Ocean Colour Multi-Mission Algorithm Prototype System (OMAPS)

Product Validation and Evaluation Report

Ref: D4.3
Date: 27/01/2022
Issue: 1.5
For: EUMETSAT
Ref: ID 995318

Table of Contents

List of Figures 3
Document Control 5
Applicable Documents 5
Reference Documents and sources 6
List of Acronyms 7
1 Purpose 9
2 Configurations compared 9
2.1 Product quality flags 9
2.2 SVC Gains 12
3 Validation Methods 12
3.1 Match-ups against in-situ 13
3.2 Atmospheric Round Robin intercomparison 14
3.2.1 Atmospheric round robin configuration 14
3.2.2 AC RR scoring 15
3.2.3 AC RR data used 18
3.3 In-water Round robin 19
3.3.1 IW RR configuration 19
3.3.2 IW RR scoring 21
3.3.3 IW RR data used 21
4 Results 22
4.1 AC RR results 23
4.2 In water Round Robin 32
4.2.1 Assessment per water class 33
4.3 Performance under known adverse conditions 34
4.3.1 Phytoplankton Bloom Granules 35
4.3.2 Aerosol Granules 36
4.3.3 Cloud Granules 37
4.3.4 Glint Granules 38
4.3.5 Sea Surface Condition Granules 38
4.3.6 Corrupted Data Granules 39
4.4 Time series analysis 51
5 Summary and recommendations 56
Acknowledgements 57

List of Figures

Figure 1: OCDB in-situ matchup locations against diagnostic granules. Locations of matchups are shown in red and are all clustered close to Hawaii, reflecting that the matchups are almost entirely from the MOBY site. 18
Figure 2: In situ spectra that were matched to POLYMER data as part of the AC RR CBQ analysis. As can be seen, almost all the spectra show a strong blue signal, typical of open ocean (case 1) waters. 19
Figure 3: Flow chart of multi-metric scoring approach for In-water algorithm comparison........ 21
Figure 4: OCDB in-situ matchup locations against diagnostic granules. Locations of matchups are shown in red and are clustered around three regions. 22
Figure 5: Example atmospheric CBQ RR total scores in a comparison of L2GEN, POLYMER, SACSO and BASELINE processors. These distributions of scores are derived by bootstrapping,........ 23
Figure 6: Example atmospheric CBQ RR performance for the mean percent difference (MPD) metric........ 25
Figure 7: Example atmospheric CBQ RR performance for the mean absolute percent difference (MAPD) metric. 26
Figure 8:Example IBQ Atmospheric Round Robin Total scores comparing POLYMER, SACSO, L2GEN and BASELINE processors........ 27
Figure 9: Example IBQ AC RR number of matchups available per processor. Box and whisker plots show variance across bootstrapping exercise, diamond points are classed as outliers (as determined by python seaborn.boxplot function). The l2gen value is non-variable across the bootstraps and so appears as a single horizontal line in this plot. 27
Figure 10: Example scatterplots of CBQ matchups between BASELINE processed OLCI data and in-situ measurements. Only points matched to all 4 processor configurations are shown. Dashed line shows 1:1 relationship........ 28
Figure 11: Example scatterplots of CBQ matchups between L2GEN processed OLCI data and in-situ measurements. Only points matched to all 4 processor configurations are shown. Dashed line shows 1:1 relationship........ 29
Figure 12: Example scatterplots of CBQ matchups between POLYMER processed OLCI data and in-situ measurements. Only points matched to all 4 processor configurations are shown. Dashed line shows 1:1 relationship........ 30
Figure 13: Example scatterplots of CBQ matchups between SACSO processed OLCI data and in-situ measurements. Only points matched to all 4 processor configurations are shown. Dashed line shows 1:1 relationship........ 31
Figure 14: Example scores per processor and per variable for CBQ analysis. 31
Figure 15: Example scores per processor and per variable for IBQ analysis. 32
Figure 16: Summary plot of normalised multi-metric scores across six candidate chl-a algorithms using in-situ data from the OCDB for matchup and metric generation. 32
Figure 17: Example comparison of matchup performance between the OC2 and OC4 algorithms. Statistics shown are correlation coefficient (r),RMSE (ψ) unbiased RMSE (Δ), bias (δ), slope of type 2 regression (S), intercept of type 2 regression (I) and number of matchups (n). 33
Figure 18: The water classes used by default in the OMAPS processor. 34
Figure 19: Comparison of (left to right) SACSO, L2GEN, POLYMER and EUMETSAT BASELINE Collection-3rd processing for Granule S3A_OL_1_EFR____20180524T102451_20180524T102751 , containing a phytoplankton bloom in the North Sea. Blue and orange outlined regions highlight areas of likely coccolith bloom and enhanced noise in L2GEN and BASELINE respectively. For closer inspection see appendix_Phytoplankton_bloom_granules.pdf for higher resolution version of image........ 40
Figure 20: Comparison of (left to right) SACSO, L2GEN, POLYMER and EUMETSAT BASELINE Collection-3 processing for Granule S3A_OL_1_EFR____20180716T073803_20180716T074103 , containing a phytoplankton bloom in the Red Sea. Orange outlined region shows area of chlorophyll-a

depicted eddy structures (visible in SACSO processing). For closer inspection see appendix_Phytoplankton_bloom_granules.pdf for higher resolution version of image. 41
Figure 21: Comparison of (left to right) SACSO, L2GEN, POLYMER and EUMETSAT BASELINE Collection-3 processing for Granule S3A_OL_1_EFR____20180730T080327_20180730T080627, containing a phytoplankton bloom in Barents and White Sea area. Orange outlined region shows area of coccolith bloom and blue region shows anomalous POLYMER behaviour. For closer inspection see appendix_Phytoplankton_bloom_granules.pdf for higher resolution version of image. 42
Figure 22: Comparison of (left to right) SACSO, L2GEN, POLYMER and EUMETSAT BASELINE Collection-3 processing for Granule S3A_OL_1_EFR____20171210T183320_20171210T183620 , containing a plume of wildfire smoke West of California. Orange outlined region shows an area of cloud shadow impact on Rrs estimates. Green outline shows an area of dust contamination. For closer inspection see appendix_Aerosol_Granules.pdf for higher resolution version of image. 43
Figure 23: Comparison of (left to right) SACSO, L2GEN, POLYMER and EUMETSAT BASELINE Collection-3 processing for Granule S3A_OL_1_EFR____20180726T215905_20180726T220205 , containing a plume of volcanic ash near Fiji. Orange outlined region shows area of differing chlorophyll-a estimates proximal to land. Green outline shows area of differing chl-a estimates between SACSO and BASELINE. For closer inspection see appendix_Aerosol_Granules.pdf for higher resolution version of image. 44
Figure 24: Comparison of (left to right) SACSO, L2GEN, POLYMER and EUMETSAT BASELINE Collection-3 processing for Granule S3A_OL_1_EFR____20180326T110646_20180326T110946 , containing a massive Saharan dust plume over the Eastern Atlantic. Orange outlined region shows area of contaminated data in POLYMER processing. Green outline shows area of contaminated data in BASELINE processing. Red outlined area shows suspiciously elevated chlorophyll-a estimates from SACSO processing. For closer inspection see appendix_Aerosol_Granules.pdf for higher resolution version of image. 45
Figure 25: Comparison of (left to right) SACSO, L2GEN, POLYMER and EUMETSAT BASELINE Collection-3 processing for Granule S3A_OL_1_EFR____20180419T205315_20180419T205615 , containing a mix of clouds and cloud shadows over the North Pacific. Orange outlined region shows area of increased SACSO coverage. For closer inspection see appendix_Cloud_type_granules.pdf for higher resolution version of image. 46
Figure 26: Comparison of (left to right) SACSO, L2GEN, POLYMER and EUMETSAT BASELINE Collection-3 processing for Granule S3A_OL_1_EFR____20180524T022054_20180524T022354 , containing a many small clouds over the ocean south of Indonesia. Orange outlined region shows area of potential cloud shadow impact. For closer inspection see appendix_Cloud_type_granules.pdf for higher resolution version of image. 47
Figure 27: Comparison of (left to right) SACSO, L2GEN, POLYMER and EUMETSAT BASELINE Collection-3 processing for Granule S3A_OL_1_EFR____20180518T195153_20180518T195453, containing a large band of high glint in the right of the image. Orange outlined region shows area with structured chlor-a field that crosses glint boundary region. For closer inspection see appendix_Glint_granules.pdf for higher resolution version of image. 48
Figure 28: Comparison of (left to right) SACSO, L2GEN, POLYMER and EUMETSAT BASELINE Collection-3 processing for Granule S3A_OL_1_EFR____20180524T134649_20180524T134949 , containing a mix of clouds and sea-ice near Greenland. Orange outlined region shows area with some speckle in 443 band. Green outlined region shows large range in chlorophyll-a estimates between the different processors. For closer inspection see appendix_Sea_surface_conditions_granules.pdf for higher resolution version of image. 49
Figure 29: Comparison of (left to right) SACSO, L2GEN, POLYMER and EUMETSAT BASELINE Collection-3 for Granule S3A_OL_1_EFR____20160614T202246_20160614T202446_20170930T194125_0119_005_185 , containing an OLCI camera anomaly. 50

Figure 30: Map showing locations of HOTs and MOBY measurements. 51
Figure 31: Example comparison of the extracted mean, median, standard deviation and number of valid pixels for chl-a for SACSO, L2GEN, POLYMER and BASELINE processors and the HOTS dataset. 53
Figure 32: Example comparison of the extracted mean, median, standard deviation and number of valid pixels for Rrs 443 (or reflectance/pi for BASELINE) for SACSO, L2GEN, POLYMER and BASELINE processors. Data extracted from a 5x5 pixel area centred on the MOBY location.For the mean and median Rrs panels we also show the MOBY in-situ measurements of Rrs at the relevant OLCI bands for comparison. MOBY data are however not fully band shifted to the OLCI nominal band at 442.5 nm. 54
Figure 33: Example comparison of the extracted mean, median, standard deviation and number of valid pixels for Rrs 510 (or reflectance/pi for BASELINE) for SACSO, L2GEN, POLYMER and BASELINE processors. Data extracted from a 5x5 pixel area centred on the MOBY location. For the mean and median Rrs panels we also show the MOBY in-situ measurements of Rrs at the relevant OLCI bands for comparison. MOBY data are however not fully band shifted to the OLCI nominal band at 510 nm. 55

Document Control

Version No.	Release Date	Author/Contributor	Reason for issue
0.1	26/03/2021	Thomas Jackson	Initial draft for internal review
0.2	26/05/2021	Thomas Jackson	Updates with statistics from multiple atmospheric processors
1.0	13/10/2021	Thomas Jackson	Updated to include 4 atmospheric processors and both AC and IW RR results.
1.1	08/11/2021	Thomas Jackson	Updates following review of v1.0 by EUMETSAT.
1.2	30/11/2021	Thomas Jackson	Updates following internal review
1.3	28/12/2021	Thomas Jackson	Updates following EUMETSAT feedback
1.4	07/01/2022	Thomas Jackson, Shubha Sathyendranath, Dagmar Mueller, Ben Calton	Internal updates/review before final submission
1.5	25/01/2022	Thomas Jackson, Shubha Sathyendranath, Dagmar Mueller, Andre Valente, Ben Calton	Final submission version within OMAPS project

Applicable Documents

ID	Document
AD-1	OMAPS Product Validation Plan (PVP)

AD-2	Input Output Data Definition (IODD)
AD-3	EUM/SEN3/DOC/19/1092968 Recommendations for Sentinel-3 OLCI Ocean Colour product validations in comparison with in situ measurements – Matchup Protocols (v8B, 25 Jan 2022)
AD-4	OMAPS Algorithm Theoretical Baseline Document for In-Water Algorithm selection, processing and blending
AD-5	OMAPS Algorithm Theoretical Baseline Document for AC-RR, Atmospheric Correction Round Robin
AD-6	OMAPS Algorithm Theoretical Baseline Document for Matchup Generation

Reference Documents and sources

ID	Document
RD-1	SACSO (2021) https://www.eumetsat.int/SACSO
RD-2	L2gen https://seadas.gsfc.nasa.gov/help-8.1.0/processors/ProcessL2gen.html
RD-3	POLYMER https://www.hygeos.com/polymer
RD-4	Sentinel-3 OLCI 3rd reprocessing collection report https://www.eumetsat.int/media/47794
RD-5	https://forum.earthdata.nasa.gov/viewtopic.php?f=7&t=1273&sid=1c0e8c49e6eb8f8a0206b9c9defd37b4 Sean Bailey's entry on Fri Apr 23, 2021, accessed on 30/11/2021
RD-6	https://www.eumetsat.int/ocean-colour-system-vicarious-calibration-tool
RD-7	OCDB (2021) https://ocdb.eumetsat.int/
RD-8	Valente, A., Sathyendranath, S., Brotas, V., Groom, S., Grant, M., Taberner, M., Antoine, D., Arnone, R., Balch, W. M., Barker, K., Barlow, R., Bélanger, S., Berthon, J.-F., Beşiktepe, Ş., Borsheim, Y., Bracher, A., Brando, V., Canuti, E., Chavez, F., Cianca, A., Claustre, H., Clementson, L., Crout, R., Frouin, R., García-Soto, C., Gibb, S. W., Gould, R., Hooker, S. B., Kahru, M., Kampel, M., Klein, H., Kratzer, S., Kudela, R., Ledesma, J., Loisel, H., Matrai, P., McKee, D., Mitchell, B. G., Moisan, T., Muller-Karger, F., O'Dowd, L., Ondrusek, M., Platt, T., Poulton, A. J., Repecaud, M., Schroeder, T., Smyth, T., Smythe-Wright, D., Sosik, H. M., Twardowski, M., Vellucci, V., Voss, K., Werdell, J., Wernand, M., Wright, S., and Zibordi, G.: A compilation of global bio-optical in situ data for ocean-colour satellite applications – version two, Earth Syst. Sci. Data, 11, 1037–1068, https://doi.org/10.5194/essd-11-1037-2019, 2019.
RD-9	MOBY "Satellite weighted" data https://www.star.nesdis.noaa.gov/socd/moby/gold/

RD-10	John E. O'Reilly, P. Jeremy Werdell, Chlorophyll algorithms for ocean color sensors - OC4, OC5 & OC6, Remote Sensing of Environment, Volume 229, 2019, Pages 32-47, ISSN 0034-4257, https://doi.org/10.1016/j.rse.2019.04.021.
RD-11	Hu, C., Lee, Z., & Franz, B. (2012). Chlorophyll a algorithms for oligotrophic oceans: A novel approach based on three-band reflectance difference. Journal of Geophysical Research, 117, C01011. https://doi.org/10.1029/2011JC007395
RD-12	Hu, C., Feng, L., Lee, Z., Franz, B. A., Bailey, S. W., Werdell, P. J., & Proctor, C. W. (2019). Improving satellite global chlorophyll a data products through algorithm refinement and data recovery. Journal of Geophysical Research: Oceans, 124, 1524–1543. https://doi.org/10.1029/2019JC014941
RD-13	IOCCG (2019) Report number 18: Uncertainties in Ocean Colour Remote sensing (chapter 4).

List of Acronyms

Acronym	Description
AC	Atmospheric Correction
ATBD	Algorithm Theoretical Baseline Document
BRDF	Bidirectional Reflectance Distribution Function
CBQ	Common Best Quality
CCI	Climate Change Initiative
ESA	European Space Agency
IBQ	Individual Best Quality
IdePIX	a multi-sensor pixel identification tool
IPF	Instrument Processing Facilities
IW	In-Water
MODIS	Moderate Resolution Imaging Spectroradiometer
MERIS	MEdium Resolution Imaging Spectrometer
NASA	National Aeronautics and Space Agency
OC-CCI	Ocean Colour Climate Change Initiative
OCDB	Copernicus Ocean Colour In-Situ Database
OLCI	Ocean and Land Colour Instrument

OMAPS	Ocean Colour Multi-Mission Algorithm Prototype System
POLYMER	POLYnomial based algorithm applied to MERIS (though it is now applicable to multiple sensors)
PVP	Product Validation Plan
RBD	Requirements Baseline Document
RMSD	Root Mean Square Difference
RR	Round Robin
SACSO	Spectral matching Atmospheric Correction for Sentinel Ocean colour measurements
SOW	Statement Of Work
SVC	System Vicarious Calibration
TOA	Top of atmosphere
VIIRS	Visible Infrared Imaging Radiometer Suite

1 Purpose

This Product Validation and Evaluation Report (PVER) presents OMAPS summary comparison between and assessment of four Ocean Colour processors applied to Sentinel-3 OLCI-A data. The goal of the document is also to demonstrate OMAPS capabilities in processor validations, inter-comparisons and ranking. The report describes the configuration of the Ocean Colour processors, the validation datasets used, including the in-situ database and the diagnostic granules, and the validation and ranking results. The four Ocean Colour processors applied in the current validations are the following:

- SACSO (Spectral matching Atmospheric Correction for Sentinel Ocean colour measurements) [RD-1]
- L2GEN [RD-2]
- POLYMER [RD-3]
- Collection-3 BASELINE operational IPF [RD-4].

The results presented in the report are not to be considered as valid algorithm validations. This is because

- very limited in situ and diagnostic datasets are used,
- the AC RR scoring method requires further development,
- the configurations of some processors are tentative, e.g. L2GEN,
- the in situ radiometric datasets can be further adjusted to OLCI nominal bands.

The results should be understood as a demonstration of OMAPS inter-comparison and validation capabilities rather than an assessment of processor performance.

2 Configurations compared

Below we will cover the settings that define each of the processor configurations compared within the report. These four configurations are referred to as 'default' configurations for POLYMER, SACSO, L2GEN and Collection-3 BASELINE atmospheric correction schemes. The BASELINE OLCI operational processor is not implemented within the OMAPS on-line processing scheme (and the code is not openly available) and the validations presented here are derived from L2 IPF standard products. The default POLYMER, SACSO and L2GEN configurations are located within the OMAPS code repository at 'online_processor/test/demo_configs/', in files 'polymer_demo.ini', 'sacso_demo.ini' and 'l2gen_demo.ini' respectively.

2.1 Product quality flags

Most ocean colour processing chains have some form of quality flagging available so that poor quality data is removed before delivery to users. In the cases presented below the recommended quality flags were applied in all cases. For the POLYMER, SACSO and L2GEN processor configurations this flagging involved a combination of IdePIX and atmospheric correction flags as given in Table 1 to Table 3. For the EUMETSAT Collection-3 BASELINE processor, the flagging applied was that recommended in the Sentinel-3 OLCI Collection-3 report [RD-4] as shown in Table 4.

