampling (1Hz / 20Hz), in general they all rely on simple statistical approaches:

- Combination of fixed thresholds applied on different variables
- Iterative filtering with local thresholds (i.e. *n*-sigma outliers, with *n* dynamically configurable)
- Correlation between consecutive measurements

Although, these methods have been proven to work, recent techniques based on Machine Learning (ML) approaches have provided interesting results that could contribute to the further improvement of the existing methods. Furthermore, since editing is also a common issue in different fields and instrument technologies, exchanges on the state-of-the-art approaches with other communities could also bring new elements for improvements. Overall, the identified gap for editing activities is small since performances (e.g. percentage of edited measurements, data quality after editing) over the different surfaces are already good. Nonetheless, we still recommend some R&D activities to test new or existing ML approaches as well as knowledge exchanges with other technology areas. **Criticality – low, effort needed – medium**

4.4. Comparisons with other altimetry missions

Such comparisons are regularly performed and provide excellent accuracy to detect small discrepancies between two altimetry missions. The main method used to compare two missions relies on the crossover approach. The method itself does not present any limitations, but the resulting matchups could be further exploited and analyzed to better characterize the observed discrepancies. Among the CalVal activities that could be further developed out, the following have been identified:

- Analyze the Sea Surface Height (SSH) residual variance as a function of the time lag between the two cross-over observations. Several components within the total SSH budget have specific temporal periodicity. Thus, their residual signature and relative contribution to the SSH is expected to vary with varying time-lags.

- Identify the contribution of different geophysical components to the observed SSH discrepancies at cross-overs. Such analysis would help to allocate an error to the different geophysical processes and corrections used to compute SSH and, hence, refine the SSH error budget.
- Analyze the temporal variability of the cross-over residuals on a regional basis. Such analysis would help refining the characterization of the residual errors and potentially identify their main controlling factors. **Criticality – low, effort needed – low**

4.5. Limitations of the in-situ comparison methods

Comparison of altimetry signal with ground measurements is an important step of the CalVal activities. It ensures that geophysical variables measured at several hundred kilometers from earth are reliable and consistent what we consider as the ground truth (this despite the several limitations related to in-situ measurements identified and discussed in D3.3).

Regarding the comparison methods, the main limitations identified are related to two factors:

- The In-situ measurements level of uncertainty. This level varies from one instrument to another, it is not always known and it is often greater than the altimeter error signal to be characterized.
- The selection of altimetry measurements to be compared with a given in-situ instrument. Indeed, the individual altimeter measurements are noisy (few centimeters at 1Hz sampling) and need to be averaged over several kilometers to get a precise local measure. However, depending on the specific region of study and the variables analyzed the measured geophysical signal could be characterized by large natural variability over short scales in both space and time. This variability can introduce a significant contribution to the total uncertainty estimation. Thus, a compromise between altimeter noise reduction and impact of averaging surfaces with different physical characteristics shall be found. In addition, for the case of SSHA analyses, the distance between altimetry and in-situ measurements can also introduce additional contributions associated with the spatial variability of the geophysical corrections: the ocean tide signal, the Mean Sea Surface (MSS) and some of the atmospheric corrections can vary by several centimeters over a few kilometers. Such variations should also be considered when comparing satellite to in-situ observations.

Different techniques have been developed and can be found in the literature. We recommend to identify with in-situ CalVal experts what would be the optimal methodology to compare altimetry and in-situ measurements and communicate about this recommendation. **Criticality – low, effort needed – medium**

4.6. Limitation of the model comparison methods

For several geophysical variables measured with altimetry, there are equivalent products resulting from global geophysical models (e.g. wave height, troposphere correction, wind speed …). The modeled products are useful to perform a large scale validation of the altimeter measurements. However, large scale models have too coarse resolution to accurately resolve the medium and small spatial scales. Comparisons are thus not relevant below a given spatial wavelength that varies regionally.

In Calval activities, when altimetry measurements are compared with model outputs, the difference is computed for each sample and averaged per spatial or temporal time steps. Such method has been used for many years and has proven to be effective. However, it could be further improved, for instance, by filtering the altimetry high frequency signal up to a given frequency consistent with the local effective resolution of the model used for validation.

The gap identified is not critical, we only recommend to improve the comparison approach and evaluate its impact. **Criticality – low, effort needed – medium**

4.7. Multiple colocation with altimetry, model and in-situ

To combine advantages of comparisons with models and in-situ and limit their drawbacks, triple colocation can be analyzed. Different configurations could be defined using one or two different altimeters, in-situ measurements and models. Such kind of analyses would help to discriminate the level of uncertainties coming from the different inputs.

Furthermore, as the models can also provide useful information about the spatial variability of the signal around the in-situ location, they could be used to optimize the selection of satellite and in-situ observations to be compared against each other. Currently, in-situ observations are compared against the closest point or section of altimetry track. While such an approach is valid in the open ocean where SSH gradients and variability are usually weak, it should be avoided in coastal regions characterized by highly variable smaller scale dynamics influenced by the coastal and bottom topography. Because of the sharp gradients, close points in space might be characterized by very different dynamical conditions. Thus, blindly comparing in-situ observations from one location with remote sensing observations from a close one, might induce important biases in the analysis. Spatial information from the models could be used to identify the area with similar dynamical characteristics around an in-situ observation, and satellite observations collected within such area should be then used for a more robust comparison against the in-situ reference.

These synergistic comparisons between satellites, in-situ and models are not usually performed in the frame of CalVal activities. We identify that they may bring additional novel elements for the characterization of altimetry products quality and deserve to be further exploited. **Criticality – low, effort needed – medium**

4.8. Intercomparison with alternative processing chains

For Copernicus altimetry missions, the development of ground segments and particularly their specification and validation are supported by the use of processing prototypes. The Sentinel-3 PDGS and Sentinel-6 PDAP ground segments have been respectively developed using the S3PAD and the GPP prototypes as reference. These prototypes are particularly useful during missions' commissioning phases to investigate potential ground segment anomalies as well as to test and validate new algorithm approaches for data quality improvements. However, from a CalVal point of view, as the two chains are very similar in terms of algorithmic choices, the comparison of their products cannot be effectively used to investigate residual errors. Nonetheless, the importance of having independent processing chains designed with different approaches has been further proved within the framework of the Sentinel 3 CalVal activities. The CNES Sentinel-3 Prototype Processing (S3PP), thanks to its numerical retracking approach, allowed the identification of the source of residual small errors detected by the Cal/Val experts. Without this alternative processing chain, the understanding of the SARM range 1 mm/yr drift error would have taken much more time to investigate. Therefore, it is important to identify this specific Cal/Val activity and, maintain the most relevant alternative processing. It provides recommendations to improve the existing ground segments as well as element of comparison to better understand the residual errors observed. **Criticality – low, effort needed – low**