Table 1: Quality Flags applied to POLYMER processing chain

POLYMER PROCESSOR: Flags that make pixel invalid (remove pixel if True)
Polymer:LAND
Polymer:CLOUD_BASE
Polymer:L1_INVALID
Polymer:NEGATIVE_BB
Polymer:OUT_OF_BOUNDS
Polymer:EXCEPTION
Polymer:THICK_AEROSOL
Polymer:HIGH_AIR_MASS
Polymer:EXTERNAL_MASK
Polymer:dust_mask
IDEPIX_INVALID
INDEPIX_LAND
IDEPIX_CLOUD
IDEPIX_SNOW_ICE
IDEPIX_CLOUD_BUFFER
IDEPIX_CLOUD_SHADOW

Table 2: Quality Flags applied to SACSO processing chain

SACSO PROCESSOR: Flags that make pixel invalid (remove pixel if True)
Sacso:LAND
Sacso:CLOUD_BASE
Sacso:L1_INVALID
Sacso:NEGATIVE_BB
Sacso:OUT_OF_BOUNDS
Sacso:EXCEPTION
Sacso:THICK_AEROSOL
Sacso:HIGH_AIR_MASS
Sacso:EXTERNAL_MASK
Sacso:dust_mask
IDEPIX_INVALID
INDEPIX_LAND
IDEPIX_CLOUD
IDEPIX_SNOW_ICE
IDEPIX_CLOUD_BUFFER
IDEPIX_CLOUD_SHADOW

Table 3: Quality Flags applied to L2GEN processing chain

L2GEN PROCESSOR: Flags that make pixel invalid (remove pixel if True) in default configuration

L2gen:ATMFAIL
L2gen:LAND
L2gen:HIGHGLINT
L2gen:HILT
L2gen:HISATZEN
L2gen:STRAYLIGHT
L2gen:CLDICE
L2gen:COCCOLITH
L2gen:HISOLZEN
L2gen:LOWLW
L2gen:CHLFAIL
L2gen:NAVWARN
L2gen:MAXAERITER
L2gen:CHLWARN
L2gen:ATMWARN
L2gen:NAVFAIL
IDEPIX_INVALID
INDEPIX_LAND
IDEPIX_CLOUD
IDEPIX_SNOW_ICE
IDEPIX_CLOUD_BUFFER
IDEPIX_CLOUD_SHADOW

Table 4: Quality Flags applied to the OLCI BASELINE operational processing data

BASELINE PROCESSING: Flags that make pixel invalid (remove pixel if True)
CLOUD
CLOUD_AMBIGUOUS
CLOUD_MARGIN
INVALID
COSMETIC
SATURATED
SUSPECT
HISOLZEN
HIGHGLINT
SNOW_ICE
AC_FAIL
WHITECAPS
ADJAC
RWNEG_O2

RWNEG_O3
RWNEG_O4
RWNEG_O5
RWNEG_O6
RWNEG_O7
RWNEG_O8
OC4ME_FAIL
NOT WATER
NOT INLAND_WATER

2.2 SVC Gains

The following SVC gains (processor specific) were used in the processing of the OLCI 3A granules for the validation purposes:

Band / Processor	400	412	443	490	510	560	620	665	673	681
POLYMER	1	0.996623	0.998205	0.984052	0.987536	0.987782	0.990961	0.987454	1	1
SACSO	0.981810	0.979885	0.981551	0.972591	0.976808	0.977369	0.984240	0.985055	0.983324	0.985154
L2GEN	0.959728	0.972313	0.97156	0.96916	0.97636	0.97951	0.97705	0.975398	0.97344	0.97597
BASELINE	0.97546	0.97406	0.97492	0.9689	0.97184	0.97571	0.98001	0.97834	0.9786	0.97908

Band / Processor	709	754	761	764	767	779	865	885	900	940	1020
POLYMER	1	1	1	1	1	1	1	1	1	1	1
SACSO	1.014972	0.995357	1	1	1	1.000770	1.004331	1.011602	1	1	0.933846
L2GEN	1.00562	0.9829	1	1	1	0.98989	1	1.01815	1	1	1
BASELINE	0.98013	0.98552	1	1	1	0.98772	0.986	0.98657	1	1	0.91316

The gains for POLYMER and SACSO were generated using the OMAPS SVC off-line module during the projects testing and development stage. The L2GEN gains were provided by NASA as documented at the Earth Data Forum [RD-5]. The BASELINE gains were derived by EUMETSAT for Collection-3 processing and documented in RD-4 and, with further detail, in RD-6.

3 Validation Methods

The primary measure of ocean colour Level-2 processor quality, its configuration and module performance is based on combined metrics derived from validations with co-incident in-situ data. It is essential that performance is assessed in a multi-metric manner as the use of single statistics such as correlation coefficient or bias implies a very narrow definition of ‘best’ performance. Additional information on processor performance can also be gained from the analysis of diagnostic granules, timeseries analysis and inter-sensor comparisons as described below. Additional details on the methods for intercomparison of processor configuration performance (for both atmospheric and in-water modules) are covered in the relevant round robin ATBDs, so will not be described here.

3.1 Match-ups against in-situ

Matchups were made against in-situ data available in the OCDB [RD-7] and used to generate a match-up dataset for validation of the generated OLCI products for multiple configurations. In this case four processing configurations will be compared, the BASELINE processing available from EUMETSAT and the default POLYMER, SACSO and L2GEN processor configurations. The matchup procedure within OMAPS is covered in detail in the dedicated ATBD [AD-6], but we have summarised the key matchup criteria involved below:

1) Granules that overlap with in-situ measurements in time ±6 hours (EUMETSAT state 3 hours but OC-CCI have a longer window) and space are identified.
2) The granules are processed through a given processor configuration.
3) A macro-pixel of 5 x 5 pixels around the single closest pixel is extracted and summary statistics are produced (median, standard deviation, number of valid pixels within the macro-pixel).
4) Matchups are filtered based on a defined set of criteria (threshold values for factors such as solar and viewing zenith angles, product value, number of valid retrievals, coefficient of variation within macro-pixel) for which the defaults in the config file are:
    a. sza_max_valid = 70.0
    b. vza_max_valid = 60.0
    c. min_valid_pixels = 40.0
    d. variance_factor = 1.5
    e. coeff_of_variation_thresh = 0.5
5) Duplicated measurements (from drift stations or other sampling anomalies) are removed.

OMAPS metrics are close to but not quite identical to the recommendations of the EUMETSAT matchup protocol (AD-3), and are also different from the OC-CCI metrics for product validation. The primary difference from EUMETSAT recommendations are that the min_valid_pixels are 40.0% not 50.0% and the coeff_of_variation_thresh should be reduced to 0.2 and used for the 560 nm band according to AD-3. By default, the OMAPS metrics consist of median (absolute) difference, MdD and MdAD, and median (absolute) percentage difference, MdPD and MdAPD, and spectral measures chi-square value and spectal angle mapper (SAM). The OC-CCI metrics like bias, RMSD, regression parameters and number of matchups are still available and can be selected. The scoring system is multi-metric in nature and combines all metrics that the user requests in the config file.

For the AC-RR exercise in section 4.1, we have used in situ measurements from MOBY (Marine Optical BuoY). MOBY is located in highly oligotrophic waters in the middle of the Pacific Ocean off Hawaii.

For the time series comparison in section 4.4 (and only for this section) we have used datasets of remote sensing reflectance from MOBY and the Hawaii Ocean Time-series (HOTs) for chlorophyll-a that were not generated using the OCDB matchup module (as we were not matching to satellite granules). Instead, the data for each of the two sites was retrieved from the OC-CCI in-situ database (which itself has contributed to the OCDB) and used for comparison to OLCI macro-pixels (using the above matchup filtering criteria and statistics) over the measurement locations. Over 700 granules with some coverage of the Hawaii area were found.

The MOBY data is "satellite-weighted" data for OLCI A/B mean spectral response functions (i.e. not the MOBY hyperspectral data) and documentation for this is provided at RD-8 and RD-9. It is noted that, with OLCI's L2 smile correction, the centre wavelengths of the mean spectral responses are slightly shifted to OLCI nominal bands reported for L2 products. However, the difference of ~0.5 nm is within the validation protocol range and perceived to be negligible for these validations. The band shifting of MOBY data within the OCDB will be performed in the future so that matchups are at nominal wavelengths.

3.2 Atmospheric Round Robin intercomparison

For this report we have conducted an atmospheric correction (AC) round robin (RR) exercise, for which the configuration and scoring metrics and data used are described below, with results presented in section 4.1.

3.2.1 Atmospheric round robin configuration

The RR intercomparison was performed between four candidate processor configurations. These configurations are:

1) The default POLYMER processor configuration used by the OMAPS in-line processing
2) The default SACSO processor configuration used by the OMAPS in-line processing
3) The default L2GEN processor configuration used by the OMAPS in-line processing
4) The EUMETSAT Collection-3 BASELINE processor

As mentioned above, the BASELINE processor is not executed within the OMAPS in-line processing infrastructure and only output standard Level-2 products are used in the RR. The BASELINE processing configuration is available from the EUMETSAT reports associated with the Collection-3 release [RD-4]. The default POLYMER, SACSO and L2GEN configurations are located within the OMAPS code repository at 'online_processor/test/demo_configs/', in files 'polymer_demo.ini', 'sacso_demo.ini' and 'l2gen_demo.ini' respectively.

The AC RR module allows analysis in both common best quality (CBQ) and individual best quality (IBQ) modes. In-situ data for this assessment was sourced from the EUMETSAT Ocean Colour Database (OCDB) [RD-7]. The assessment follows the algorithm of the AC-RR described in the RR-AC ATBD [AD-5]. In both cases the assessment looked at seven visible wavebands (412, 443, 490, 510, 560, 620 and 665 nm). The exact set of metrics used for assessment is defined in the AC-RR configuration file (e.g 'rrac_config_IPF_SACSO_POLYMER_L2GEN_CBQ.ini') and in this case was set to use median absolute difference (MdAD), median difference (MdD), median absolute percent difference (MdAPD), median percent difference (MdPD), spectral angle SAM and $\chi 2$.

3.2.2 AC RR scoring

The results of all statistical properties, assessed per wavelength, and the spectral tests need to be distilled into a single value for easy comparison. Simple conversions are made using the statistical value itself and its standard error or 95% confidence interval. The statistical values are converted in the following way, so that the ranking of the values becomes easy: the best statistical value is closest to zero and receives the highest score. Full details of the AC RR scoring methodology are given in the corresponding ATBD [AD-5], with a summary given below.

The scoring method used in this report is tentative and it requires further development. The scoring results should be considered with caution. The variability in ocean colour radiometry often causes large ranges of confidence intervals, here used as a 95%. The large confidence intervals then result in the same score values awarded to algorithms with vastly different statistical results and, otherwise, statistically superior algorithms are not identified.

In the scoring system, all statistical parameters derived on absolute differences (MAD, MAPD or the respective medians of absolute differences, MdAD and MdAPD) are used as they are, while statistics which yield both positive and negative values (MD and MPD, or the respective median-based scores MdD and MdPD) are interpreted as absolute values (e.g. $|\mathrm{M}d\mathrm{D}|$ and |MdPD|) so that a simple ranking of the statistics can be applied. The best algorithm is the one for which the statistical property in question has the value closest to zero (after taking the absolute if necessary) e.g. lowest absolute bias or dispersion, lowest values in spectral angle or $\chi 2$. This 'best' algorithm receives a score of 2 points for the parameter and variable combination in question. Then, the other algorithms are scored as follows:

- If the statistical value of another algorithm for this band falls within the confidence interval of the best, this algorithm is not significantly different from the best and receives 2 points as well.
- If the value of another algorithm lies outside the confidence interval of the best but their confidence intervals overlap, this algorithm receives 1 point.
- If the confidence interval of an algorithm does not overlap with the best algorithm, this algorithm receives 0 points.

For each wavelength, the statistics of the ACs are compared and turned into scores independently (this applies to the single value statistics like MdD etc.) and an example of this is explained below for a single wavelength.

Table 6 shows the conversion of performance statistics from Table 5 into scores for the band at 412nm. The starting point are the statistics for one band for all ACs in the comparison. The best value for median absolute difference (MdAD) is found for the AC sacso_1.0, which then defines the upper threshold of statistically similar results; the upper threshold is 0.001047+0.000886=0.001933. All other MdAD values from other ACs are below this threshold, so all of them are statistically similar and each one receives 2 points in the scoring (Table 6). The same reasoning applies to the MdAPD values.

The values of MdD and MdPD are converted into their absolutes, before the minimum value is chosen as the best. For MdD, the AC sacso_1.0 has the lowest value, which then defines the upper limit of statistical similarity as 0.00019+0.000886=0.001077. The MdD of polymer_4.17 and l2gen lie within this range, so the three ACs receive 2 points. The values of ipf_Collection-3 with 0.001528 is larger than this upper limit of the confidence interval of the best result. Though it does not fall within the confidence interval of the best model, the lower limit of the ipf_Collection-3 confidence interval is 0.001528-0.000867=0.000661, which does fall below the upper limit of the best AC confidence interval. Therefore, ipf_Collection-3 receives 1 point in the scoring as the confidence intervals overlap. The same reasoning applies to MdPD.

Tables 5 and 6 clearly show that large confidence intervals result in algorithms to be assigned the same score even though their median statistics are significantly different and, otherwise, there could be clear winners. For example, sacso_1.0 MdPD is 1.128697, while polymer_4.17 and l2gen_9.5.1 MdPD are more than 5 times higher, nevertheless all three algorithms obtain the same score. The confidence interval approach should be further investigated.

In a second step, each set of scores for a single statistical parameter is scaled by their column sum, so that each statistical parameter gains equal weight when the scores are summed up for all algorithms (Table 7). The sum of scores is scaled again, so that the column sum is equal to the number of ACs in the study (i.e. no change in this case, as the number of statistical parameters equals the number of ACs).

After calculating the scores for each band in the comparison, the scaled sums of scores per band are combined with the scores for the spectral statistical measurements SAM and $\chi 2$ (Figure 7). An example is given, how the $\chi 2$ values are directly translated into scores (Table 8). In a first step, the $\chi 2$ values are normalised by dividing by their sum, so that values are now in the range between {0,1} (see column 'chi_sqr_norm'). The smallest normalised $\chi 2$ value is the best, so values are simply converted by calculating the difference score=1- chi_sqr_norm. The score is scaled again, so that its sum equals the number of ACs. In this fashion, the spectral statistics receive as much weight in the total sum of scores of each band.

Table 5: Statistics for band 412nm (MdAD, MdAPD, MdD, MdPD) and the width of the confidence interval (MdAD95, MdAPD95, MdD95, MdPD95).

processor	MdAD	MdAD95	MdAPD	MdAPD95	MdD	MdD95	MdPD	MdPD95
polymer_4.17	0.001306	0.000889	10.29495	5.780118	0.000846	0.000889	6.020694	5.780118
sacso_1.0	0.001047	0.000886	6.330624	6.346729	0.00019	0.000886	1.128697	6.346729
ipf_Collection-3	0.001528	0.000867	10.84188	4.44587	0.001528	0.000867	10.84188	4.44587
l2gen_9.5.1-V2021.2	0.001566	0.001129	9.580587	8.263558	-0.00101	0.001129	-6.46565	8.263558

Table 6: Statistics converted into scores according to the general scheme. Intermediate step in scoring for band 412nm.

processor	MdAD	MdAPD	MdD	MdPD
polymer_4.17	2	2	2	2
sacso_1.0	2	2	2	2
ipf_Collection-3	2	2	1	1
l2gen_9.5.1-V2021.2	2	2	2	2

Table 7: Each set of scores for one statistical parameter is scaled by their column sum. The column Sum is scaled afterwards, so that its sum equals the number of ACs in the study. These are the scaled scores for the statistics of band 412nm in the example.

processor	MdAD	MdAPD	MdD	MdPD	Sum
polymer_4.17	0.25	0.25	0.2857	0.2857	1.07

sacso_1.0	0.25	0.25	0.2857	0.2857	1.07
ipf_Collection-3	0.25	0.25	0.1429	0.1429	0.79
l2gen_9.5.1-V2021.2	0.25	0.25	0.2857	0.2857	1.07

Table 8: Conversion of chi-square values into scores (normalising the chi-square values, transformation, and final scaling).

processor	chi_square	chi_sqr norm	score	score_scaled
polymer_4.17	0.357552	0.253335	0.746665	1.00
sacso_1.0	0.25506	0.180716	0.819284	1.09
ipf_Collection-3	0.39959	0.283119	0.716881	0.96
l2gen_9.5.1-V2021.2	0.399181	0.28283	0.71717	0.96

Processor	varname	MAD	MD	MAPD	MPD	N
polymer_4.17	rrs_412	9.65e-04	7.68e-04	6.9	5.1	92
polymer_4.17	rrs_443	5.84e-04	3.52e-04	5.7	3.0	92
polymer_4.17	rrs_490	2.71e-04	-1.30e-06	4.6	-0.4	92
polymer_4.17	rrs_510	3.18e-04	2.94e-04	9.0	8.2	92
polymer_4.17	rrs_560	6.54e-05	-1.26e-05	5.4	-1.4	92
polymer_4.17	rrs_620	5.73e-05	-5.73e-05	33.8	-33.8	92
polymer_4.17	rrs_665	5.41e-05	-5.36e-05	62.5	-62.1	92
sacso_1.0	rrs_412	9.94e-04	-3.74e-04	7.5	-3.5	90
sacso_1.0	rrs_443	6.48e-04	-2.87e-04	6.8	-3.5	90
sacso_1.0	rrs_490	7.46e-04	-7.43e-04	12.9	-12.8	90
sacso_1.0	rrs_510	1.95e-04	-1.00e-04	5.9	-3.3	90
sacso_1.0	rrs_560	2.65e-04	-2.65e-04	21.4	-21.4	90
sacso_1.0	rrs_620	1.03e-04	-1.02e-04	60.5	-59.8	90
sacso_1.0	rrs_665	1.06e-04	-1.06e-04	121.5	-121.4	90
ipf_Collection_3	rrs_412	1.39e-03	6.77e-04	10.2	4.5	66
ipf_Collection_3	rrs_443	8.74e-04	3.06e-04	8.8	2.6	66
ipf_Collection_3	rrs_490	4.66e-04	7.12e-05	7.8	0.9	66
ipf_Collection_3	rrs_510	4.93e-04	3.38e-04	14.3	9.6	66
ipf_Collection_3	rrs_560	1.79e-04	2.39e-05	14.3	1.7	66
ipf_Collection_3	rrs_620	9.78e-05	-2.92e-05	57.6	-18.8	66
ipf_Collection_3	rrs_665	5.81e-05	-2.07e-05	67.8	-27.2	66
l2gen_9.5.1-V2021.2	rrs_412	1.12e-03	4.39e-04	8.4	3.1	94
l2gen_9.5.1-V2021.2	rrs_443	7.43e-04	-2.99e-04	7.7	-3.3	94
l2gen_9.5.1-V2021.2	rrs_490	6.34e-04	-5.93e-04	10.8	-10.1	94
l2gen_9.5.1-V2021.2	rrs_510	3.25e-04	-2.28e-04	9.5	-6.8	94
l2gen_9.5.1-V2021.2	rrs_560	4.10e-04	-4.05e-04	32.7	-32.3	94
l2gen_9.5.1-V2021.2	rrs_620	1.69e-04	-1.5e-04	98.7	-87.7	94
l2gen_9.5.1-V2021.2	rrs_665	9.5e-05	-8.90e-05	110.1	-104.0	94

Figure 1: Example of automatically generated table in a LaTeX document of statistics for the single representation of the matchup data. The "Processor" column lists the ACs with their type and version, the variable name ("varname") lists the name of the bands. The statistical values are defined above, the number of valid macropixels per band ("N") is included as additional information.

Processor	CHI2	SAM
polymer_4.17	0.2027	0.0213
sacso_1.0	0.2275	0.0335
ipf_Collection_3	0.5848	0.0188
l2gen_9.5.1-V2021.2	0.8707	0.0525

Figure 2: Example of automatically generated table in a LaTeX document of spectral statistics for the single representation of the matchup data.

Processor	rrs_412	rrs_443	rrs_490	rrs_510	rrs_560	rrs_620	rrs_665	CHI2	SAM	Total.Scores
polymer_4.17	1.04	1.17	2.33	0.0	3.0	2.67	1.0	1.19	1.11	13.5
sacso_1.0	1.2	1.17	0.0	3.33	0.0	0.0	0.0	1.17	0.98	7.85
ipf_Collection_3	0.73	0.83	1.67	0.0	1.0	1.33	3.0	0.92	1.13	10.61
l2gen_9.5.1-V2021.2	1.04	0.83	0.0	0.67	0.0	0.0	0.0	0.72	0.78	4.03

Figure 3: Example of automatically generated table in a LaTeX document of the scores assigned to the statistics for the single representation of the matchup data.

In this particular analysis this means a maximum score would be 36 (=9x4), in the unlikely situation where a single atmospheric correction scheme outperforms all others in all metrics at all bands (with no confidence interval overlap).

3.2.3 AC RR data used

As this is a demonstration case that makes use of a limited in situ and diagnostic datasets (those matched through the combination of predefined diagnostic granules and the in-situ data available within the OCDB) the matchups available for the demonstration AC-RR are not global in coverage. A map of the matchup locations for each of the processor configurations is shown in Figure 1. The data are almost exclusively what we might consider typical 'open-ocean' spectra (Figure 2) sourced from the clear waters near Hawaii from the MOBY instrument [RD-9]. This means that nothing approaching case 2 waters is included in this demonstration AC-RR.