4.9. Uncertainties evaluation

The overall error associated with each altimetry measurement is a result of the contribution of several components. These components are introduced at different stages along the altimetry processing chain and can be broadly grouped into three different categories based on their source as well as spatio-temporal scales:

- **Instrumental errors** are intrinsic to the instrument. They contribute to the white noise component of the altimetry error and thus are uncorrelated in space and time (high-frequency errors).
- **Correction errors** are associated with the geophysical corrections (atmospheric and from the ocean surface) applied to convert the recorded altimetry range into sea level elevation and include uncorrelated (i.e. white noise/ high-frequency errors) as well as correlated errors (low-frequency errors). The correlated errors are due to the insufficient resolution and accuracy at which the geophysical processes determining the spatio-temporal variability of a given correction field can be represented/observed. Because of that, correlated errors can span a broad range of scales and their spatial and temporal scales are directly correlated: errors at higher-frequencies (smaller temporal-scales) are correlated over shorter wavelengths (due to fast but localized geophysical processes); those at lower frequencies are correlated over longer wavelengths (due to slower but broader geophysical processes). For example, over the oceans these errors can typically span spatio-temporal scales from O(10) km - O(1) week to O(1000) km - O(1) year.

- **Retracking errors** are associated with way the various geophysical parameters (i.e. range, sigma0 and significant wave height) are derived from each returned altimetry echo. While retracking accuracy is tightly related to the algorithm (and the associated assumptions) applied to fit the model waveform to the observations, retracking performances can vary both spatially (for instance depending on the satellite geometry along each orbit) as well as temporally (for instance based on the ageing of the STM components). These variations lead to basin-scale low-frequency errors as well as to long-term drifts, both of which have an important direct impact for climate scale applications (e.g. global mean sea level trend).

The first type of errors are the ones we can identify and quantify with the highest accuracy. Different approaches can be adopted for their analysis. Such approaches are mostly based on the direct analysis of the altimetry record (e.g. along-track power density spectra; analysis of high frequency variance). Results from different approaches can be intercompared, further increasing the robustness of the results.

Long-term trends from retracking errors can also be accurately assessed either via cross-comparison with other altimetry reference missions (e.g. Jason for the GMSL) or via an uncertainty analysis of the trends obtained at different temporal scales.

Currently, the largest gaps are in the quantification of the errors at the scales intermediate between these two. These spatio-temporal scales are currently the area of focus of several altimetry applications (small scale ocean dynamics; hydrological seasonal cycles; flooding surveys; ice cap dynamics and melting). Thus, proper Cal/Val assessment at those scales is a much needed requirement. Unfortunately, while current analytical methods indicate that altimetry record at those scales include both signal and error, it remains extremely challenging to quantify how much of the altimeter record is error and how much is associated with the geophysical signal of interest. Although possible, it is unlikely that this gap will be filled in the near future through the development of new individual methods. A more reasonable assumption is to develop synergistic analyses that will be able to detangle the two by combining the info that can be retrieved from existing diagnostics.

A complete characterization of uncertainties for the different spatio-temporal scales and for the different surfaces (ocean, inland waters, land & sea ice) is important to provide feedbacks toward agencies and end-users. Some of these uncertainties are estimated and documented (the ones used to verify missions’ requirements), they however do not answer all the needs. This gap is identified as critical. We recommend to define an uncertainty matrix for each surfaces addressing the main geophysical variables and identifying the existing work (uncertainty characterization), its maturity level and scales for which an estimation is required. We recommend then R&D activities to fill the identified gaps. **Criticality – medium/high, effort needed – high**

4.10.Gaps in Copernicus Cal/Val processes

To answer the end-user needs, a key input to the Copernicus services is the temporal and spatial sampling of the satellite constellation. The more satellites are assimilated, the more precise the altimetry-derived products become, especially for the restitution of small and rapid topography variations or small space/time variations of the inland water levels. Therefore, it is of great interest to also assimilate non-Copernicus missions (also known as Contributing Missions).

As example, in addition to the Copernicus missions, the CMEMS sea surface height products derived from altimetry technique also integrate observations from the SARAL (CNES and ISRO), CRYOSAT-2 (ESA), CFOSAT and Haiyang-2 unit A to C (CNSA/NSOAS and CNES). In other words, there are currently more Collaborative altimeters than Sentinel satellites as an input to the Copernicus Services. With SWOT (NASA and CNES) and HY-2 unit D to H, the number of Collaborative missions will remain significant for the next 10 years.

Before combining all these different data sources, Cal/Val activities are performed to cross-validate the products and to verify that they are compatible with user requirements: to remove spurious measurements, to characterize the product performances and errors, to verify platform or instrument events do not affect the output products… These activities can be considered as a lighter version of Cal/Val activities with a specific focus on the overall consistency of CEOS' Ocean Surface Topography Virtual Constellation, and on the metrics of interest for a given Copernicus Service.

For the Copernicus missions, these Cal/Val activities are naturally performed by EUMETSAT for the ocean and by ESA for inland waters and land ice. However, for the non-Copernicus missions, this is a responsibility of the Space Agencies not involved in the Copernicus Program, leading to potential discrepancies between Cal/Val strategies, methods and cross-comparisons (e.g. biases and drifts, undocumented standard changes…). This lack of coordination on Cal/Val activities may impact the production quality of the Copernicus services.

To illustrate, it has been the case, recently, with the Chinese HY-2A and HY-2B missions, for which there is no communication process from NSOAS to Copernicus. Similar discrepancies happened in the past with ESA's Earth Explorer mission CRYOSAT-2 for which unexplained biases were observed over ocean at LRM/SAR mode transitions. As of today, to mitigate this risk, the Copernicus services have to rely on non-operational services from collaborative Agencies (e.g. CNES for HY2) without any Service Level Agreement nor long-term commitment. This essential extra-activity of homogenezing, completing the Cal/Val metrics is not funded by the Copernicus program, and there is no formal agreement for collaborative Agencies to coordinate their Cal/Val activities with the Copernicus Program in the coming years.