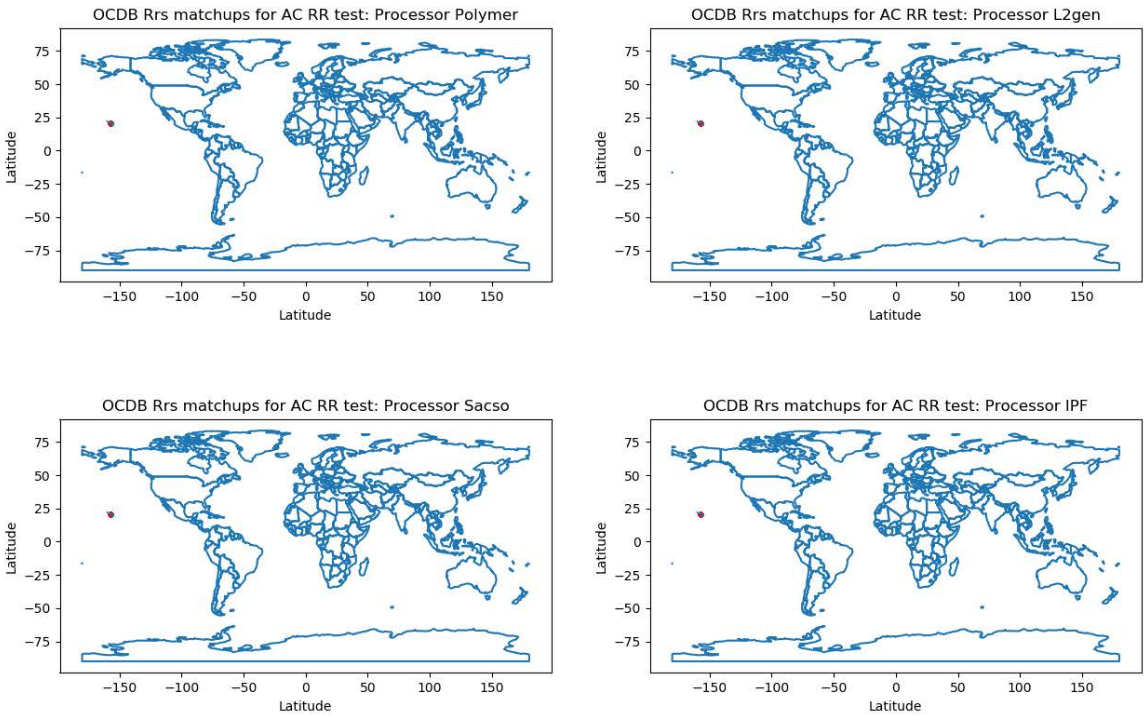


Figure 1: OCDB in-situ matchup locations against diagnostic granules. Locations of matchups are shown in red and are all clustered close to Hawaii, reflecting that the matchups are almost entirely from the MOBY site.

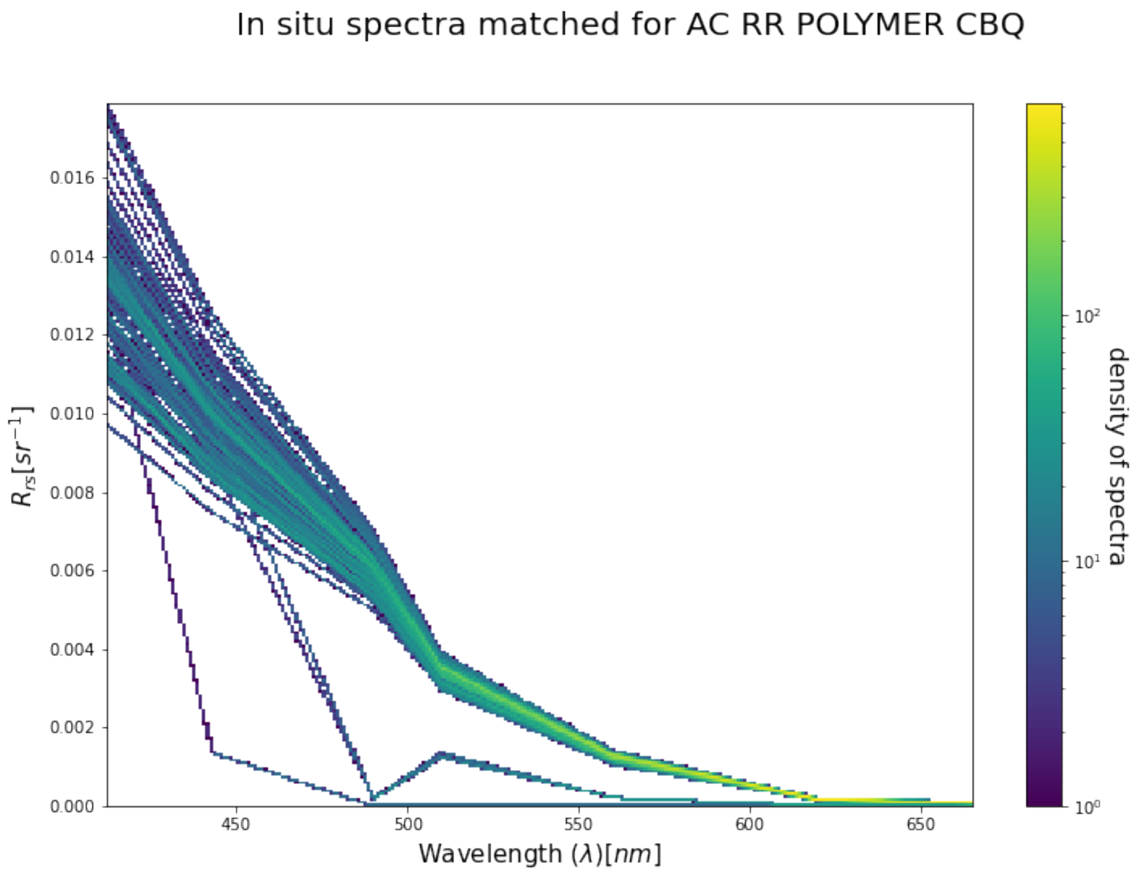


Figure 2: In situ spectra that were matched to POLYMER data as part of the AC RR CBQ analysis. As can be seen, almost all the spectra show a strong blue signal, typical of open ocean (case 1) waters.

3.3 In-water Round robin

For this report we have conducted an In-water algorithm (IW) round robin (RR) exercise, for which the configuration and scoring metrics and datasets are described below, with results presented in section 4.2.

3.3.1 IW RR configuration

The following configuration options were set, in the config file, for the IW RR results presented:

- match_file_chl: MDB_S3A_OLCI_L2_OCDB_IW_RR_L2_FULLTEST_CHLA.csv (Generated from OCDB)
- chl_algorithms: chlor_oc2,chlor_oc3,chlor_oc4,chlor_ocx,chlor_oci,chlor_oci2
- Sensor: OLCI
- RR_type: chl
- insitu_column_string: chla_fluor
- bootstraps: 100
- LogNormVar: True
- SplitByWaterClass: True

- WeightedByWaterClass: False
- threshold_memb: 0.3
- Split_Criteria: dominant

As shown in the configuration above, the IW RR was performed using six candidate empirical chlorophyll-a algorithms. The OC2, OC3, OC4, and OCX algorithms all follow a common formulation with the OCX being effectively equivalent to the OC4 for OLCI single sensor implementation (the OCX swaps parameterisations and number of bands depending on sensor). Chlorophyll (C) is estimated according to Eq. 3. The coefficients a0 to a4 are sensor dependant. X is the log of the maximum band ratio between 2, 3 or 4 bands (OC2, OC3 and OC4 respectively) with the general form given in Eq.4, where λ_{green} is the closest wavelength to 555nm.

$$C = 10^{\left(a_0 + \sum_{i=1}^{4} a_i X^i\right)} \tag{3}$$

$$X = log_{10}\left[\frac{\max\left(R_{rs}(\lambda_1, \lambda_2 \ldots)\right)}{R_{rs}\left(\lambda_{green}\right)}\right] \tag{4}$$

The OCI and OCI2 algorithms are identical to the OC4 algorithms only differ from the OC4 algorithm below a predefined estimated chlorophyll-a concentration. At these lower chlorophyll-a concentrations C is estimated according to

$$C = 10^{a_0 + a_1 CI}$$

where,

$$CI = R_{rs}(\lambda_2) - \frac{\left(R_{rs}(\lambda_1) + \lambda_2 - \lambda_1\right)}{(\lambda_3 - \lambda_1) * \left(R_{rs}(\lambda_3) - R_{rs}(\lambda_1)\right)},$$

with λ_1, λ_2 and λ_3 being 412, 443 and 665 in the case of OLCI.

Not all these empirical algorithms had been explicitly parameterised for OLCI. When parameters tuned to OLCI data were not available, we have used the parameterisation derived for MERIS, but only for testing the processing chain.

Algorithm	Source	a0	a1	a2	a3	a4
OC2	MERIS proxy https://oceancolor.gsfc.nasa.gov/atbd/chlor_a/	0.2389	-1.9369	1.7627	-3.0777	-0.1054
OC3	MERIS proxy https://oceancolor.gsfc.nasa.gov/atbd/chlor_a/	0.2521	-2.2146	1.5193	-0.7702	-0.4291
OC4	OLCI specific parameterisation RD-9	0.42540	-3.21679	2.86907	-0.62628	-1.09333
OCI	OLCI specific parameterisation RD-10	-0.5379	180.9642			
OCI2	OLCI specific parameterisation RD-11	-0.4287	230.47			

3.3.2 IW RR scoring

The full details of the IW RR scoring are documented in the relevant ATBD [AD-4] but we have provided a summary here for better understanding and interpretation of the scores presented. As with the AC RR, the assessment is performed using matchups between the satellite products and an in-situ database of variable measurements.

The IW RR scoring provides a total score for each algorithm which is the sum of scores per performance metrics in the comparison. Each algorithm can score 0 to 2 points per metric in the assessment with a higher score meaning better performance. The scores are then normalised to the highest total score so that all models have a score between 0 and 1. This assessment is performed in a bootstrapped manner so that a suite of scores are generated with a mean and percentile range (2.5-97.5 percentile) stored and plotted as output. The set of statistics used in the results presented here were correlation coefficient (r), unbiased RSME (Δ), bias (δ), slope of type 2 regression (S), intercept of type 2 regression (I) and number of retrievals (n). RMSE (ψ) is also shown in Figure 3 but we did not use it as this is effectively a compound metric, reproducing information that we are already testing with the bias and unbiased RMSE.

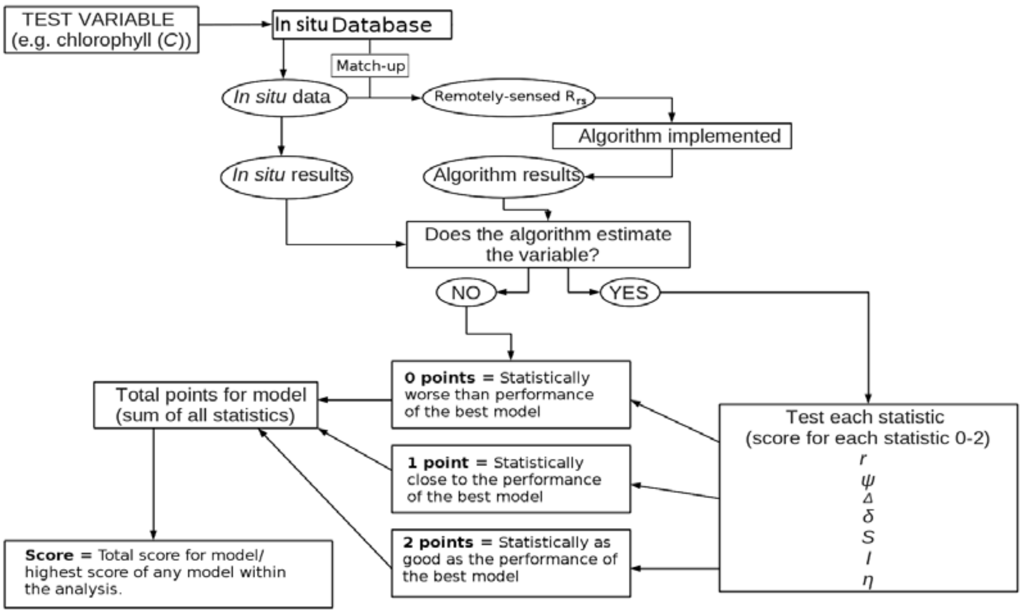


Figure 3: Flow chart of multi-metric scoring approach for In-water algorithm comparison

3.3.3 IW RR data used

As this is a demonstration case that makes use of a limited in situ and diagnostic datasets (those matched through the combination of predefined diagnostic granules and the in-situ data available within the OCDB) the matchups available for the demonstration IW-RR are not global in coverage. A map of the matchup locations for each of the processor configurations is shown in Figure 4. Looking at the histogram of in-situ chlorophyll-a measurements that were matched for analysis (Figure 17) the matchups contain data from oligotrophic to productive waters (chl-a concentrations <0.1 and >5.0) with a peak between 0.3 and 1.0 mg m^{-3}.

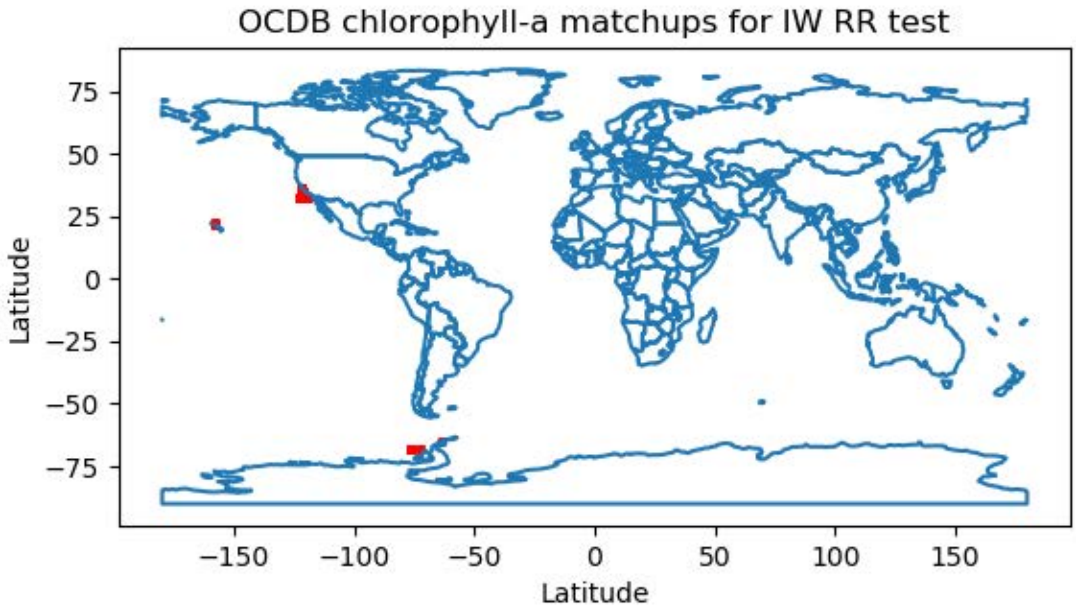


Figure 4: OCDB in-situ matchup locations against diagnostic granules. Locations of matchups are shown in red and are clustered around three regions.

HPLC is considered the most accurate measurement of chlorophyll-a (compared to spectrophotometric and fluorometric methods) but fluorometric measurements are much more common. As this study was limited by the number of matchups available, we chose to use fluorometric chlorophyll-a measurements (of extracted, *in vitro*, chlorophyll-a) to try to maximise the number of available matchup points. Though generally considered to have higher uncertainties than HPLC, laboratory measurements based on fluorometry or spectrophotometry performed after pigment extraction can achieve comparable performance for chlorophyll (though the presence of chlorophyll-*b* may introduce a bias between fluorometry and HPLC determinations) [RD-13]. Fluorescence measurements taken *in vivo* (not used here) are subject to much higher biases and should be interpreted with great care.

4 Results

As described in Introduction, the results presented in this report should be understood as a demonstration of OMAPS inter-comparison and validation capabilities and not as an assessment of processor performance. This is because:

- very limited in situ and diagnostic datasets are used,
- the scoring method requires further development,
- the configurations of some processors are tentative, e.g. L2GEN,
- the in situ radiometric datasets can be further adjusted to OLCI nominal bands.

4.1 AC RR results

We will consider the CBQ case first, where only matchups available to all the processor configurations are used in the assessment. Starting at the final output of the RR, the total score performance distributions generated by the bootstrapping analysis are shown in Figure 5. Refer to AC-RR ATBD for the description of bootstrapping.

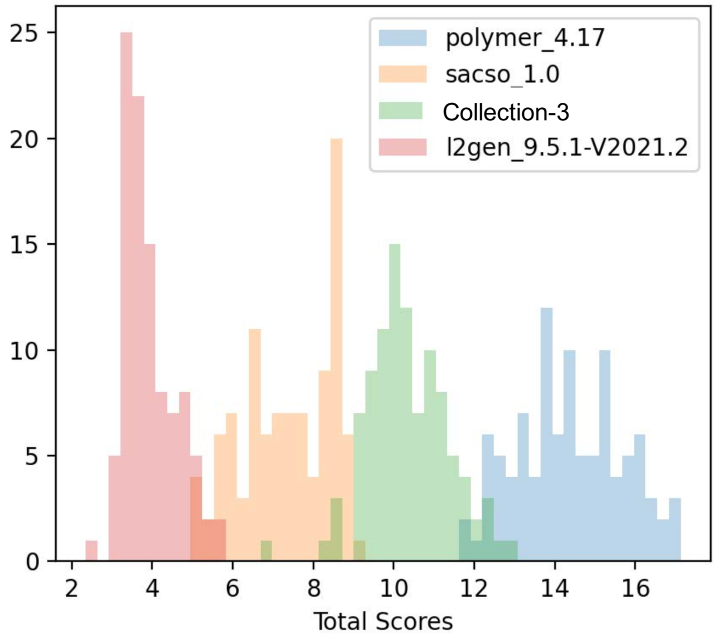
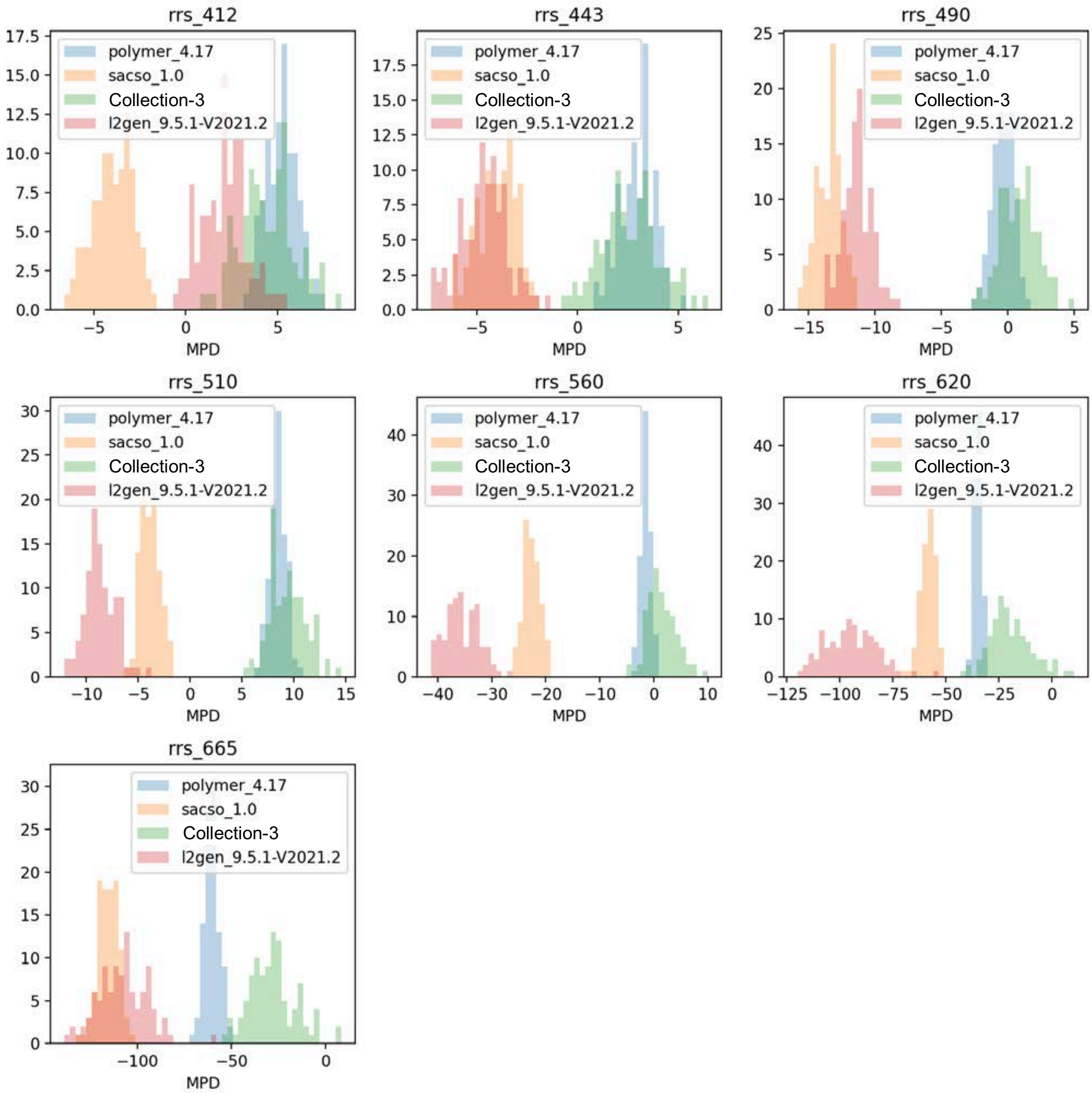


Figure 5: Example atmospheric CBQ RR total scores in a comparison of L2GEN, POLYMER, SACSO and BASELINE processors. These distributions of scores are derived by bootstrapping,

Example interpretation of AC-RR procedure exercises is described in the following text and, as explained, it should not be understood as valid processor validation results.

The exercise round robin seems to suggest that POLYMER performs best for remote-sensing reflectance (Rrs) retrieval, with the BASELINE processor coming second, SACSO third and L2GEN being the least capable. This result depends on multiple performance metrics so it is important to evaluate the individual results in a more detail. Looking at the median percent difference (MdPD) across all bands (as shown in Figure 6) it might seem surprising that POLYMER has come out on top in this exercise total assessment, as POLYMER is not the best performing candidate (i.e. closest to zero) in all cases. It is worth noting (as documented in AD-5) that here for 'difference' statistics such as (shown in Figure 6) the difference between in-situ and satellite estimates is calculated as $x_{in\_situ} - x_{satellite}$, meaning that a consistent overestimate by the satellite would be expressed as a negative difference value. POLYMER often has similar or worse performance for MdPD than the BASELINE processing and we even see L2GEN preforming best (measured by MdPD) for the 412 band. However, if we look at the median absolute percent difference (MdAPD), as shown in Figure 7, it becomes clear that the errors in the POLYMER and BASELINE processors have a different structure. The MdAPD for the BASELINE is worse than for POLYMER for almost all bands. This likely results from the fact that the strong performance of the BASELINE processor in the MdPD statistic may be resulting from the averaging of larger positive and negative percent difference errors, that are symmetrically distributed around zero, for the BASELINE processor more than for POLYMER. If the sign of the error is ignored and only the mean magnitude (percent) of the error is considered, then POLYMER exhibits a smaller mean error. This demonstrates the importance of a multi-metric approach and that the choice of different single statistics can lead to a different 'optimal' algorithm result.

The CBQ assessment may be considered a 'fair' assessment of the candidate processors as it assesses the performance based on a common set of measurements. However, this may do some processors a disservice if they are able to produce good results under conditions that other algorithms cannot handle. In order to assess the performance, allowing the processors to use all points that they process for matchup, we can look at the IBQ results. In the IBQ comparison increased coverage from some configurations may degrade the quality of the overall product and decrease the performance metrics (see the AC-RR ATBD for more details



on the scoring system) so increased coverage without a decrease in product quality can improve a processors score in the IBQ relative to the CBQ. It is important to consider both cases for a thorough analysis of performance.

Figure 6: Example atmospheric CBQ RR performance for the mean percent difference (MPD) metric.

The total scores for the IBQ comparison (Figure 8) are similar to those of the CBQ, and POLYMER still leads the field in this example AC-RR. Looking at the number of matchups per processor (Figure 9) we can see that in the IBQ comparison the BASELINE processing seems to provide a lower number of matchups than all the other processors (which are relatively similar in the total points provided).

Full AC-RR results for this limited AC-RR exercise (both figures and statistical tables) are available in the annex and can be regenerated (allowing for small changes in exact values due to random bootstrapping) by following the instructions in the user guide. However, to provide more context for the scores shown in figures 2-5 we have also included the summary matchup scatter plots (CBQ) for each of the 4 processor configurations in Figure 10 to Figure 13. A summary table of the scores for both the IBQ and CBQ analysis across the set of variables is show in Figure 14 and Figure 15. The overall 'winner' in this limited analysis appears to be POLYMER, scoring highest in 5 out of 9 categories for both the CBQ and IBQ assessments.

Obviously, these results are only a demonstration of OMAPS validation capabilities, as explained before, and should not be considered as a true algorithm performance. Particularly, these results are limited by a low number of matchups compared to similar analyses performed for historical sensors. This means that the conditions beyond oligotrophic waters around Hawaii are not sampled and the processor performance may be significantly different in the other water types. It is of note that the CBQ and IBQ results are relatively consistent and that different processors seem to have certain bands for which they are particularly strong or weak. The 510 nm band seems to be particularly interesting as the MOBY in situ measurements used in the validation depart by about 0.4 nm from the nominal OLCI wavelength at 510 nm and both POLYMER and the BASELINE processor consistently score 0 points for this variable. Given the results presented here for the atmospheric round robin exercise we will fix the atmospheric processor as POLYMER for the in-water round robin exercise.

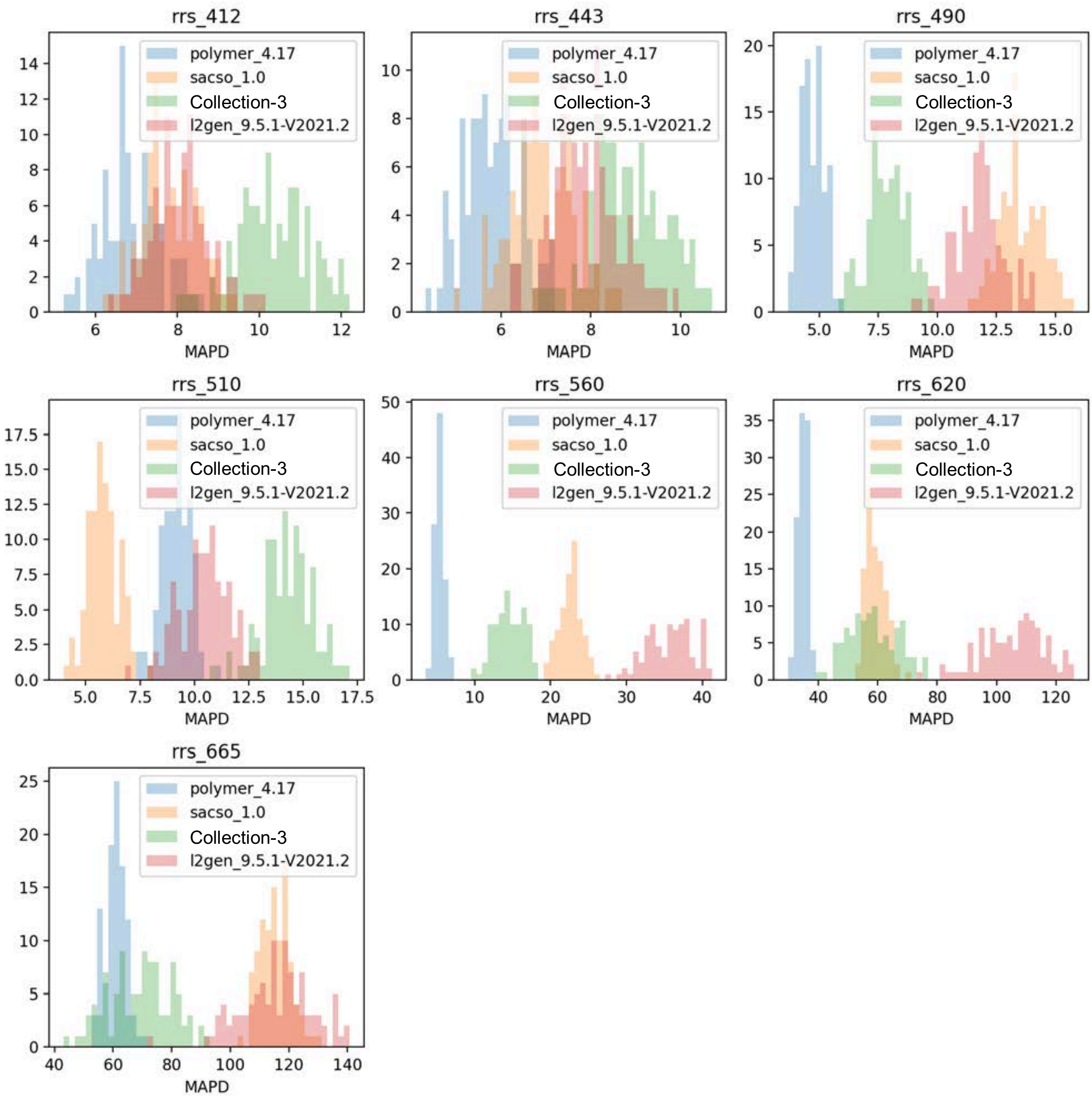


Figure 7: Example atmospheric CBQ RR performance for the mean absolute percent difference (MAPD) metric.

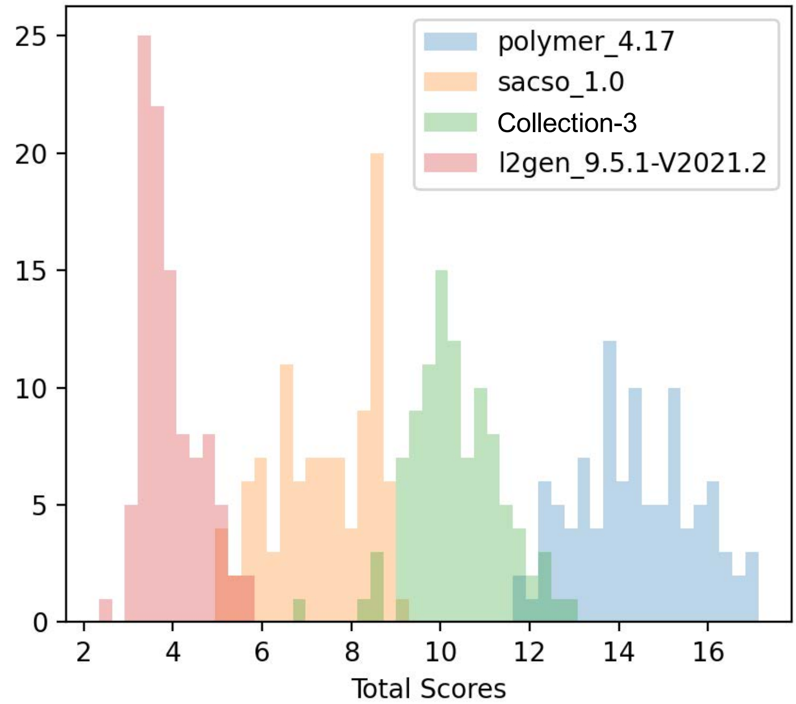


Figure 8:Example IBQ Atmospheric Round Robin Total scores comparing POLYMER, SACSO, L2GEN and BASELINE processors.

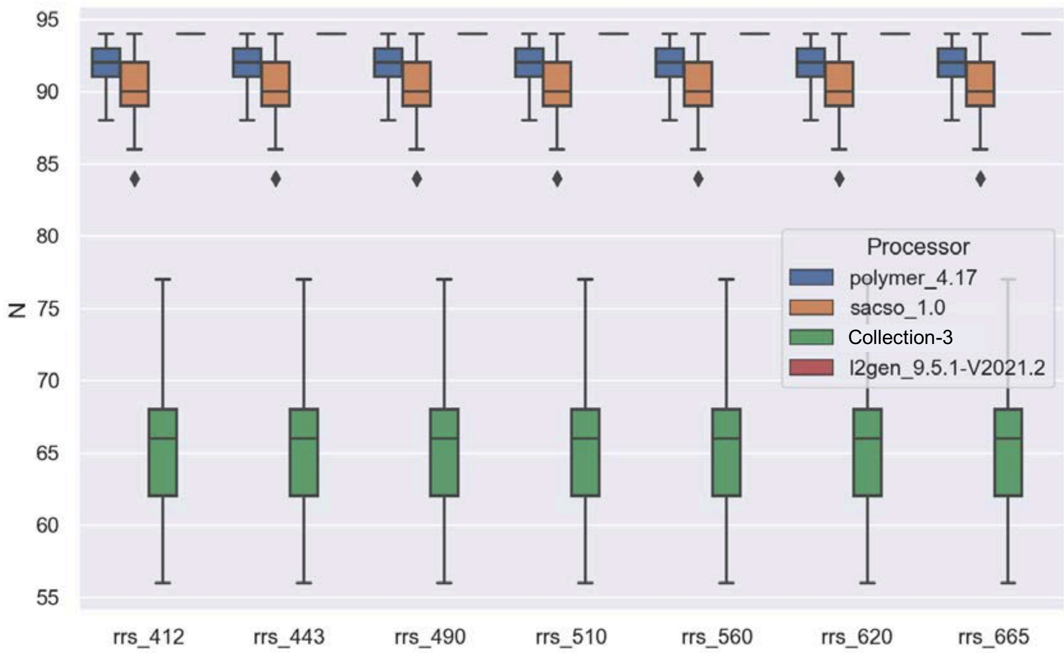


Figure 9: Example IBQ AC RR number of matchups available per processor. Box and whisker plots show variance across bootstrapping exercise, diamond points are classed as outliers (as determined by python seaborn.boxplot function). The l2gen value is non-variable across the bootstraps and so appears as a single horizontal line in this plot.

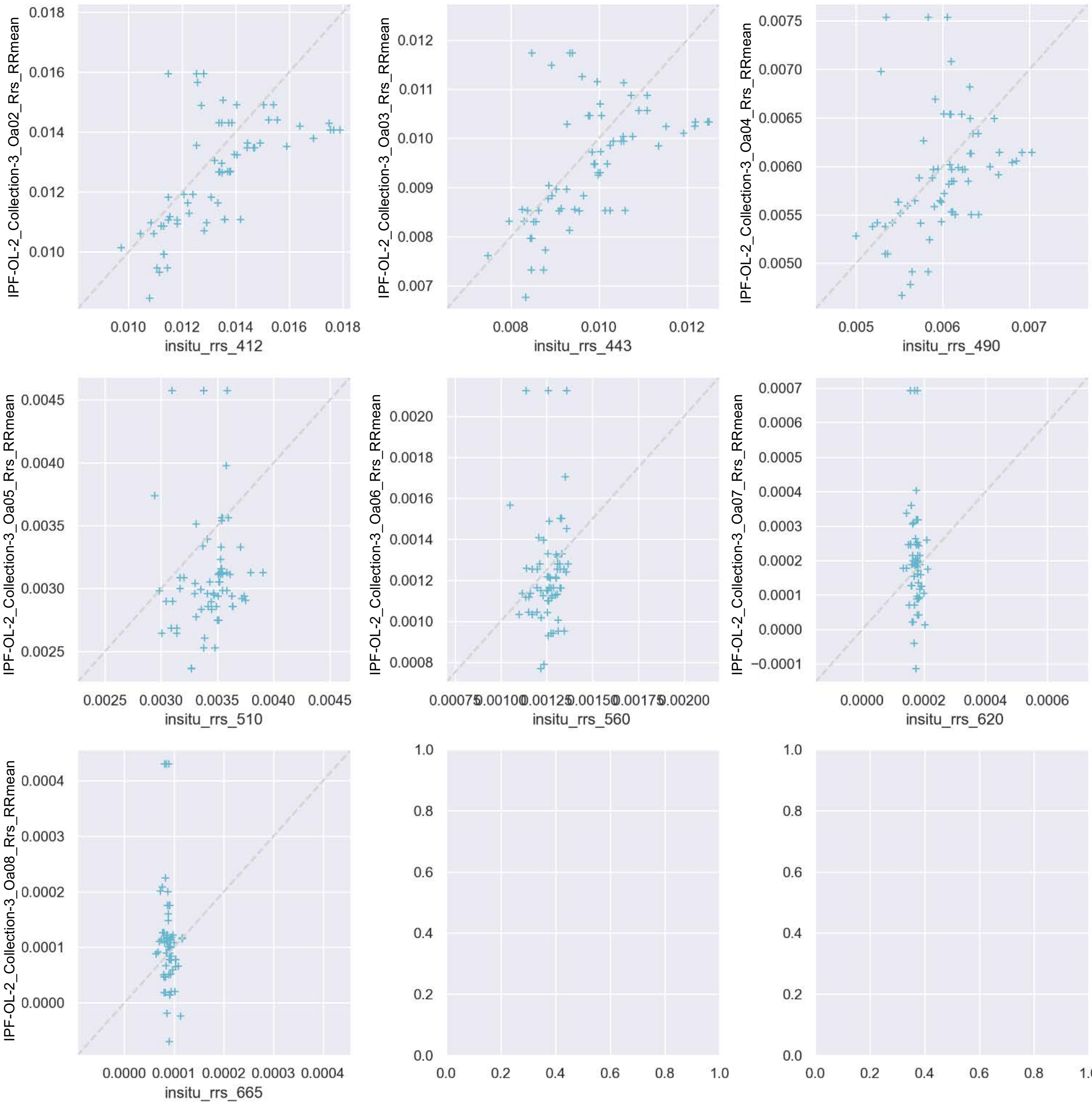


Figure 10: Example scatterplots of CBQ matchups between BASELINE processed OLCI data and in-situ measurements. Only points matched to all 4 processor configurations are shown. Dashed line shows 1:1 relationship.