In the event that the external service is stopped, one of two scenarios might happen:

- The Services would have to stop using any unsecured Collaborative altimeter. The guaranteed inputs would be limited to Jason-3, Sentinel-3 and Sentinel-6 (for which Cal/Val activities are supported by the Copernicus program), thus degrading significantly the spatial and temporal resolution and the overall product quality of Copernicus Services.

- An alternative would be that the Copernicus Services fill the Cal/Val gap by performing additional cross-mission Calibration or Validation on non-Copernicus L2 products before assimilation. This scenario would be detrimental to other core activities of the Copernicus Services, and possibly performed by non-Cal/Val experts.

An optimal response to fill this gap for collaborative missions would be twofold:

- Setup a multi-Agency forum about between EUMETSAT and ESA for the Copernicus Program and collaborative Programs and Agencies (e.g. CNES and NASA for SWOT, NSOAS and CNES for the 6 HY2 satellites) to ensure that Cal/Val activities are consistent and coordinated, plus NRT communication channels to meet the timeliness requirements of Copernicus Services.
- Define and set up and fund complementary Collaborative Cal/Val activities needed to meet the Service requirements independently from Collaborative Agencies (e.g. near real time monitoring on essential multi-mission metrics, delayed time comparisons between Sentinels and Collaborative altimeters).

Another identified similar gap in the Copernicus processes specifically concerns the periodic reprocessing of the CMEMS & C3S Sea Level products for which specific Cal/Val studies are needed. These studies are needed to identify the optimal standards (geophysical corrections) and methods (editing, filtering…) to be implemented for the L2/L2P reprocessing. The standard must be aligned for Copernicus and non-Copernicus altimetry missions. These preliminary activities are currently not covered by the Copernicus program.

To illustrate, in the past, they were handled by the ESA CCI Sea Level project in 2017 (for the 2018 CMEMS/C3S reprocessing) or FDR4ALT for ENVISAT/ERS (ongoing), and by CNES fundings in 2020 (for the 2021 CMEMS/C3S reprocessing). For future reprocessing, it will be essential to anticipate the need of this essential undertaking and to ensure that the Copernicus Services still benefit from aligned standards and quality Cal/Val for reprocessed products. **Criticality – medium, effort needed – medium**

4.11. Section summary

- For altimetry Calibration and validation activities no critical gaps or limitation are identified.
- Most of the methods used are mature and well established. They mainly require minimal improvements and evolutions.
- Land surfaces represent a partial exception as the exploitation of altimetry observations over these surfaces have only recently started.
- The main criticality for altimetry observation over the ocean is the uncertainty definition at all spatio-temporal scales (i.e. individual observations, scales ranging between 10 and 100 kilometers eg the mesoscale, and global climate trends).

5. SAR component

5.1. Gaps in SAR Cal/Val vicarious methods

Two types of vicarious methods are used for the calibration and validation of the Sentinel-1:

- Using the backscattering of rain forest canopy to calibrate the elevation antenna pattern under assumption of flat gamma (radar brightness)
- Using the backscattering of sea surface at medium wind speed and a geophysical model of radar cross section to monitor the absolute radiometric calibration over the instrumented swath and the beam-to-beam offset. Those methods are originally related to the algorithm used for evaluation of wind speed over ocean as used in the generation of the Level 2 Ocean products.

Those methods are well controlled and are effective in C-Band for large range of wind speed in C Band and co polarisation VV. Still additional improvements of such models both for co polarisation HH and cross polarisation (HV or VH) on one side, and for high wind speed may be required.

In addition, the accuracy of the wind models and their effectiveness will have to be reassessed for L-Band, and explicitly for the ROSE-L mission. Most specifically, the maturity of the L-Band GMF (Geophysical Model Function) providing the relationship between wind speed/direction and radar cross section needs to be consolidated. While for C-Band multiple versions of such GMF were derived in the previous decades, benefiting from the scatterometer community, and improving with time, there is up to now and to our knowledge only one wind GMF in L-band that may need to be challenged.

5.2. Gaps in FRM Operations

Two types of FRM are used for calibration of Sentinel-1 SAR: Corner Reflectors and Transponders.

- The Corner Reflectors are used to validate and whenever appropriate to calibration the absolute geolocation accuracy of the product and to some extent to validate and calibrate the absolute radiometry.
- The Transponders are used to validate and calibrate the absolute radiometry of the products.

Multiple Corner Reflectors are used either as opportunistic targets (For instance, fields of corner reflectors deployed in the Surat Basin in Australia, without dedicated pointing matching the Sentinel-1 orbit) or optimised targets (Corner Reflectors operated by DLR in Germany and pointed specifically toward Sentinel-1 unit during overpass).

Only a limited set of Transponders are used for now, operated by DLR in Germany. They are pointed toward the Sentinel-1 unit during overpass.

The collection of data over those FRM is performed depending on actual mission acquisition plan. The Sentinel-1 mission can operate using four exclusive acquisition modes (Extra Wide Swath, Interferometric Wide Swath, Stripmap and Wave modes) using a set of configurable polarization configuration (Single polarisation H or V, Dual polarization H or V). The objective of the mission leads to (almost) constant acquisition plan over dedicated area to ensure continuity of measurements. For instance, over Europe mainland (out of the northern part), the acquisitions are performed in Interferometric Wide Swath Mode in Dual Polarisation V (IW DV) configuration. The data acquired over the DLR transponders and corner reflectors in Germany are then nominally acquired in this configuration, excluding the other ones.

The other modes are then calibrated using opportunistic measurements out of Europe (over corner reflectors or rain forest) using variability of acquisitions in different modes and extending the calibration of IW DV performed over European measurements. This is however an indirect calibration not using RFM as reference. The direct absolute calibration of all mode will require placing acquisitions over FRM in all the operated modes, that can only be possible for very limited period as conflicting with the nominal acquisition scenario. A short period of acquisition of IW Dual Polarization H (IW DH) over the DLR Transponders was set up in January/February 2021.

Operating Transponders in other areas of the globe for which there are no strong restrictions on switching from one acquisition mode to another will benefit to the overall calibration of the mission.

A set of transponders are operated by the Canadian Space Agency for the calibration and validation of the Radarsat Constellation Mission. Ensuring operation of those transponders and the acquisition of corresponding Sentinel-1 data could be a way forward.

5.3. Gaps in validation SAR GSLC and ARD products

The core SAR Level-1 products portfolio from the Sentinel-1 mission do not include GSLC (Geocoded Single Look Complex) and ARD (Analysis Ready Data). However, such products are under specification for instance as part of the CARD4L initiative (CEOS Analysis Ready Data for Land). This initiative is currently addressing the product family specification for SAR Normalised Radar Backscatter for which specific products could be generated separately for Land and Sea surface.

The CEOS group is currently defining guidelines for the CEOS ARD data for Land and Sea. Presentations of the status of those guidelines were exposed during the Living Planet Symposium 2022 [Rosenqvist et al 2022] and [Albinet et al 2022].

Those products are mostly defined with consideration of provided the information with the most available set of calibration, correction, projection, etc then enabling one user to be ready to analyse it without applying any heavy post processing and reformatting. The type of measurements provided in those products is by design the same as the one from the originating Level 1 product. Thus, the calibration and validation methodology of the main measurements should be the same as the originating Level 1 products. However, the set of correction/calibration/projection applied in those products may not be revertible, and this may prevent to replace them by more advance or local processing used for the calibration of originating L1 products.

5.4. Gaps in Inter-satellite comparison methods

5.4.1. Intercalibration methods

For Level 1 products (SAR images) there is no direct intercalibration method applicable to SAR for now as this would require acquiring data with very short-term revisit between two similar instrument units, thus with the risk of each of them interfering to the other. Intercalibration of two spacecraft operating in two distinct frequency bands cannot be considered as the backscattered energy over the same surface will not be the same depending on the frequency band.

For Level 2 products (derived information), intercalibration method can be foreseen considering two acquisitions of data from unit operating in different frequency bands (for instance Sentinel-1 vs ROSE-L) and considering the same derived information. However, this would require that the accuracy of the extraction of each derived information is similar (suitability of each band for the same measurement) and well defined.

5.4.2. Intercomparison with models

No specific gaps on intercomparison with models are identified so far.

Wind and oceanic models are used for comparison with SAR Level 2 products. The methodology is well defined. The improvements of those models benefit to the validity of comparison with SAR measurement, still using the same methodology.

5.5. Gaps in calibration/validation of new L2 products

For now, the only SAR operated for the Copernicus constellation is composed of the Sentinel-1 unit. This mission allows deriving one single Level 2 product dedicated to ocean measurements.

The future ROSE-L mission is designed to generate multiple other Level 2 products (to be confirmed with progress for ROSE-L mission) not only related to ocean measurement. The Calibration and Validation of those future Level 2 products will have to be characterised and defined with the progress of the ROSE-L mission design.

The targeted Level 2 products are listed in the ROSE-L mission requirement document. They are separated as Primary Objectives and Secondary Objectives. A summary of those products/applications is provided below.

Primary ROSE-L Level 2 Products:

- Geohazard Monitoring
- Forestry
- Land cover change and agriculture
- Soil moisture
- Floating ice
- Ice sheets
- Snow
- Maritime Surveillance (Safety & Security)

Secondary ROSE-L Level 2 Products

- Land surface Topography below vegetation
- Glacier and ice cap subsurface mapping
- Line of sight surface current

At the time of the preparation of this report, this list of primary and secondary products is not yet confirmed and then not fully defined. The corresponding calibration validation methodology will have to be addressed as part of their detailed definition and probably further refined.

5.6. Recommendations

This section provides a list of recommendations for R&D on Cal/Val method based on the gaps identified above, with discussion on the associated effort and criticality. The levels of efforts and criticality are defined as exposed in Table 2-1and Table 2-2 Define direct wind geophysical model functions (GMF) for co polarisation HH in C Band:

- For now, very accurate wind GMF are only defined for co polarisation VV. For co polarisation HH, the state of the art is to use the VV GMF as a baseline and to apply a so-called polarisation ratio, enabling to predict NRCS in HH based on NRCS in VV. This approach is not satisfactory as requiring placing a set of hypotheses on the backscattering in HH vs backscattering in VV. For instance, the dependencies of this polarisation ratio with respect to the wind direction is considered or not depending on the different models considered. A better methodology would be to derive a direct C Band HH GMF. This would require collecting large number of colocations of NRCS in HH together with wind information (speed and direction) and geometry of observation, and then to derive an empirical model using the same methodology as the one use to build the VV GMF. This may require placing specific acquisition campaigns with calibrating airborne radar instrument over ocean and capturing large range of wind speed (from very low to extreme high), with ground truth, and to complement this with large set of similar observations from spaceborne SAR instrument. Such HH GMF can then be used to support calibration of SAR NRCS measurement over ocean.

The effort is set to Medium to High as collecting large set of collocated data from SAR measurement is a long process, and from this defining a new GMF requires multiple steps of characterisation and validation. This can be High effort if airborne acquisition campaigns are set.

The criticality is Medium, as data acquired in HH polarisation over open sea is only a subset of the acquisition scenario of current SAR mission in C Band. **Criticality – medium, effort needed – medium/high**

- Define wind geophysical model functions (GMF) for C band in cross polarisation

The capacity to measure wind speed over ocean from cross polarisation (HV or VH) was demonstrated from SAR data over high wind (typhoon and hurricanes). However, the Normalised Radar Cross Section (NRCS) over ocean is low, and then associated to a low signal to noise ratio. Defining a cross polarisation GMF requires first to very accurately compensate for the sensor noise that was not possible up to now. The recommendation is to define an accurate cross polarisation GMF taking benefit of the noise compensation of the sensor. Such cross polarisation GMF can then be used to support calibration of SAR NRCS measurement over ocean.

The effort is set to Medium to High (as for GMF in C band in HH).