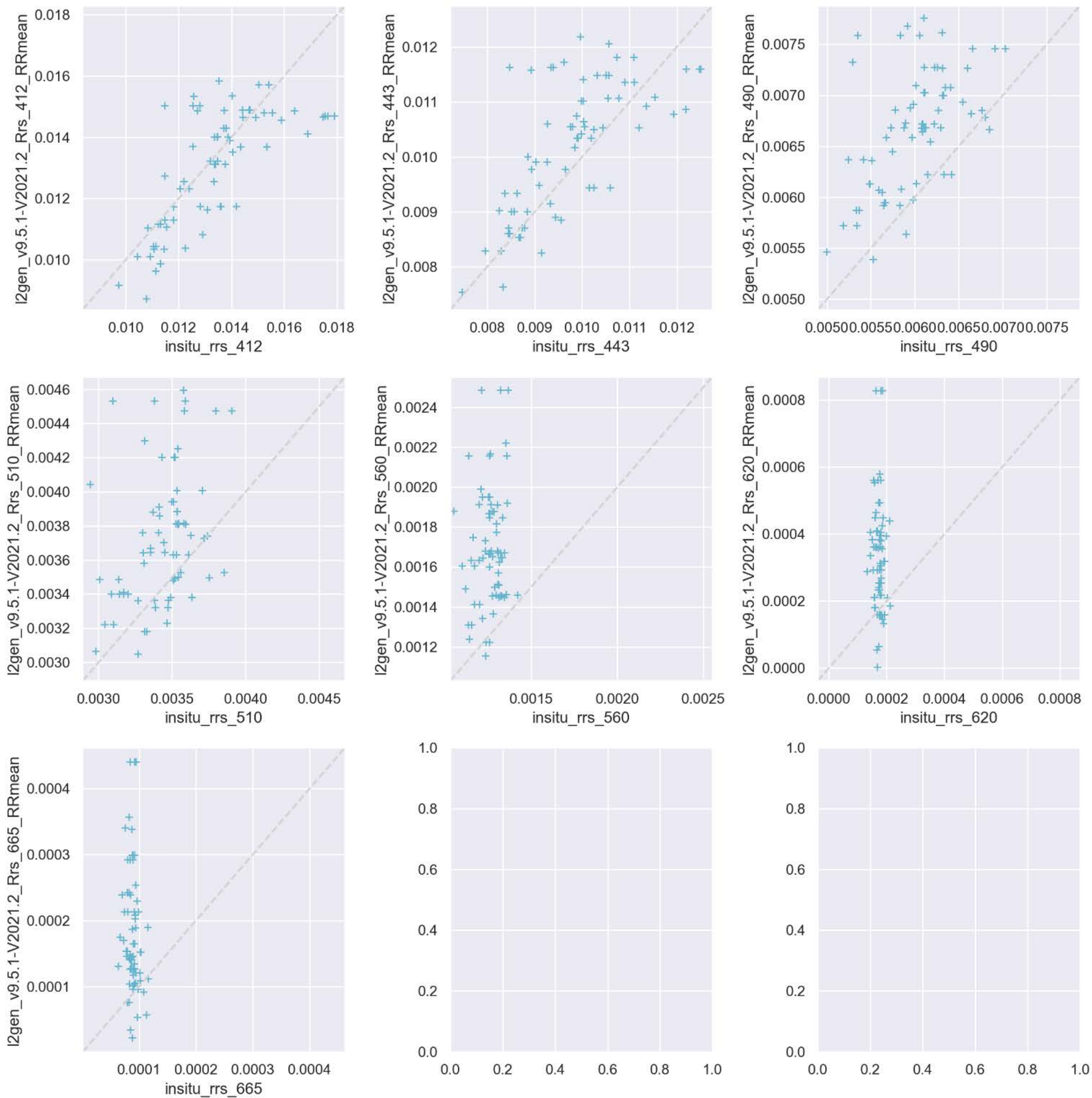


Figure 11: Example scatterplots of CBQ matchups between L2GEN processed OLCI data and in-situ measurements. Only points matched to all 4 processor configurations are shown. Dashed line shows 1:1 relationship.

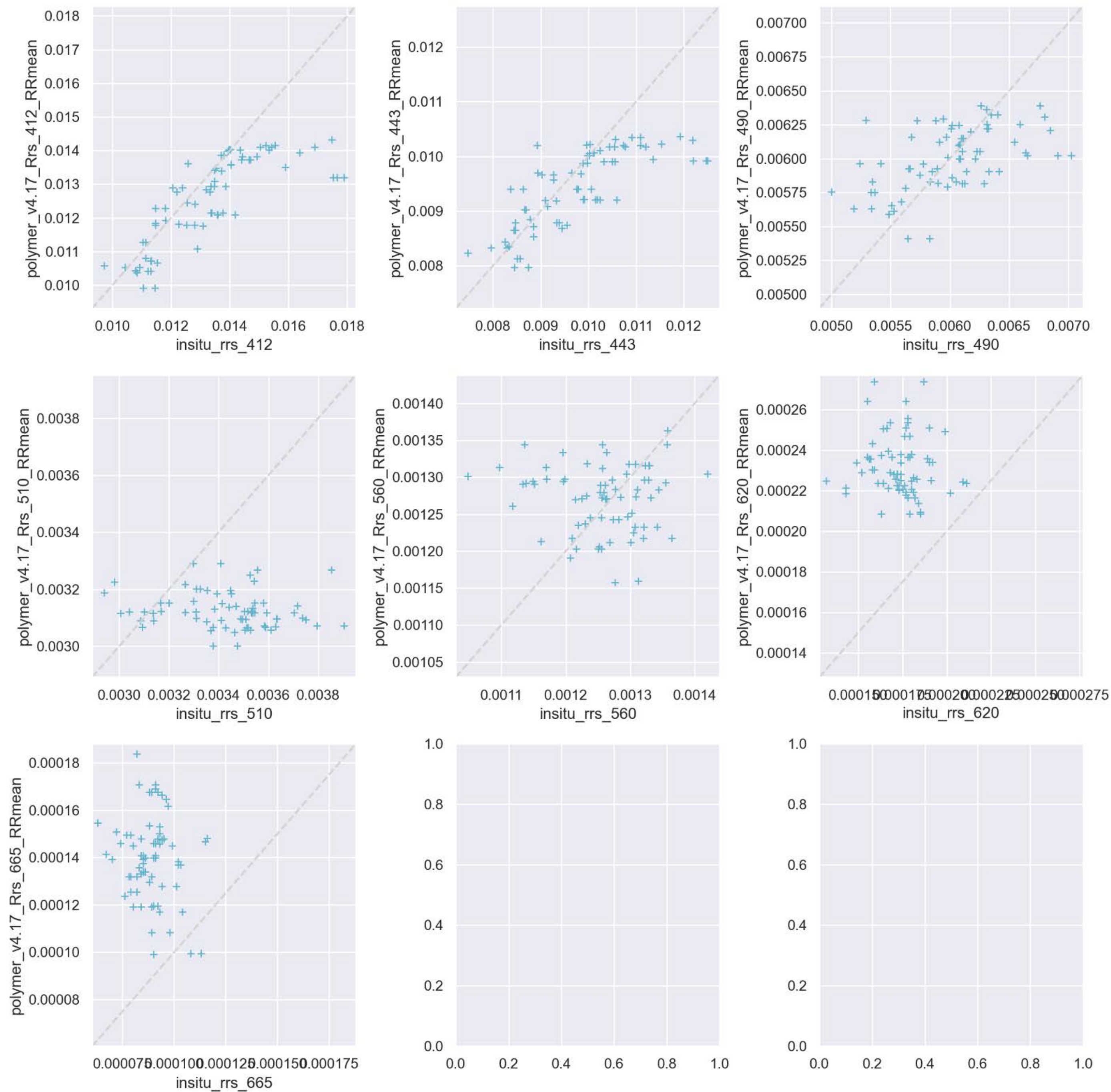


Figure 12: Example scatterplots of CBQ matchups between POLYMER processed OLCI data and in-situ measurements. Only points matched to all 4 processor configurations are shown. Dashed line shows 1:1 relationship.

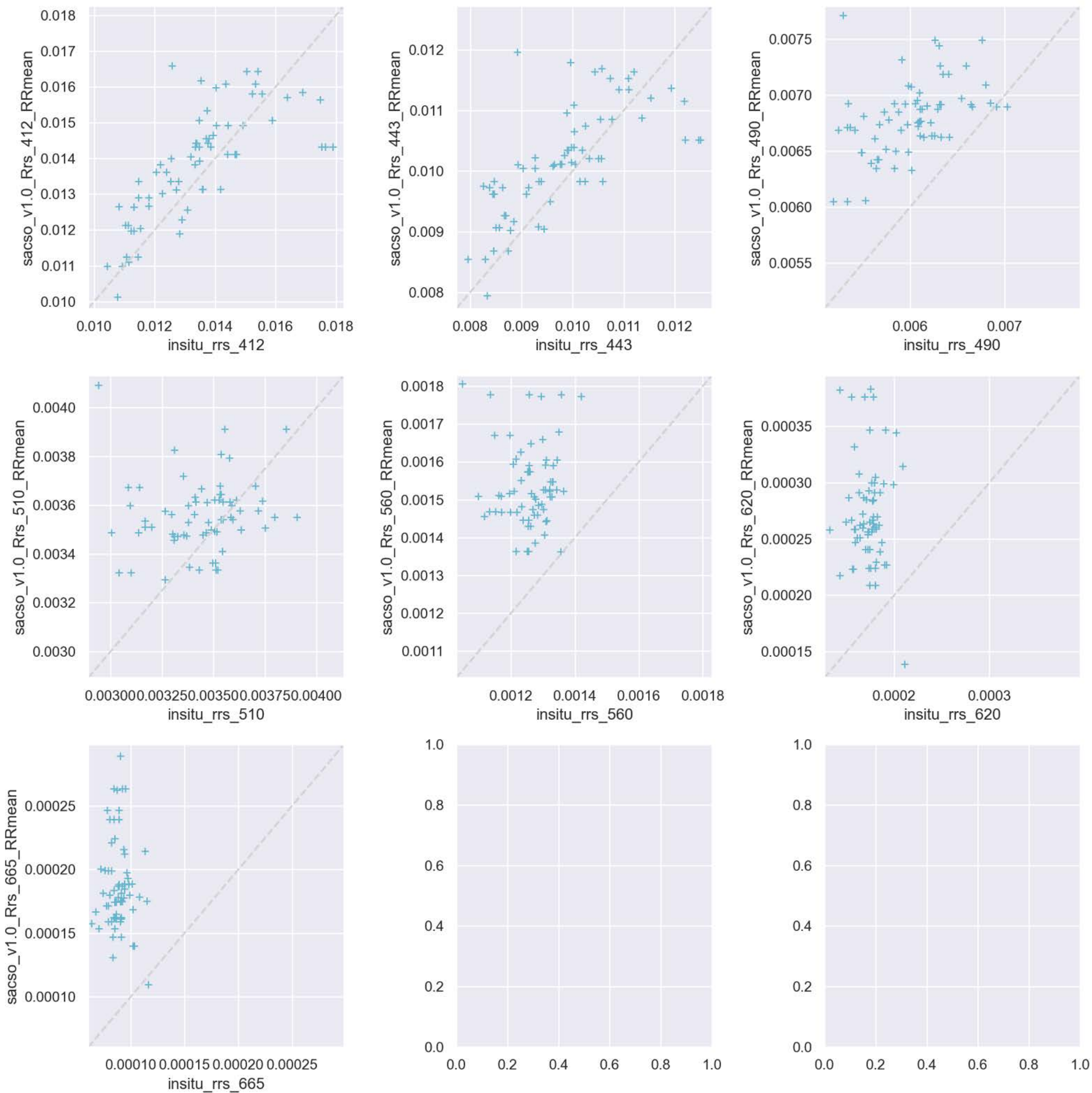


Figure 13: Example scatterplots of CBQ matchups between SACSO processed OLCI data and in-situ measurements. Only points matched to all 4 processor configurations are shown. Dashed line shows 1:1 relationship.

Processor	rrs_412	rrs_443	rrs_490	rrs_510	rrs_560	rrs_620	rrs_665	CHI2	SAM	Total.Scores
polymer_4.17	0.88	1.17	2.33	0.0	3.0	2.33	1.67	1.2	1.11	13.69
sacso_1.0	1.19	1.17	0.0	4.0	0.0	0.0	0.0	1.17	0.98	8.51
ipf_Collection_3	0.74	0.83	1.67	0.0	1.0	1.67	2.33	0.95	1.14	10.32
l2gen_9.5.1-V2021.2	1.19	0.83	0.0	0.0	0.0	0.0	0.0	0.68	0.78	3.48

Figure 14: Example scores per processor and per variable for CBQ analysis.

Processor	rrs_412	rrs_443	rrs_490	rrs_510	rrs_560	rrs_620	rrs_665	CHI2	SAM	Total.Scores
polymer_4.17	1.04	1.17	2.33	0.0	3.0	2.67	1.0	1.19	1.11	13.5
sacso_1.0	1.2	1.17	0.0	3.33	0.0	0.0	0.0	1.17	0.98	7.85
ipf_Collection_3	0.73	0.83	1.67	0.0	1.0	1.33	3.0	0.92	1.13	10.61
l2gen_9.5.1-V2021.2	1.04	0.83	0.0	0.67	0.0	0.0	0.0	0.72	0.78	4.03

Figure 15: Example scores per processor and per variable for IBQ analysis.

4.2 In water Round Robin

The demonstration in-water round robin (IW RR) performed for this report made use of in-situ data from the OCDB for the matchups, and granules processed using the default OMAPS POLYMER configuration for Rrs generation. The OMAPS matchup module was used to extract the relevant matchup data and the absolute matchups (central pixels of the macro pixel) were used for comparison. Full details of the IW RR multi-metric scoring approach are covered in the relevant ATBD.

As with the AC RR it is perhaps best to begin with an example of the final normalised score comparison between a set of candidate chlorophyll-a algorithms (Figure 16).

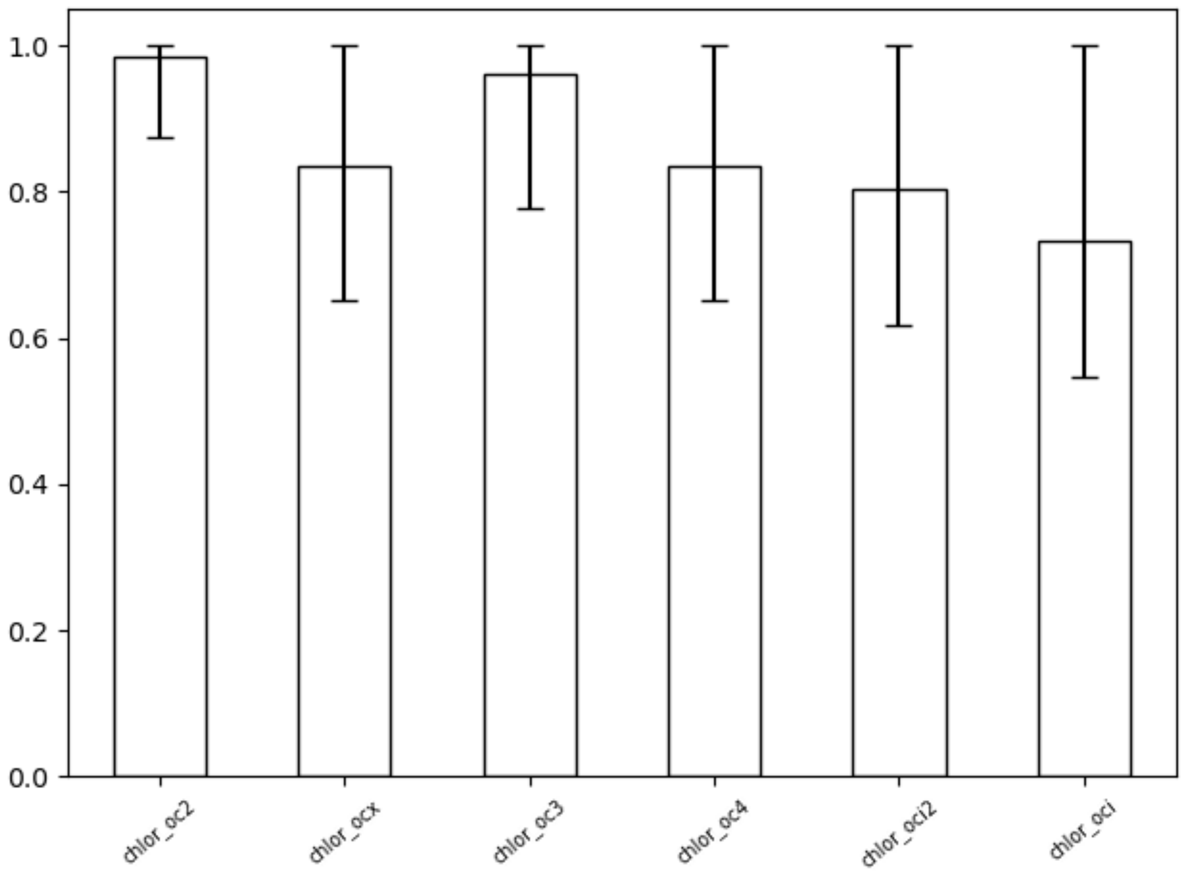


Figure 16: Summary plot of normalised multi-metric scores across six candidate chl-a algorithms using in-situ data from the OCDB for matchup and metric generation.

Here it appears that the simpler OC2 and OC3 algorithms seem to outperform the OC4 and OCI algorithms. This may seem surprising at first, but we must consider that the number of matchups is relatively small and there is a sampling bias towards coastal waters and complex Antarctic water conditions (while the OCI and OCI2 algorithms were designed for the clearest open ocean waters), meaning that overall performance may favour the coastal algorithms. The geographic location of the matchups are shown in Figure 4.

It is of note that the OCI and OCI2 algorithms revert to the OCx algorithm as chlorophyll-a concentrations increase, meaning that it should not be a surprise that they exhibit a similar performance to the OCx overall. Looking at Figure 17 we can see that indeed, according to the matchup information, the OC2 algorithm gives a better correlation coefficient, root mean square error, bias, intercept and number of retrievals than the OC4 (which only 'wins' on the slope metric).

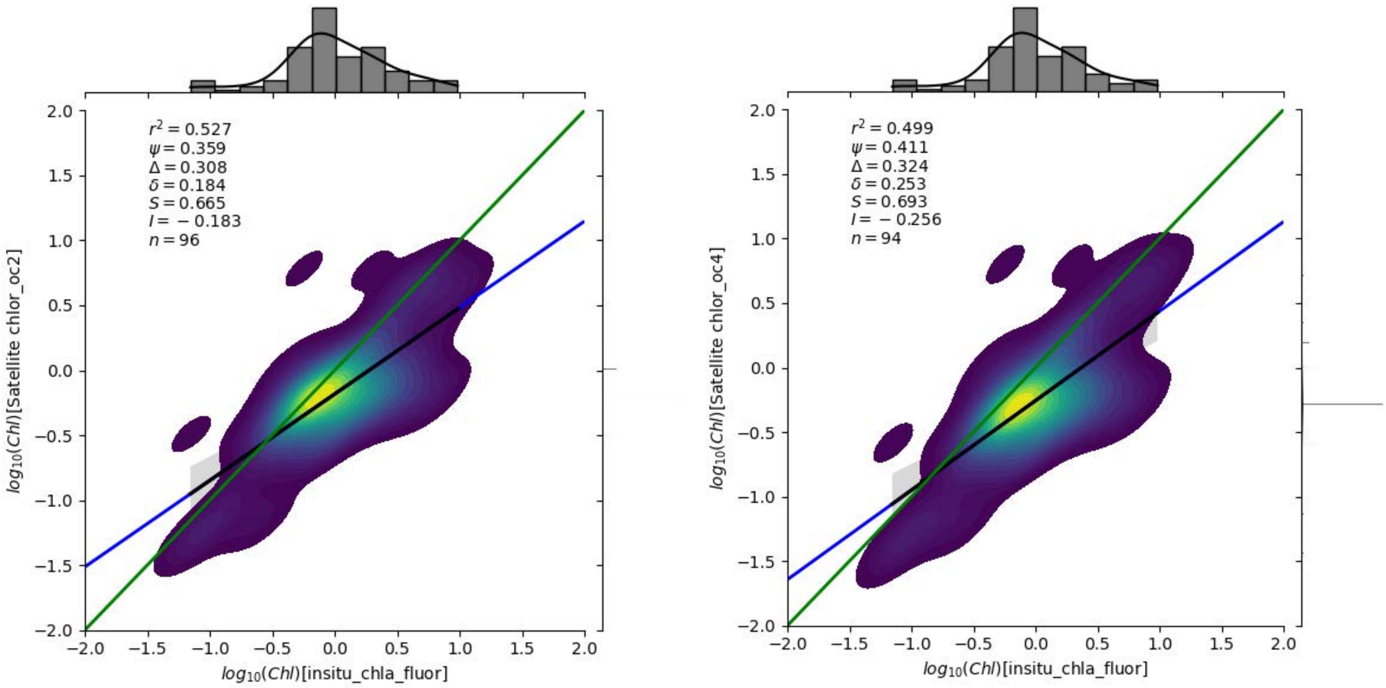


Figure 17: Example comparison of matchup performance between the OC2 and OC4 algorithms. Statistics shown are correlation coefficient (r),RMSE (ψ) unbiased RMSE (Δ), bias (δ), slope of type 2 regression (S), intercept of type 2 regression (I) and number of matchups (n).