The criticality is set to Medium, as for the acquisition plan for Sentinel-1 the cross-polarisation data is only available in dual polarisation acquisition, for which the co-polarisation (HH or VV) is available. Having such fine GMF available should allow to inter compare the calibration of cross polarisation NRCS, but not necessarily to validate it (depending on noise level) and to derive better wind speed over hurricanes, but still with limitations related to the availability of ground truth for such events. **Criticality – medium, effort needed – medium/high**

- Define wind geophysical model functions (GMF) for L band in cross polarisation

The capacity to measure wind speed over ocean from L Band data SAR was demonstrated by [Isoguchi 2009]. However, from the literature only one such model is available based ALOS 2 / PALSAR instrument only. This model needs to be challenged and compared with capabilities of other L Band instruments. Defining an alternative GMF needs to collect large colocation of L Band SAR calibrated data with proper denoising with ground truth of wind speed over ocean. Specific airborne campaigns of NRCS measurements over large range of wind speed will be required as well. The availability of a robust L Band GMF will benefit to L Band SAR calibration over ocean.

The effort is set to High. The activity is similar to the one related to GMF in C band HH or cross polarisation, but requiring to collect large set of non-Copernicus Data (ALOS, SAOCOM…), to check their calibration, consistency, etc before defining the model. In addition, the need of airborne campaign shall be anticipated.

The criticality is set to High, as not having access to fine L band GMF may prevent usage of the NRCS over ocean to contribute to radiometric calibration of the Level 1 products, and of the L2 wind over the ocean products. **Criticality – high, effort needed – high**

- Define calibration/validation of native L2 products of the ROSE-L mission together with their detailed definition

The native L2 products of ROSE-L are under definition. It is recommended to define the methodology for their calibration and validation as part of their detailed definition. **Criticality – high, effort needed – high**

5.6.1. Recommendations summary

Thus we summarize the above recommends in the following points to be considered for R&D:

- Improved GMF for C band HH; **Criticality – medium, effort needed – medium/high**
- Improved wind GMF for C band cross- polarisation; **Criticality – medium, effort needed – medium/high**
- New GMF for L band; **Criticality – high, effort needed – high**
- Definition of ROSE-L Cal/Val methodology of new L2 products as part of their detailed definition; **Criticality – high, effort needed – high**

6. Atmosphere component

6.1. Gaps in atmosphere On-board calibration methods

S5P TROPOMI (Folkert Boersma and Gijsbert Tilstra, KNMI)

A few vicarious techniques using natural Earth targets and other natural radiation sources have been developed successfully and are applied, sometimes in routine, to the Cal/Val of atmospheric composition sounders. Natural targets and vicarious techniques mentioned hereafter in this section have been used for the calibration of:

- In the UV-VIS-NIR: SCIAMACHY, OMI and Sentinel-5p TROPOMI, the precursors of the Sentinel-4 and Sentinel-5 series
- In the SWIR: SCIAMACHY, OCO-2/3 and GOSAT-1/2, heritage missions for CO2M

Ludewig (TROPOMI) identified as a principal gap in the use of vicarious methods for TROPOMI that “such methods do not provide possibilities to achieve accuracies better than on-ground. All Earth-based measurements rely to some extent on external models, which affects the absolute accuracy”.

- Rayleigh Scattering over Ocean
- Deep convective clouds (DCC)
- Desert Pseudo-Invariant Calibration Sites (PICS)
- etc

6.2. Gaps in atmosphere ground-based Cal/Val methods

6.2.1. Calibration

For TROPOMI, ground-based calibration methods have been used to limited extent. This is rooted in the focus on in-flight instrument calibration, and scepticism expressed on the possibility to improve calibration key data from on-going ground-based cal/val activities. The main ‘gap’ here could be described as a lack of desire or confidence in using natural targets, intercomparisons with other satellite measurements, and inferences based on level-2 analysis as means to update calibration data.

6.2.2. Validation

For each primary product (i.e., those to be measured by S5p, S4, S5 and CO2M and for which user requirements are formulated in CCVS D1.4), the maturity of the (ground-based) validation methods was assessed along (1) the different steps in the end-to-end validation chain defined by the generic validation protocol (Table 3 in D1.4) and (2) the different maturity levels and corresponding requirements as defined in the CEOS Data Management and

Stewardship Maturity Matrix for data product validation (CEOS-DMSMM, Table 4 in D1.4). To that end, validation experts for each specific product were contacted, either within the project team or external, and requested to complete a survey template built to gauge the maturity of the validation methods. As a summary of that survey, an overview of the maturity levels for each product and step of the validation protocol is visualized in Table 6-1.

Table 6-1: Maturity assessment for the validation methodology for each atmospheric composition product, for the 12 steps in the generic validation protocol (Table 3 in D1.4). The colour scale corresponds to the different maturity levels of the CEOS-DMSMM, where red = not managed, orange = limit managed, yellow = managed and green = well managed.

Product		Design	Data selection	Data content	Information content	Co-location	Co-location content	Harmonization	Comparison	QI	Acceptance	Reporting	Feedback
L1b		⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜
Ozone (O_3) total column		🟩	🟩	🟨	🟥	🟩	🟨	🟨	🟩	🟩	🟩	🟩	🟨
Ozone (O_3) tropospheric column		🟨	🟩	🟨	🟥	🟧	🟧	🟧	🟨	🟨	🟨	🟩	🟨
Ozone (O_3) vertical profile		🟨	🟩	🟨	🟨	🟨	🟧	🟧	🟨	🟩	🟨	🟩	🟨
Nitrogen dioxide (NO_2) (sub-) columns		🟨	🟩	🟨	🟥	🟨	🟨	🟨	🟨	🟨	🟨	🟩	🟨
Formaldehyde (HCHO) column		🟨	🟩	🟨	🟥	🟨	🟧	🟨	🟨	🟨	🟨	🟩	🟨
Glyoxal (CHOCHO) column		🟧	🟩	🟧	🟥	🟧	🟥	🟧	🟨	🟧	🟧	🟨	🟧
Sulfur dioxide (SO_2) column		🟧	🟧	🟧	🟥	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧
Carbon monoxide (CO) column		🟨	🟩	🟨	🟥	🟨	🟧	🟨	🟨	🟨	🟨	🟩	🟨
Methane (CH_4) column		🟨	🟩	🟨	🟥	🟨	🟧	🟨	🟨	🟨	🟨	🟩	🟨
Carbon dioxide (CO_2) column		🟨	🟩	🟨	🟥	🟨	🟧	🟨	🟨	🟨	🟨	🟩	🟨
Water vapor (H_2O) column		🟨	🟩	🟧	🟥	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨
Cloud properties		🟧	🟧	🟧	🟥	🟨	🟧	🟧	🟧	🟧	🟧	🟨	🟨
Lambertian Equivalent Reflectivity (LER)		⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜
Surface albedo		⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜
Aerosol properties	AOD	🟨	🟩	🟨	🟥	🟨	🟧	🟨	🟨	🟨	🟨	🟩	🟨
	other	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜
Solar-Induced Fluorescence (SIF)		⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜

Below, a synthesis of the identified gaps is provided following the different steps in the validation procedure.