4.2.1 Assessment per water class

The IW RR module can be run in a 'per-waterclass' assessment mode. This mode performs the same statistical analysis on the variable of interest as when run in the standard 'all-data' mode, but it creates subsets of the dataset based on the dominant optical water-type associated with each matchup. The optical water types used as default within the OMAPS processor are those created for a MERIS/OLCI waveband set within the OC-CCI project during the creation of the OC-CCI V5.0 processor. These water classes are shown in Figure 18 and were created by performing a cluster analysis of the OC-CCI 'MERIS like' v5.0 spectra, sampled across a range of biogeochemical provinces. The default config file and cluster definition file can be found within the OMAPS repository at 'online_processor/module_scripts/OWT_classification/example_files/'.

Running the IW RR in 'per-waterclass' mode for OLCI-only matchups highlights the sampling bias and limited number of matchups currently available (this might be improved soon as more data is added to the OCDB and corresponding granules are added to the matchup comparison). From 101 matchups, none show water classes 3,4,6,7 or 10 as dominant; water classes 1,2,5,8,9,12 and 14 have less than 10 matchups that they dominate and 56 of the points are dominated by a single water class, the class 11. This confirms the point made earlier that the current samples available for matchup are not representative of the optical diversity of the oceans.

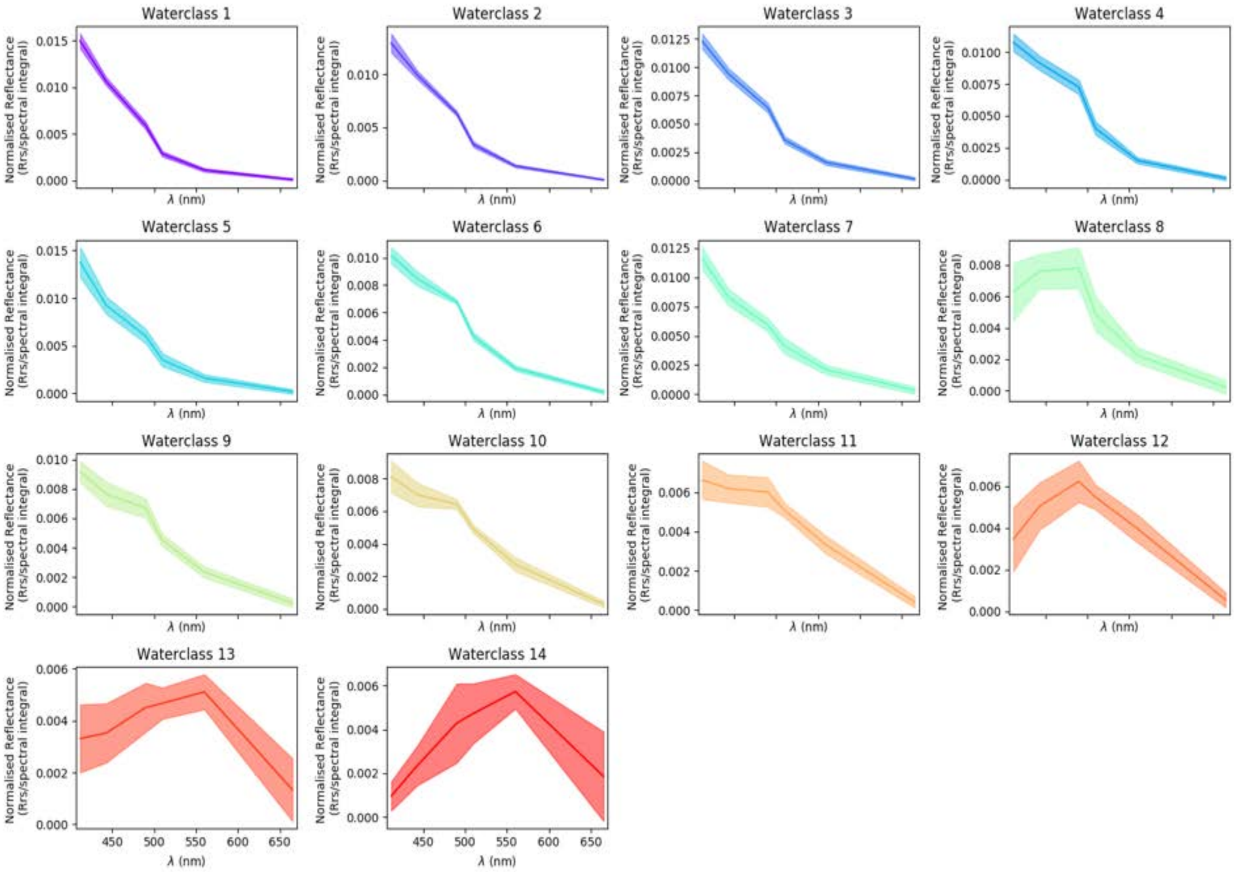


Figure 18: The water classes used by default in the OMAPS processor.

It is of note that within the OC-CCI V5 algorithm round robin (performed using a merged 20 year dataset of matchups against a MERIS-reference sensor merged-dataset) the OC2 was also found to be the optimal algorithm for water class 11. Though far from conclusive, this could suggest that using the same algorithm per water class assignment as used in OC-CCI would be suitable for OMAPS POLYMER processed data. This OC-CCI assignment is the default set for product blending within the OMAPS processor configuration:

- chlor_oci weighting = water class 1
- chlor_oci2 weighting = water classes 2,3,4,5
- chlor_oc2 weighting = water classes 7,8,9,10,11
- chlor_ocx weighting = water classes 6,12,13,14

Given the current limitations of a match-up based assessment it is important to use the diagnostic granule set, described in this product validation plan, to give some insight into the performance of both atmospheric correction and in-water algorithms.

4.3 Performance under known adverse conditions

In addition to match-up statistics it is also of interest to process (to level 2) diagnostic granules that contain known 'troublesome' pixels to test the robustness of the atmospheric processor or module configuration in conditions that may not be represented in the in-situ database. Below are demonstrative examples across a range of potentially difficult in-water and atmospheric optical conditions of interest. In all cases below we are comparing the SACSO, L2GEN, POLYMER (all with IdePIX masking) against the EUMETSAT BASELINE Collection 3 processing using the recommended OLCI masks. All the images shown below are available at a higher resolution, allowing readers to zoom in and closer inspect the data fields, along with additional examples for each condition type in the annexes to this document.

Examples can be found for each of the processing configurations for which they appear to be the most suitable processor but looking across many diagnostic granules the overall points to summarise across the various conditions are:

1) The largest single difference in coverage between the processors is the ability of the POLYMER and SACSO to provide data in the glint region.
2) SACSO is most able to cope with extremely difficult atmospheric conditions (but it does take much longer to process granules with SACSO than the POLYMER and L2GEN processors).
3) The BASELINE Collection 3 processing lies somewhere between SACSO and POLYMER in its ability to provide data coverage in dusty conditions but it can let corrupted data through (this might be improved with the application of an additional masking scheme such as IDEPIX).
4) The BASELINE Collection 3 and L2GEN seem to exhibit less smooth field of Rrs than POLYMER and SACSO in a significant number of granules (for a prime example see the Rrs443 in the images for the Bay of Biscay on 2018/07/09 10:35:19).
5) The BASELINE Collection 3 appears to often provide data further into estuarine and riverine areas. This may be driven by differing masking approaches.

From the analysis of diagnostic granules below, it seems that we might generally state that:

POLYMER provides a good coverage and smooth Rrs fields across much of the observation conditions. However, SACSO is the optimal configuration for retrievals in complex atmospheric conditions and the BASELINE Collection 3 seems to provide the best coverage in coastal areas.

4.3.1 Phytoplankton Bloom Granules

Example processed granules representing ‘Phytoplankton Bloom’ conditions are shown in comparison to the EUMETSAT BASELINE processing in Figure 19, Figure 20, and Figure 21. Higher resolution and additional examples available in appendix_Pythoplankton_bloom_granules.pdf.

Across all three of these figures it is clear that the POLYMER and SACSO processors provide enhanced coverage in the eastern portions of the granules due to their decreased sensitivity to glint. In non-glint areas POLYMER and SACSO seem to provide more data than both L2GEN and the BASELINE processor but this is not uniformly the case. In some areas it can be seen that POLYMER, SACSO or both are lacking data that is present in the L2GEN or BASELINE processor configurations. Given that these granules were chosen because of the presence of distinct phytoplankton and in-water features we should pay particular attention to those areas.

4.3.1.1 Assessment for Figure 19 (20180524T102451)

There is a strong feature, likely a coccolithophore shedding event, in the lower right corner of Figure 19 which shows up clearly as high reflectance in both the POLYMER and SACSO images, with little signal being seen in the chlorophyll-a estimates (blue outlined region). Both the L2GEN and BASELINE images have most of this feature masked out.

In the left section of the image (orange outlined region) the POLYMER and SACSO results for reflectance (443nm) seem less noisy than those from the L2GEN and BASELINE processing configurations. This left section also seems to show better coverage in the SACSO and BASELINE processing than the POLYMER products.

4.3.1.2 Assessment for Figure 20 (20180716T073803)

The most complete picture of the Red Sea is clearly provided by the SACSO processor in this case. The POLYMER, L2GEN and BASELINE processor all show significant regions of masked data compared to the SACSO results. The areas provided solely by SACSO seem to show reasonable quality data with realistic looking oceanographic features (bloom eddies) in the southern section of the image (orange outlined region). Once again, we see that the L2GEN processor configuration provides a more cautious approach to masking with more data removed in coastal and hazy areas.

4.3.1.3 Assessment for Figure 21 (20180730T080327)

As with the images already discussed, Figure 21 shows that the L2GEN processing has the most conservative masking and lowest coverage. This image also contains a coccolith bloom which presents as high reflectance in the BASELINE, POLYMER and SACSO results (masked in L2GEN). In all three of the processors that leave this area unmasked the chlorophyll-a products don't appear to show corruption or false artifacts in the coccolith bloom region (orange outlined area). Where the POLYMER, SACSO and BASELINE results differ most strongly is in the mid to eastern section of the image. The BASELINE results show high chlorophyll-a values as you move from the Barents to the White Sea. SACSO does not show such high chlorophyll-a, looking consistent with the L2GEN chlorophyll-a estimates and SACSO has better coverage than other products in the White Sea. POLYMER shows anomalously low chlorophyll-a results with a blocky artifact pattern (blue outlined area).

4.3.2 Aerosol Granules

Example processed granules representing complex 'Aerosols' conditions are shown in comparison to the EUMETSAT BASELINE processing in Figure 22, Figure 23 and Figure 24. Higher resolution and additional examples available in appendix_Gaseous_correction_granules.pdf

4.3.2.1 Assessment for Figure 22 (20171210T183320)

In Figure 22 there is a clear plume of wildfire smoke that extends offshore that is largely (successfully) masked out in both the POLYMER, SACSO and L2GEN processing. The BASELINE processor with advised flags applied seems to have significant area of bright reflectances that are caused by smoke contamination in the core of the atmospheric eddy (green outlined area). As in a number of other images the masking in the L2GEN appears a little noisier than some of the other processors. One point of note is the line of small clouds that can be seen trending northeast (orange outlined area). These clouds are casting shadows on the surface of the ocean which are masked in the L2GEN, apparent in the BASELINE processing and present in the POLYMER and SACSO (but not as strong as in the BASELINE). The SACSO processor seems to push the furthest into the smoke plume (without the massively contaminated reflectance estimates of the BASELINE), showing the strength of the updates made to this processor for aerosol conditions. Ideally further validation in these areas would be good to test the quality of the additional SACSO data, though the values appear reasonable given the nearby pixels.

4.3.2.2 Assessment for Figure 23 (20180726T215905)

This image contains volcanic ash and aerosols, which contaminate the image horizontally from about the middle of the image upwards. All processors seem to mask the volcanic aerosols with POLYMER masking the most data in the upper half of the image, followed by L2GEN, then the BASELINE with SACSO providing the most retrievals in this region. There are two further areas of interest in this image. Firstly, in the eastern section there is an island, around which elevated chlorophyll-a concentrations can be seen in the BASELINE and SACSO processors but not in the POLYMER and L2GEN processors (orange outlined area). Secondly there is a small patch of data (green outlined area) that shows slightly elevated Chl-a in the SACSO data, less elevated Chl-a in the BASELINE processing and no data in the L2GEN and POLYMER configurations. Unfortunately, we do not have validation data in this area where we might be able to determine which approach is providing the estimate closest to the truth.

4.3.2.3 Assessment for Figure 24 (20180326T110646)

This image contains a large plume of Saharan dust over the Atlantic Ocean. The dust contamination has led to much of the image being masked with increasing order of coverage being POLYMER, L2GEN, BASELINE and SACSO. There is a small area of data in the POLYMER processing (south-eastern corner) that looks dubious and should probably have been masked out (orange outlined area). Though the BASELINE configuration provides the greatest coverage in this case it might not be considered a ‘good’ trait as there is a band of relatively high reflectance values (See 443 band) extending from the middle of the image to towards the south-west that are likely giving an artificially high estimate of surface reflectance due to the aerosols as the spatial pattern follows the dust cloud more than any oceanographic features (green outlined area). The SACSO configuration also provides reflectance estimates in this area (that is suspicious in the BASELINE processing) but the patterns in the data and the values look more representative of ocean conditions. However, there is a large area of high chlorophyll-a identified in the south-east corner of the image in the SACSO processing (red outlined area) that looks like it might be from SACSO over-reaching with its aerosol correction (offshore water reflectance at 665 nm looks too high).

4.3.3 Cloud Granules

Example processed granules representing ‘Clouds’ conditions are shown in comparison to the EUMETSAT BASELINE processing in Figure 25 and Figure 26. Higher resolution and additional examples available in appendix_Cloud_type_granules.pdf

4.3.3.1 Assessment for Figure 25 (20180419T205315)

The granule in Figure 25 was chosen due to the presence of clouds and cloud shadows over clear, deep-ocean waters. The left two-thirds of the image shows very similar results between POLYMER, SACSO, L2GEN and BASELINE processors. None of the processors seems to show the speckle or halo effects around the cloud masked areas that are sometimes associated with cloud shadow contamination, so it seems that in these conditions the cloud masking for all four processors is performing nominally. As with other granules, SACSO seems to provide a little more coverage in this western region (orange outlined region). The eastern (right) third of the image shows a stark difference between POLYMER/SACSO and L2GEN/BASELINE processors. This is due to the ability of POLYMER and SACSO to cope with glint conditions which are largely masked in L2GEN and BASELINE processing configurations. That said, the POLYMER and SACSO reflectances (see 443 band) in the glint region are visibly different but it is not possible to easily determine which is closer to the truth.

4.3.3.2 Assessment for Figure 26 (20180524T022054)

Much of Figure 26 backs up the conclusions drawn from Figure 25, though there is a couple of additional points of note in this image. Firstly, there are two areas of hazy cloud in the image, the first in the top left corner and the second extending horizontally across the mid-left of the granule. In both these areas the POLYMER and L2GEN configurations seem to mask more data the SACSO and BASELINE configurations. Secondly, there is an example of a strong cloud shadow in this image that might be making a visible impact in the SACSO and BASELINE processing configurations (orange outlined area), seen mostly in the 443 and 510 reflectance bands.

4.3.4 Glint Granules

The strength of POLYMER (and SACSO) in glint contaminated areas has already been mentioned in the discussion of other granules but here we discuss an example processed granule specifically dedicated to the 'Glint' diagnostic. Higher resolution and additional examples available in appendix_Glint_granules.pdf

4.3.4.1 Assessment for Figure 27 (20180518T195153)

The ability of POLYMER to cope with glint is well documented and was the original 'problem' that the algorithm was designed to solve. The image in Figure 27 contains a large, strongly glinted area in the eastern quarter of the image. This strong glint has no appreciable impact on the reflectance and chlorophyll-a fields that are produced by the POLYMER and SACSO processors. The entire region is masked out by the BASELINE and L2GEN processors. It is of note that the POLYMER and SACSO processing seems to resolve realistic looking oceanographic structures within this glint region and those structures straddle the glint/non-glint boundary in some cases (example in orange highlighted region).

4.3.5 Sea Surface Condition Granules

Mixed sea-ice and cloud example is shown in Figure 28. Higher resolution and additional examples available in appendix_Sea_surface_conditions.pdf

4.3.5.1 Assessment for Figure 28 (20180524T134649)

This image contains sea-ice which can cause both straylight contamination and sub-pixel contamination issues for the atmospheric correction processor. All four of the processing configurations seem to show a generally similar level of cloud masking with the L2GEN processor perhaps showing very slightly less data (more conservative flagging) in the central block of data, the BASELINE processor masking data in the upper right region that is present in the other processors and POLYMER perhaps letting a few bright pixels through just south of Greenland (handful of brighter specks within dark Rrs_443 region, highlighted in orange).

One significant difference between the four processors in this image is the chlorophyll-a concentrations. There appears to be a significant phytoplankton bloom located south of Greenland (green highlighted region) and which shows very different concentrations across the four configuration options. The L2GEN processor gives the lowest chlorophyll-a concentrations for the bloom, with POLYMER and SACSO showing concentrations perhaps twice that of the L2GEN estimate. The BASELINE processing on the other hand has chlorophyll-a concentrations that are an order of magnitude higher than the L2GEN estimates. As with other granules discussed above, a lack on co-located in-situ data means that we cannot say for certain which processor might be providing the most correct estimates of chlorophyll-a for this bloom, but it highlights these conditions as worthy of further investigation. The fine scale structure of the bloom is also more clearly depicted by the other processors compared to the BASELINE configuration, perhaps partly owing to the reduction of the total Chlorophyll-a scale to only 256 available values in the BASELINE processing.

4.3.6 Corrupted Data Granules

Beyond the range of naturally observed atmospheric and in-water conditions observed by ocean colour sensors we must also consider the ability of an atmospheric processor to correctly handle missing or corrupted data.

4.3.6.1 Assessment for Figure 29 (20160614T202246)

A known example of such ‘corruption’ in the Level-1B data is the ‘missing camera’ anomaly which re-occurs on OLCI every half-year on average (and for which users are informed accordingly). It is clear from this image that all the processors and masking schemes handle the corrupted data correctly.

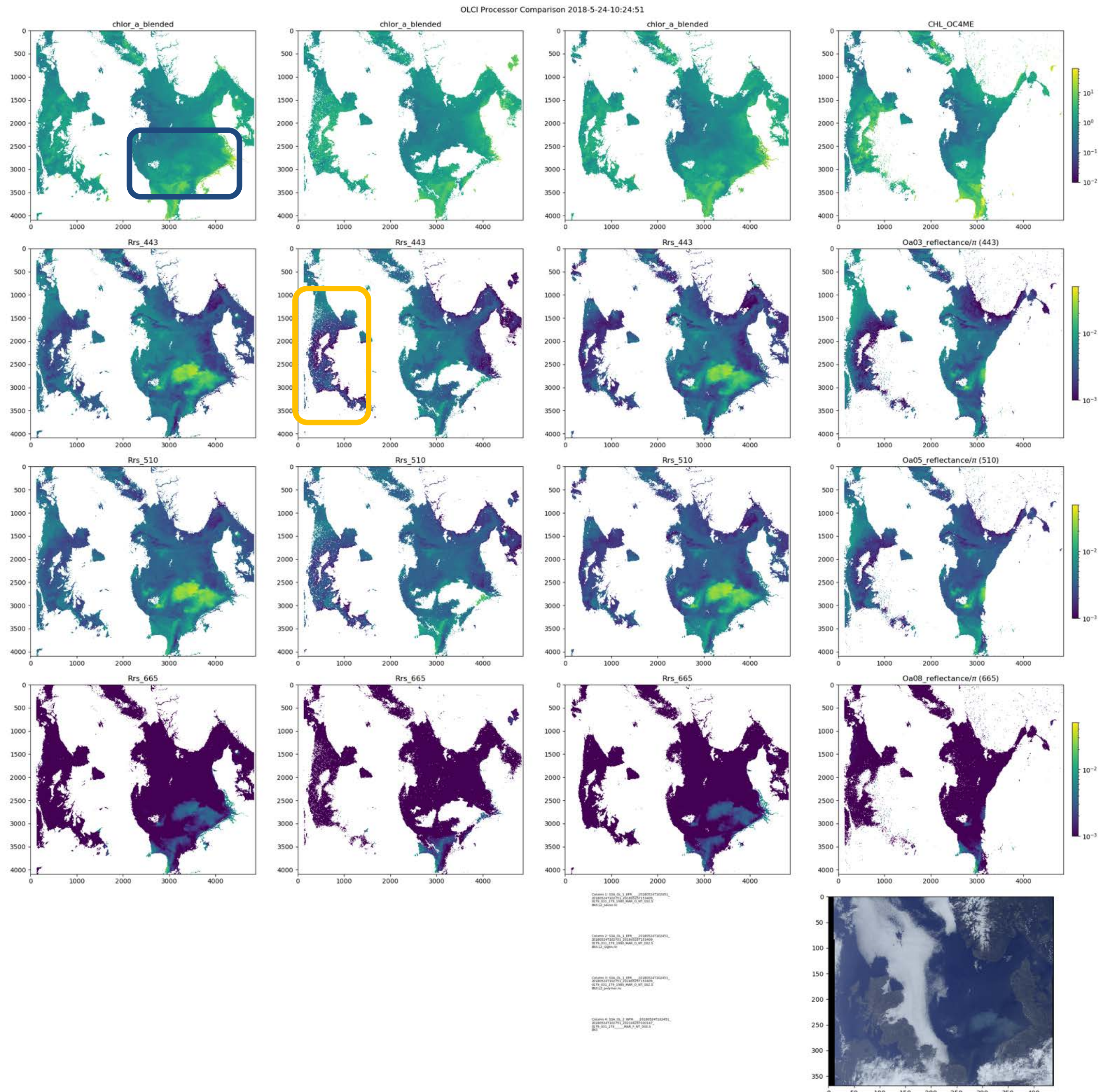


Figure 19: Comparison of (left to right) SACSO, L2GEN, POLYMER and EUMETSAT BASELINE Collection-3rd processing for Granule S3A_OL_1_EFR____20180524T102451_20180524T102751 , containing a phytoplankton bloom in the North Sea. Blue and orange outlined regions highlight areas of likely coccolith bloom and enhanced noise in L2GEN and BASELINE respectively. For closer inspection see appendix_Phytoplankton_bloom_granules.pdf for higher resolution version of image.