Design of the validation study

- Only for S5P-TROPOMI total ozone are ex-ante/reported uncertainties now being validated along with the measurements (e.g., MPC-VDAF Routine Operations Calibration and Validation Report, #14, 2022). These methods need to be adapted and implemented for all other products in order to reach full metrological traceability. This will also require advances in the uncertainty reporting for both the satellite and ground-based data (see for instance the AMT special issue on “Towards Unified Error Reporting”, and the advances in uncertainty characterisation of ground-based reference measurements in EUBREWNET and in the Pandonia Global Network.).
- Incomplete traceability for the validation process due to insufficient focus on the complete uncertainty budget in the study design. This concerns both the propagation of measurement uncertainties through the validation process, and the characterization of errors and uncertainties introduced by the various manipulations and by the use of auxiliary data such as meteorological data for unit conversions or target product climatologies used for vertical harmonization of measurements with an incomplete vertical coverage (Keppens et al., 2019).
- Meta studies, bringing together independent validation studies on comparable products, are hampered by unharmonized methodology. Community guidelines and/or detailed and broadly accepted protocols would be beneficial (cf. the CEOS Land Product Validation protocols).
- The application of some steps of harmonization between satellite and ground-based measurements depends on the target user (in particular the use of vertical Averaging Kernels and replacement of prior profiles). The operational validation of the Copernicus missions should produce a quantified assessment of the impact of these optional harmonization steps.
- For several products, the detection of some (systematic) errors requires a validation of aggregated satellite pixel data (GHG, CHOCHO, HCHO). Research is needed on how best to translate user/mission requirements to validation requirements on these aggregated data, including for instance spatial error correlation length studies.
- There is a need for more awareness on - and uptake of - more advanced validation methods such as triple co-locations, structure functions, process validation, indirect validation (see Loew et al., 2017 for a review of such methods). This could be facilitated through the publication of community guidelines or protocols, and the organization of dedicated training such as summer schools. Dedicated funding for proof-of-concept studies would be beneficial.

Data and information content analysis

- Impact of satellite data filtering (using the recommended thresholds on various quality indicators/flags) on data content, i.e., spatio-temporal coverage of the data set, is rarely investigated (O3P, GHG, NO2, CHOCHO, HCHO). This should be an explicit part of the protocol.

- Information content analyses, e.g., by studying retrieval properties accessible through the Averaging Kernels such as vertical resolution, vertical sensitivity offsets, prior contribution, etc., are often (O3P) or always (GHG, NO2, CHOCHO, HCHO) missing.
- Averaging Kernel analysis for aggregated products needs theoretical work to ensure correct propagation of the information contained in the underlying AKs. (cf. the discussion on an “averaged AK”, by von Clarmann & Glatthor, AMT, 2019). This has further implications for the use of the AKs to harmonize satellite and ground-based vertical sensitivity.

Spatio-temporal co-location

- Co-location mismatch errors are non-negligible for most of the current validation studies. As such, co-location strategies need to be optimized to reduce mismatch uncertainties using effective FOV and/or wind information. This requires research on the actual spatio-temporal sensitivity of the measurements (H2020 GAIA-CLIM gaps G3.0x).
- The optimal compromise between mismatch errors and sufficient co-locations for robust statistics needs to be quantitatively assessed (proof-of-concept for total O3 available in Verhoelst et al., AMT, 2015)
- The results from the research recommended above needs to be translated to standards and incorporated in the protocols. The resulting harmonization would facilitate meta-analyses.
- Data content analysis after co-location needs to be performed to assess the representativeness of the validation results in terms of geographical and influence quantity coverage.
- Mismatch amplitude (e.g., typical separation in time and space) and related uncertainties due to geophysical variability at those scales need to be quantified. Requires information on natural variability on those scales, which can often only be obtained from in-situ or remote sensing from aircraft campaigns (e.g., Sparling et al., JGR, 2006).

Data harmonisation

- Uncertainty introduced through the use of auxiliary data for harmonization needs to be quantified (related to H2020 GAIA-CLIM gap G5.07).
- Optimal vertical resolution/sampling harmonization should be studied and a harmonized use adopted.
- Mutual smoothing needs to be adopted for SAT-SAT intercomparisons of vertical profiles.
- Harmonization procedures need operationalization (GHG, NO2, HCHO)
- Vertical misalignment (mountain-top stations) should be addressed with estimates of the unobserved column/profile, taking into account the actual sensitivity profile of the satellite measurement and the shape of the prior profile (if applicable).

- Quality of photochemical adjustments for temporally separated measurements of products with a strong diurnal cycle needs to be assessed (stratospheric NO2).

Data comparison and analysis

- The set of comparison statistics needs to be improved to (1) make them robust, (2) take into account the ex-ante uncertainties and the comparison uncertainties, and (3) assess specifically extreme values/events. This is in particular the case for the regression techniques used for drift estimates.
- Differentiation between multiplicative and additive biases is needed.

Reporting & Feedback

- Feedback to FRM data providers is for many products considered a weak point, in particular when no single-point-of-contact exists for the ground-based network. A more formal feedback mechanism may be desirable.

Analysis per product

Looking at the results of the maturity assessment per product type, some specific gaps can also be grouped as follows:

- For the vertically resolved species (O_3 profile and tropospheric column, and to some extent also the NO_2 subcolumn products), vertical harmonization between satellite and ground measurements remains a challenge: how (and to what extent) can and should we take into account different assumed (prior) profiles, different vertical sensitivity, different definitions of the tropopause, … This requires both research and - based on those outcomes - the definition of protocols.
- For several species, the maturity is overall very low. This is either due to a lack of heritage, e.g., as a consequence of the unavailability of reference measurements (glyoxal and SO2), or due to the very different nature of the satellite and ground-based products (cloud properties). For glyoxal and SO2, support for methodological development should follow that for the procurement of reference measurements. For cloud properties, dedicated research is required to assess and improve intercomparability.
- For some species, no systematic validation is performed (yet) in the context of the ATM-MPC, SAFs, or C3S/CAMS: L1b (planned with the ATM-MPC for S5P), LER, surface albedo, aerosol properties (besides AOD), and SIF. This is a gap at a level above methodology.

6.3. Gaps in atmosphere Inter-satellite comparison and validation

This part includes inter-satellite direct comparisons as well as model/satellite comparisons and double differences comparisons.

6.3.1. Assessment template design

The same template as used for the ground-based cal/val methods has been used because most answers can apply to both ground-sat and sat-sat comparisons (ex: considerations about averaging kernels, colocations, uncertainties etc).

Moreover, feedback was collected from exchanges with several experts (CNES, KNMI, LSCE, LMD, JPL...) in informal way, as well from the literature.

People in charge of inter-comparisons were asked about the limitations and gaps they face in the methods and algorithms for the following elements:

- Spatio-temporal aspects: definition of Co-location (use of airmass trajectories?), and their limitations (latitudinal sampling depends on the orbits)
- Algorithms:
  - for the filtering of the data (selection of quality flags, cloud/clear scenes, before or after bias-correction...) (also critical for massive averaging differences),
  - for assessing the uncertainties (how to estimate error bars in the inter-comparisons and separate calibration vs natural geophysical variations, statistical approaches, representativity of the bias with limited number of samples...)
- For Level 1 comparisons only: methods for comparing instruments with different spectral response functions, interpolations for spectral gaps...
- For level 2 comparisons only: techniques for comparing columns with different weighting functions and a priori, comparisons of profiles and columns...

6.3.2. Results

Results are per type of inter-comparisons and, if needed, per product:

- Level 1: satellite inter-comparisons of radiance spectra (in UV/VIS/NIR/SWIR, in TIR), for radiometric and spectral calibration.
  - The main challenge is to extend the operational TIR radiometric and spectral inter-comparisons as coordinated by GSICS to UV/VIS/NIR/SWIR domain. Indeed, the level 1 spectra satellite-to-satellite comparisons are operational for thermal infrared spectrometers but not for UV-VIS-SWIR spectrometers. There is a single study, reported in the paper by Kataoka et al. which shows comparisons between OCO-2 and GOSAT spectra, for a 3 year period. The maturity is thus very low. It seems that there is currently no plan in the framework of CEOS AC/VC to develop operationally such comparisons (only XCO2 and XCH4 satellite-to-satellite comparisons are noticed in the CEOS AC/VC white paper, chap 6.3 TBC).
  - Kataoka et al. has identified several difficulties, including the fact that, because of different viewing angles between both sensors, the BRDF of the surface has to be taken into account. This requires the use of MODIS BRDF data. Also, one limitation comes from the local equatorial crossing time similar between missions, close to

noon, which might lead to less coincidence of orbits except near to the poles where there are no satellite observations because of the low sun angle (for example, in Kataoka et al, only about 400 co-locations have been found in almost 3 years for OCO-2 and GOSAT). Geostationnary satellite compared with LEO satellite might reduce this difficulty (for example, for S4 and S5 future missions).

  - Ludewig (TROPOMI) stated skepticism towards inter-satellite comparisons. For UV/Vis irradiance calibration, successful intercomparisons have been made, and led to improved calibration key data. For other aspects of calibration, intercomparison activities would be *"useful if the calibration of the compared instrument is truly independent. It is not always evident which radiometric calibrations have been applied (to the reference instrument). If the calibration is linked to vicarious calibration, a comparison can become meaningless and misleading."*
  - Other potential limitations (to be investigated):
    - high heterogeneity of surface reflectance combined with different ground pixel size imply either to filter the co-location on very homogeneous scenes, or to average a very high number of co-locations to get this effect far below the expected accuracy.
    - From the lessons learned in the TIR domain, attention must also be paid on methods for comparing instruments with different spectral response functions, with complex methods necessary for interpolations for spectral gaps (Xu et al., 2018).

- Satellite to model comparisons :
  - CO2 and CH4: Comparisons with CAMS have been conducted (Chevallier et al., figure 2 and 3, Tu et al. 2020), however, the challenge relies in the fact to go from limited study (in space/time) to systematic comparisons with a purpose of satellite validation.
  - O3 : Inness et al. The comparison between satellite level 2 product and CAMS value is realized as a first step for the data assimilation in CAMS. However, it is not analyzed as a tool to realize the validation of the satellite products.

Use of model for OCO-2 bias correction: the bias-correction is a critical step for CO2 mission as it removes regional biases caused by, e.g. surface pressure, aerosols, variations of ground albedo etc (O'Dell et al., 2018). Indeed, at global scale, regional biases have a strong negative impact on the level 4 (source and sinks estimates) obtained through assimilation. The bias correction on XCO2 could be considered as a kind of "calibration" of the level 2. This is realized by several methods, one of them takes benefits from comparisons between satellite XCO2 and 6 model fields to get a training data set (O'Dell et al. 2018, part 4.1). Indeed, they wrote that models may disagree to up to 1.5 ppm. Using only one model would lead to less robust results.

- Level 2 : satellite-satellite inter-comparisons

About the spatio-colocations aspects, same limitations are identified for level 2 comparisons as for level 1 comparisons above.

Satellite-to-satellite comparisons have the main advantage, compared to ground-satellite comparisons, that the averaging kernels or weighting function are closer one to each other because of the similar “down-up” geometrical path. For identical instruments on different platforms, this is a huge advantage. However, this it not true when comparing level 2 products from different sensors, especially with different spectral range (e.g. methane product from a SWIR spectrometer and from a TIR spectrometer).

In that case, same limitations as comparison with ground instrument are faced: no shared guidelines on how to take into account different averaging kernels for each satellite product when comparing different sat even if some theoretical work exists (Rodgers and Connor, 2003). Mutual kernel smoothing can be applied, especially e.g. when UV-VIS-SWIR and IR measurements are compared (ozone, methane...). This approach is not systematically applied nor have its effects been thoroughly examined.

The influence of the a priori in the smoothing equation of retrieved columns for CO is also something that should be considered more systematically, as shown by (Wizenberg et al, 2021).

A misrepresentation of the wavenumber-dependent surface emissivity or albedo in the level 2 retrieval is also pointed as a source of discrepancy for, e.g. comparisons of ozone products (Boynard et al., 2018). Thus, improving spectral model of surface should help level 2 satellite to satellite comparisons when different spectral bands are used.