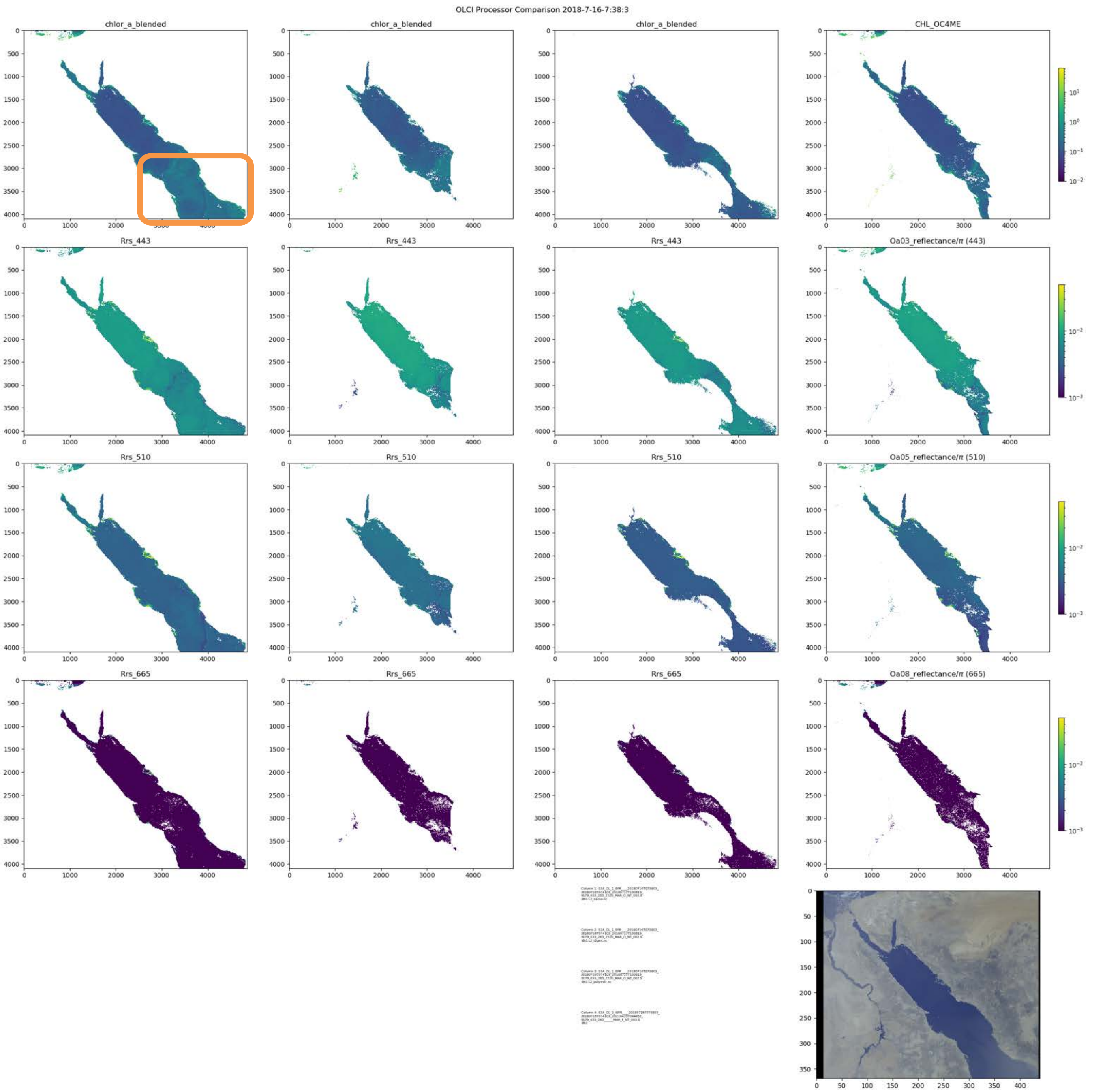


Figure 20: Comparison of (left to right) SACSO, L2GEN, POLYMER and EUMETSAT BASELINE Collection-3 processing for Granule S3A_OL_1_EFR____20180716T073803_20180716T074103 , containing a phytoplankton bloom in the Red Sea. Orange outlined region shows area of chlorophyll-a depicted eddy structures (visible in SACSO processing). For closer inspection see appendix_Phytoplankton_bloom_granules.pdf for higher resolution version of image.

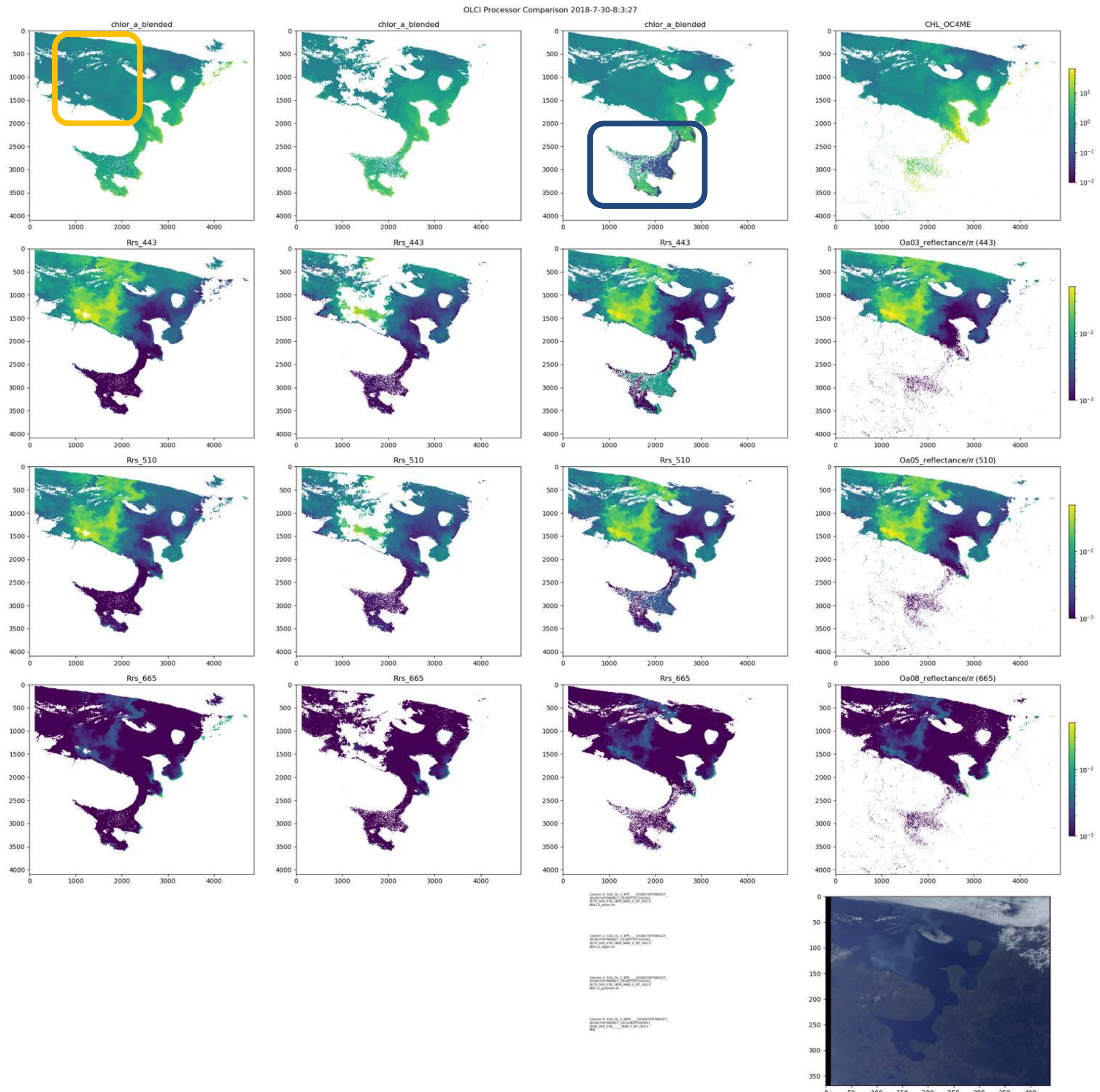


Figure 21: Comparison of (left to right) SACSO, L2GEN, POLYMER and EUMETSAT BASELINE Collection-3 processing for Granule S3A_OL_1_EFR____20180730T080327_20180730T080627, containing a phytoplankton bloom in Barents and White Sea area. Orange outlined region shows area of coccolith bloom and blue region shows anomalous POLYMER behaviour. For closer inspection see appendix_Phytoplankton_bloom_granules.pdf for higher resolution version of image.

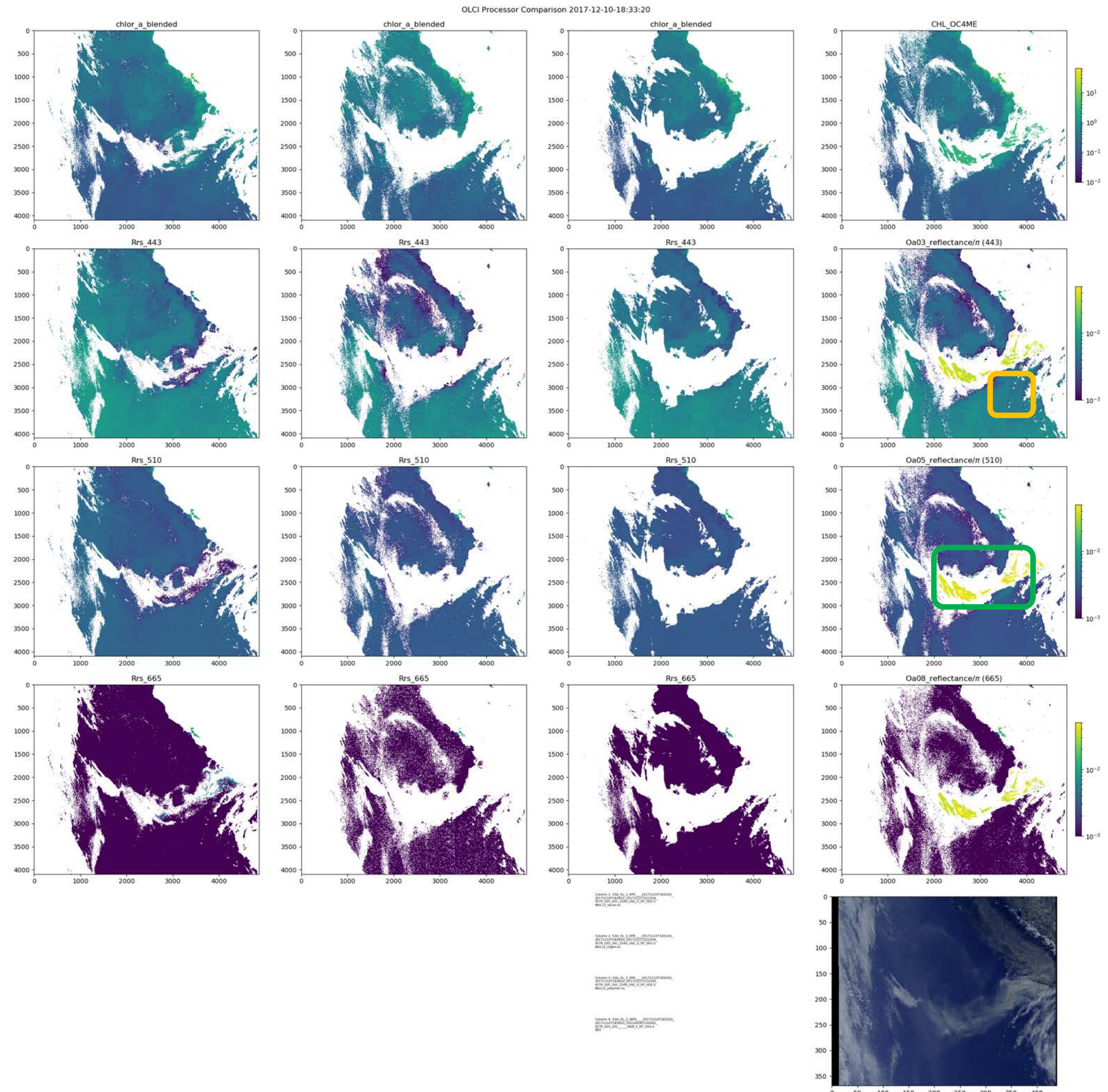


Figure 22: Comparison of (left to right) SACSO, L2GEN, POLYMER and EUMETSAT BASELINE Collection-3 processing for Granule S3A_OL_1_EFR____20171210T183320_20171210T183620 , containing a plume of wildfire smoke West of California. Orange outlined region shows an area of cloud shadow impact on Rrs estimates. Green outline shows an area of dust contamination. For closer inspection see appendix_Aerosol_Granules.pdf for higher resolution version of image.

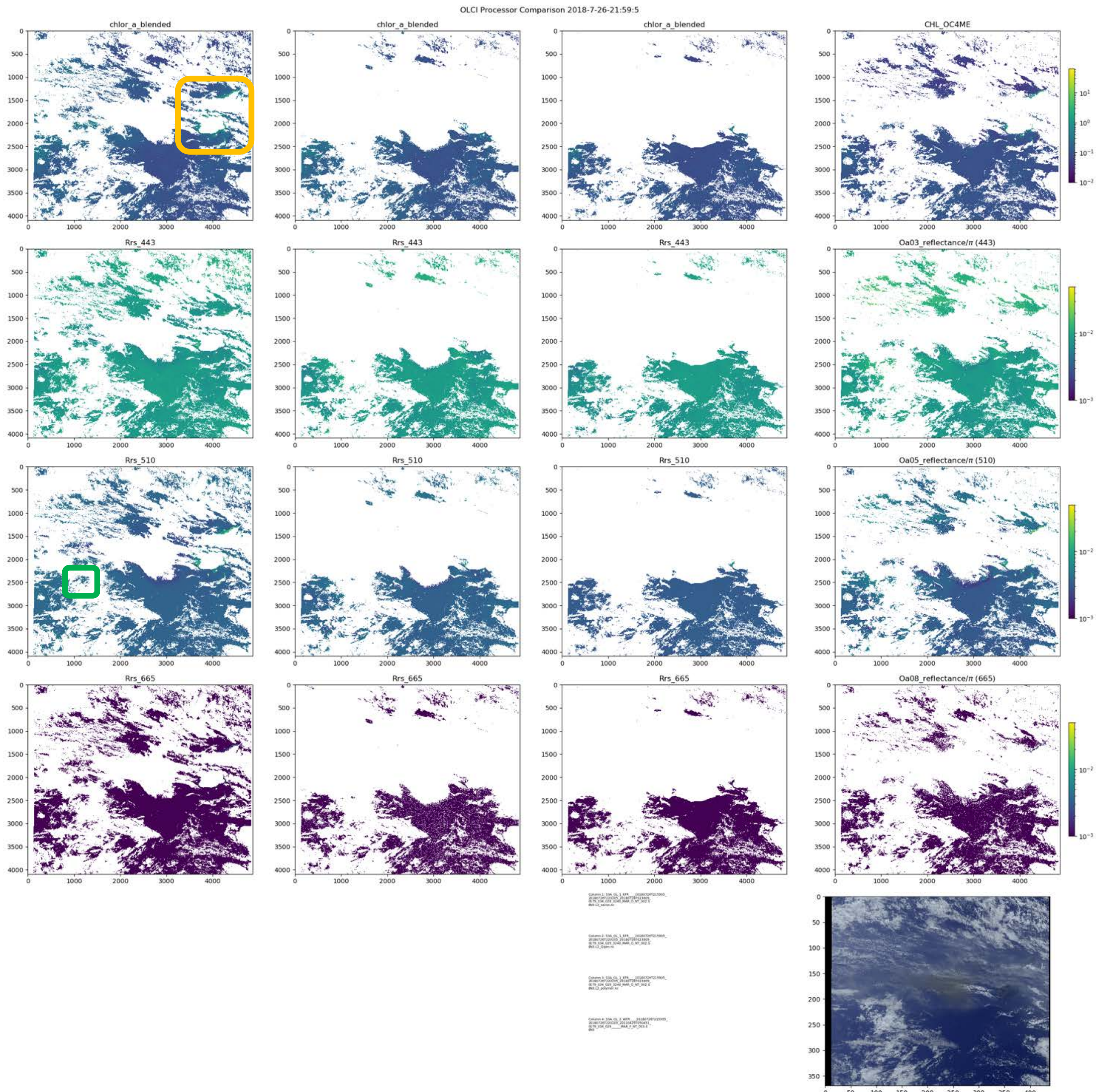


Figure 23: Comparison of (left to right) SACSO, L2GEN, POLYMER and EUMETSAT BASELINE Collection-3 processing for Granule S3A_OL_1_EFR____20180726T215905_20180726T220205 , containing a plume of volcanic ash near Fiji. Orange outlined region shows area of differing chlorophyll-a estimates proximal to land. Green outline shows area of differing chl-a estimates between SACSO and BASELINE. For closer inspection see appendix_Aerosol_Granules.pdf for higher resolution version of image.