Finally, comparison of level 2 products from satellites can hardly serve as a validation tool if the inputs of the retrieval algorithms differ too much (whereas ground description, a priori profiles, cloud filtering, spectroscopy).

- <u>Level 3 : satellite-satellite inter-comparisons</u> It is sometimes mentionned that Level 3 products may facilitate satellite-satellite inter-comparisons (as a way to counterbalance the limited number of co-locations) (e.g. for NASA/AIRS atmospheric mission Granger et al., 2004) However, no such work has been performed to our knowledge for atmospheric missions.

6.4. Synthesized recommendations

- Target the simultaneous validation of the measurements and their reported (ex-ante) uncertainties. From this follows the need to develop a complete end-to-end metrological uncertainty budget of the validation process (reference data, auxiliary data, co-location mismatch, impact of various harmonization and aggregation operations…). **Criticality – high, effort needed – high**
- Integrate systematic data and information content analyses on the satellite data sets in the validation protocols, also post co-location to assess the representativeness of the validation study, not only in geographical coverage but also in influence quantity coverage. **Criticality – high, effort needed – medium**

- Satellite-to-ground co-location and harmonization strategies need to be optimized, harmonized (to allow meta-analyses) and standardized; irreducible co-location mismatch needs to be quantified. **Criticality – high, effort needed – medium**
- Facilitate the uptake of advanced validation methods, i.e. those going beyond baseline statistical analysis on pair-wise comparison of directly comparable measurements, in operational systems. Includes an educational component. **Criticality – low, effort needed – medium**

7. Multi-sensor synergies

For cases where the same geophysical variable is measured by different satellite-based sensors, several analyses have already been carried out. However, they remain punctual and could be performed in a more systematic way. This kind of analyses requires exchanges between the different CalVal communities to share information about sensors characteristics, limitations and uncertainties. The gap identified is small since there are several valuable alternatives cross-comparison methods (e.g. with in-situ observations and model simulations). Nonetheless, a strengthening of such comparisons could bring additional outcomes and improve the error characterizations. As examples, we recommend the following cases of study:

- Analyze the quality of altimetry Significant Wave Height (SWH) and Sea State Bias (SSB) correction through the comparisons with Sentinel-1 Level-2 Ocean products. These products provide information about the wave spectrum, the wind direction and intensity as well as the surface radial velocity. Analogously, such comparisons can also be performed with CFOsat products.
- Compare the altimetry wind speed estimation with scatterometer products (such as the ones from ASCAT satellites).

For activities that combine sensors that do not measure the same physical variable, the objectives are slightly different. Such analyses can be used to bring additional, complementary information to better characterize the geophysical environment around the altimetry measure. Several cases of study have already been investigated, for instance to characterize the internal waves signature in the altimetry signal (internal wave signal confirmed by optic images), or to confirm the leads detection (leads location confirmed with Sentinel-2 optic or Sentinel-1 SAR images). These analyses mixing different technologies are interesting: they bring to the analysis different variables and diagnostics that can provide additional context for the interpretation of the altimetry signal and, thus, help its validation and understanding.

Some potential synergies between SAR and altimetry imagery for sea ice detection. The altimeter data can be used detect open sea vs sea ice, which requires to perform a classification of the wave forms between those two classes. Validation of the classification requires a reference dataset regularly updated. SAR imagery allows mapping arctic areas covered with sea ice, and to segment the area depending on the type of ice. Crossovers between altimetry and SAR observations can be used for this purpose. Here, crossovers are defined as observation of the same area at the same time or with a limited time shift. [Longépé 2019]

- Probably some synergies between atmosphere and optical missions for calibration, and for aerosol products (To be completed in the final version)

8. Conclusion

The gap analysis of the vicarious methods used in the calibration and validation of Earth Observation satellites relevant to the Copernicus programme has been presented in this document over the four components optical, altimetry, Radar/SAR and atmospheric compositions. The recommendations for the research and developments required to fulfil this gap when/where possible have been suggested per section and component as well.

The document identifies the gaps and limitations of the CalVal methods, assesses their criticality and required efforts in order to provide a prioritized list of recommendations for R&D activities on the CalVal methods.

Vicarious CalVal methods are an essential part of Cal/Val solutions. It corporates a large set of algorithms and technologies for the calibration and assessment of the on-board calibration and products validation. Thus these methods must be of the highest possible accuracy to cover the aspects of Cal/Val activities. Although the vicarious methods are used since decades, they need updating, and new developments are expected to cover the gaps.

The nature of gaps and, therefore, the recommendations vary across the mission types as presented in the previous chapters. However, some gaps are common in different fields, and actions should be taken to overcome them such as:

- Estimation of uncertainty and traceability to the SI, however, a complete characterization of uncertainties for the different spatio-temporal scales and for the different surfaces (ocean, inland waters, land & sea ice) is important to provide feedbacks toward agencies and end-users.
- Further development required in the validation coming from radiative transfer model
- Developments on new validation methods for special atmospheric correction steps like terrain correction.
- Define a community agreed validation strategy for classification-based mapping that ensures representativeness, in terms of landscape, in time and space
- Provide (an objective) definition of “cloud”, which (ideally) includes a numerical metric including definitions for transparent clouds and cloud boarder.
- Develop an open-source cloud masking validation reference archive for Sentinel sensors in order to foster the development of new algorithms
- To foster the interoperability of systems, it is important that the viewing directions of the different operational sensors (COPERNICUS and beyond) are well calibrated; so sharing some public dense and accurate references data between satellite operators would benefit this objective.
- Tandem phase opportunities should be implemented whenever practical (during commissioning phase or at end of mission). Similarly, dedicated cal/val activities

should be planned to benefit from tandem configurations with a Copernicus satellite (e.g. FLEX with Sentinel-3).

- Development of a community processor providing reference data from ground measurements. This processor should consider the calibration procedure, modelling of surface BRDF from field measurements, necessary QA-checks and yield measurement uncertainties in a clearly defined procedure. This includes permanent updating of best practices documents with necessary measurement protocols.
- Satellite-to-ground co-location and harmonization strategies need to be optimized, harmonized (to allow meta-analyses) and standardized; irreducible co-location mismatch needs to be quantified.

End of Document