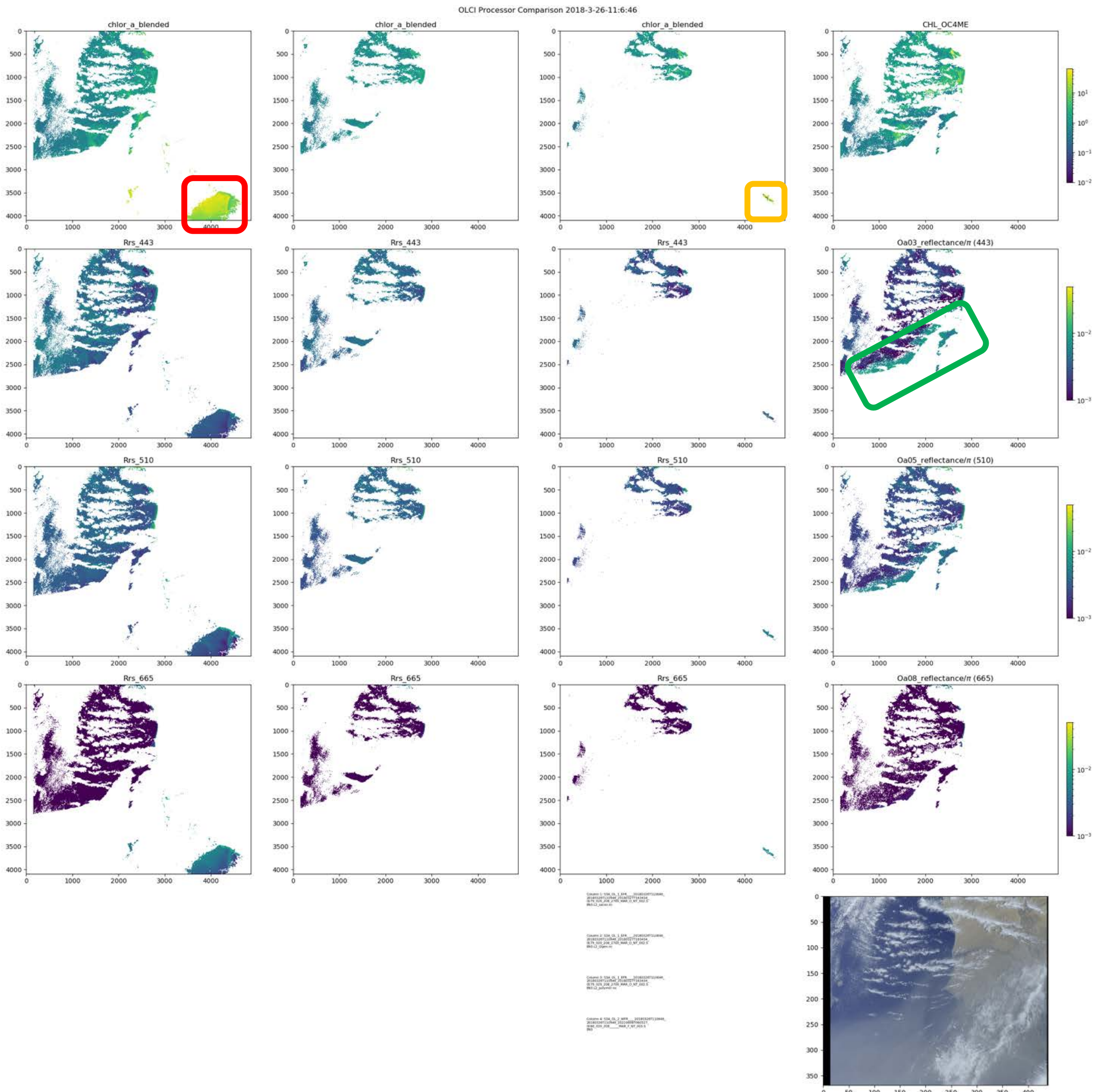


Figure 24: Comparison of (left to right) SACSO, L2GEN, POLYMER and EUMETSAT BASELINE Collection-3 processing for Granule S3A_OL_1_EFR____20180326T110646_20180326T110946 , containing a massive Saharan dust plume over the Eastern Atlantic. Orange outlined region shows area of contaminated data in POLYMER processing. Green outline shows area of contaminated data in BASELINE processing. Red outlined area shows suspiciously elevated chlorophyll-a estimates from SACSO processing. For closer inspection see appendix_Aerosol_Granules.pdf for higher resolution version of image.

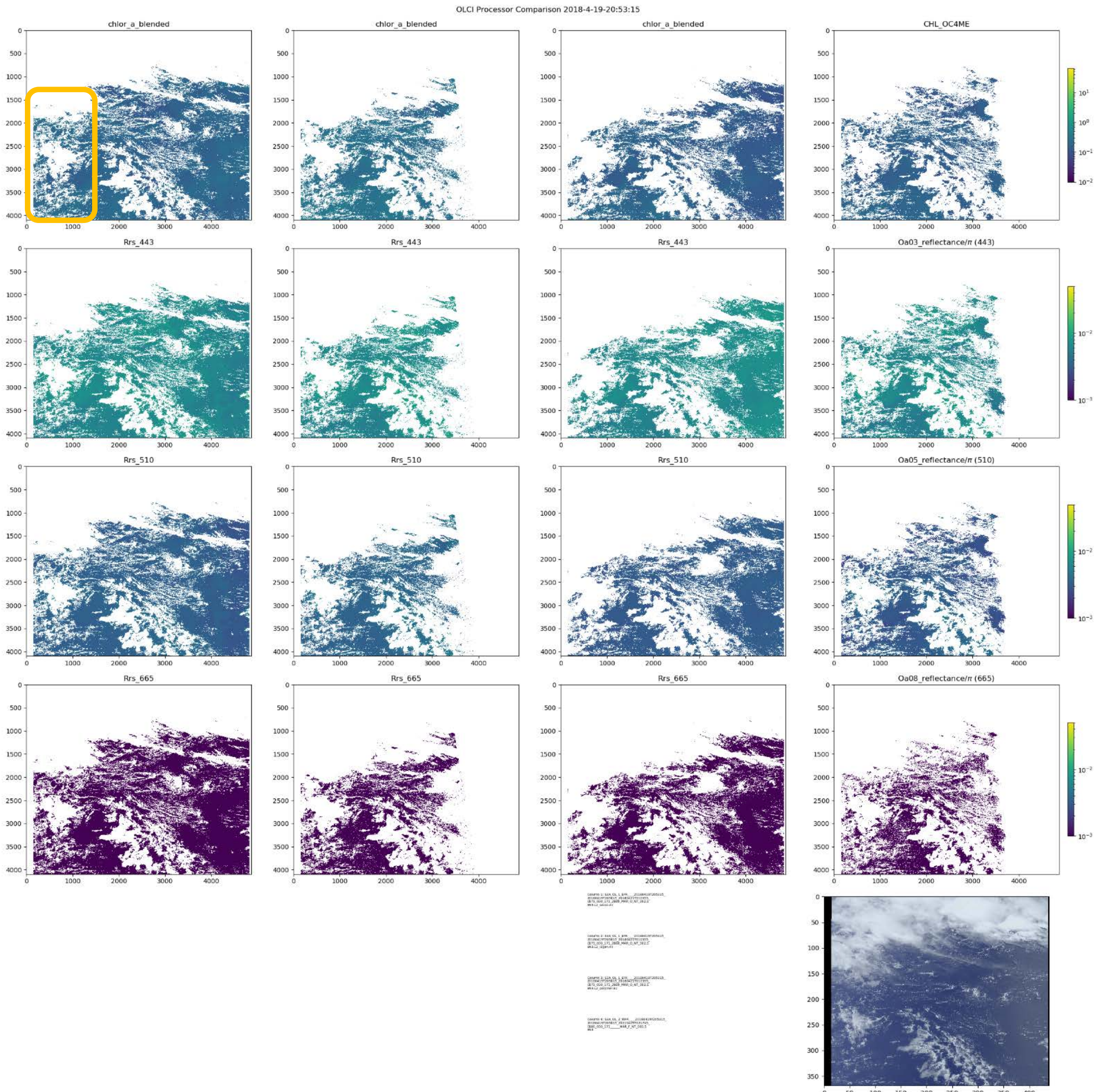


Figure 25: Comparison of (left to right) SACSO, L2GEN, POLYMER and EUMETSAT BASELINE Collection-3 processing for Granule S3A_OL_1_EFR____20180419T205315_20180419T205615 , containing a mix of clouds and cloud shadows over the North Pacific. Orange outlined region shows area of increased SACSO coverage. For closer inspection see appendix_Cloud_type_granules.pdf for higher resolution version of image.

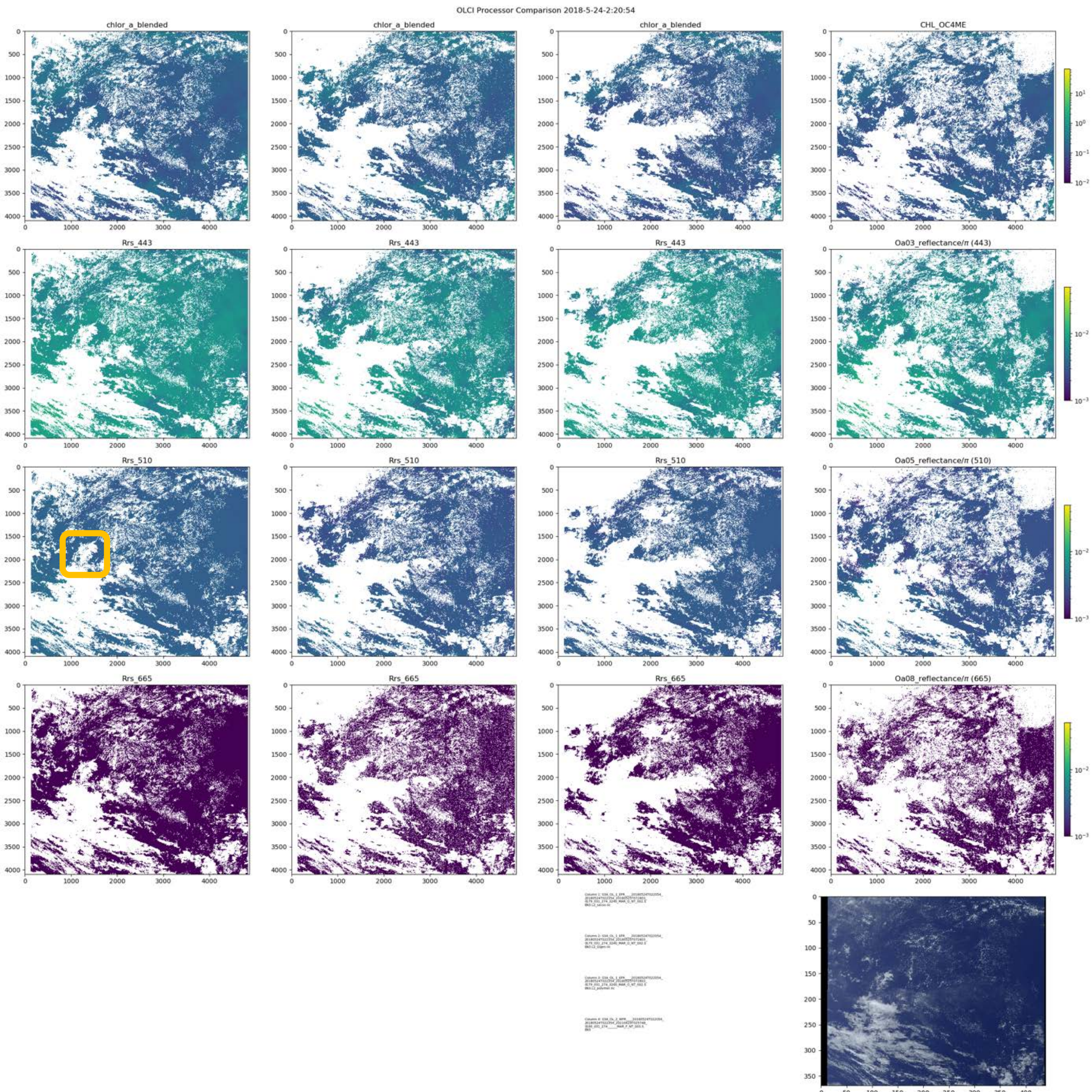


Figure 26: Comparison of (left to right) SACSO, L2GEN, POLYMER and EUMETSAT BASELINE Collection-3 processing for Granule S3A_OL_1_EFR____20180524T022054_20180524T022354 , containing a many small clouds over the ocean south of Indonesia. Orange outlined region shows area of potential cloud shadow impact. For closer inspection see appendix_Cloud_type_granules.pdf for higher resolution version of image.

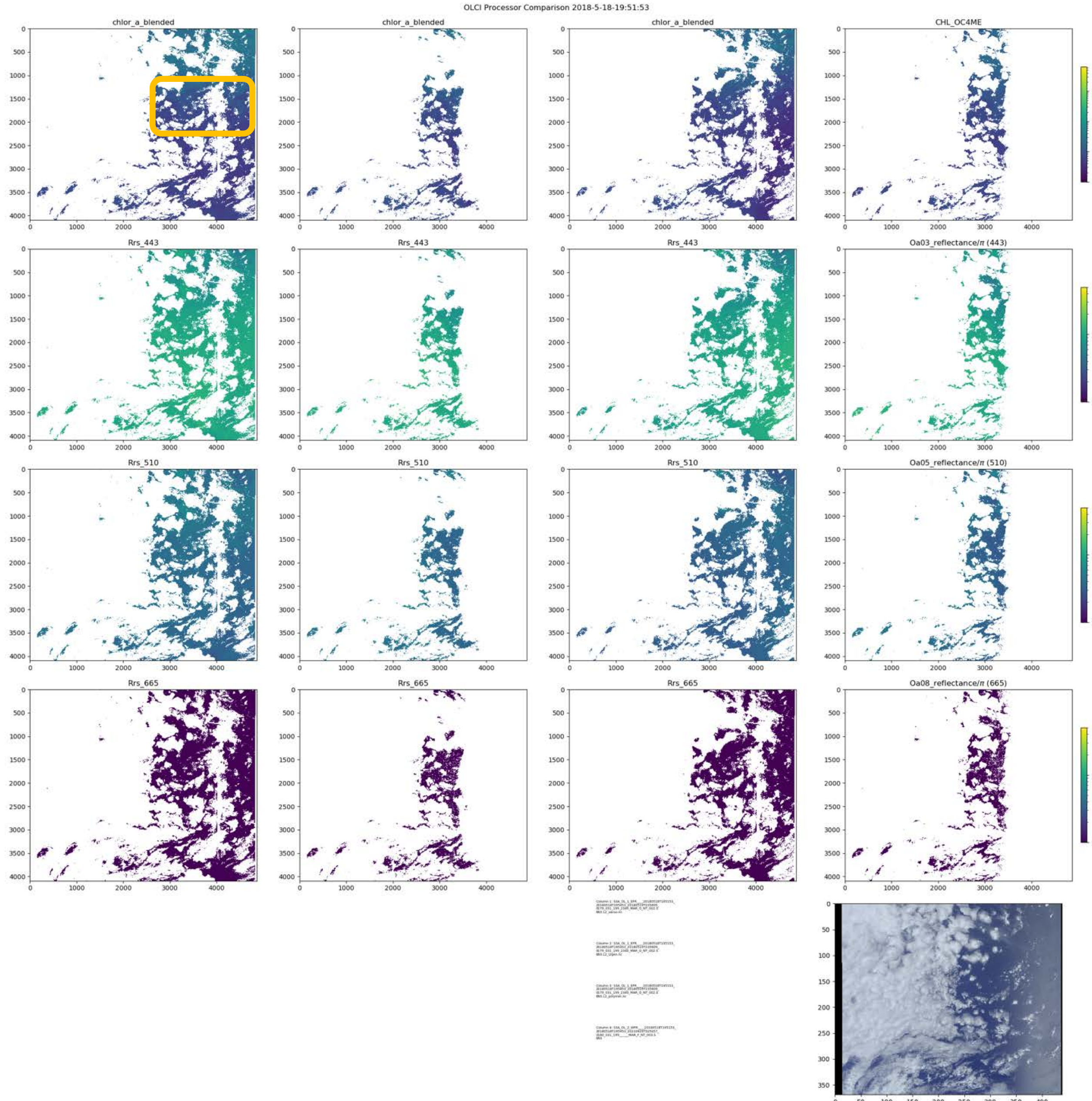


Figure 27: Comparison of (left to right) SACSO, L2GEN, POLYMER and EUMETSAT BASELINE Collection-3 processing for Granule S3A_OL_1_EFR____20180518T195153_20180518T195453, containing a large band of high glint in the right of the image. Orange outlined region shows area with structured chlor-a field that crosses glint boundary region. For closer inspection see appendix_Glint_granules.pdf for higher resolution version of image.

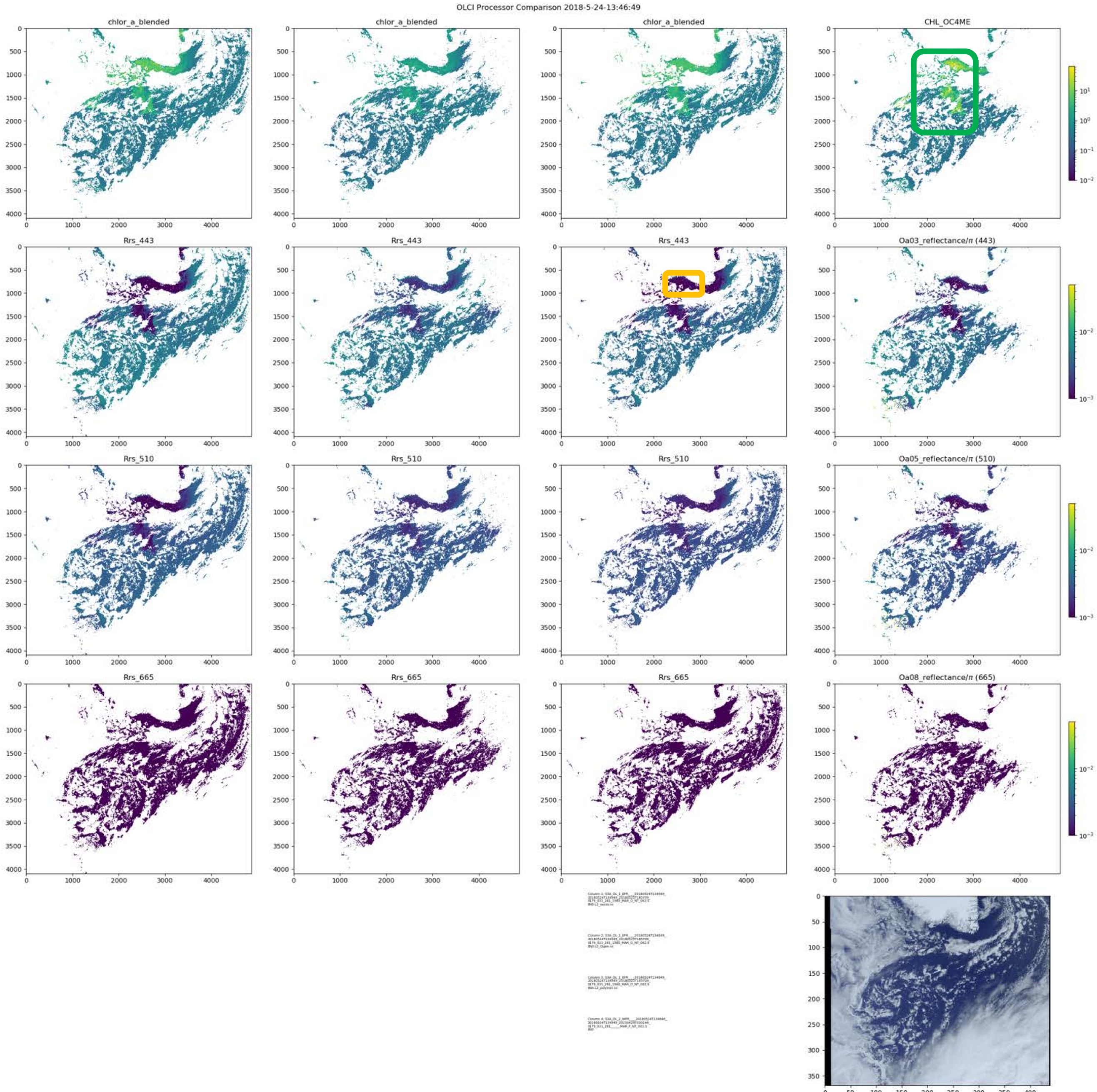


Figure 28: Comparison of (left to right) SACSO, L2GEN, POLYMER and EUMETSAT BASELINE Collection-3 processing for Granule S3A_OL_1_EFR____20180524T134649_20180524T134949 , containing a mix of clouds and sea-ice near Greenland. Orange outlined region shows area with some speckle in 443 band. Green outlined region shows large range in chlorophyll-a estimates between the different processors. For closer inspection see appendix_Sea_surface_conditions_granules.pdf for higher resolution version of image.

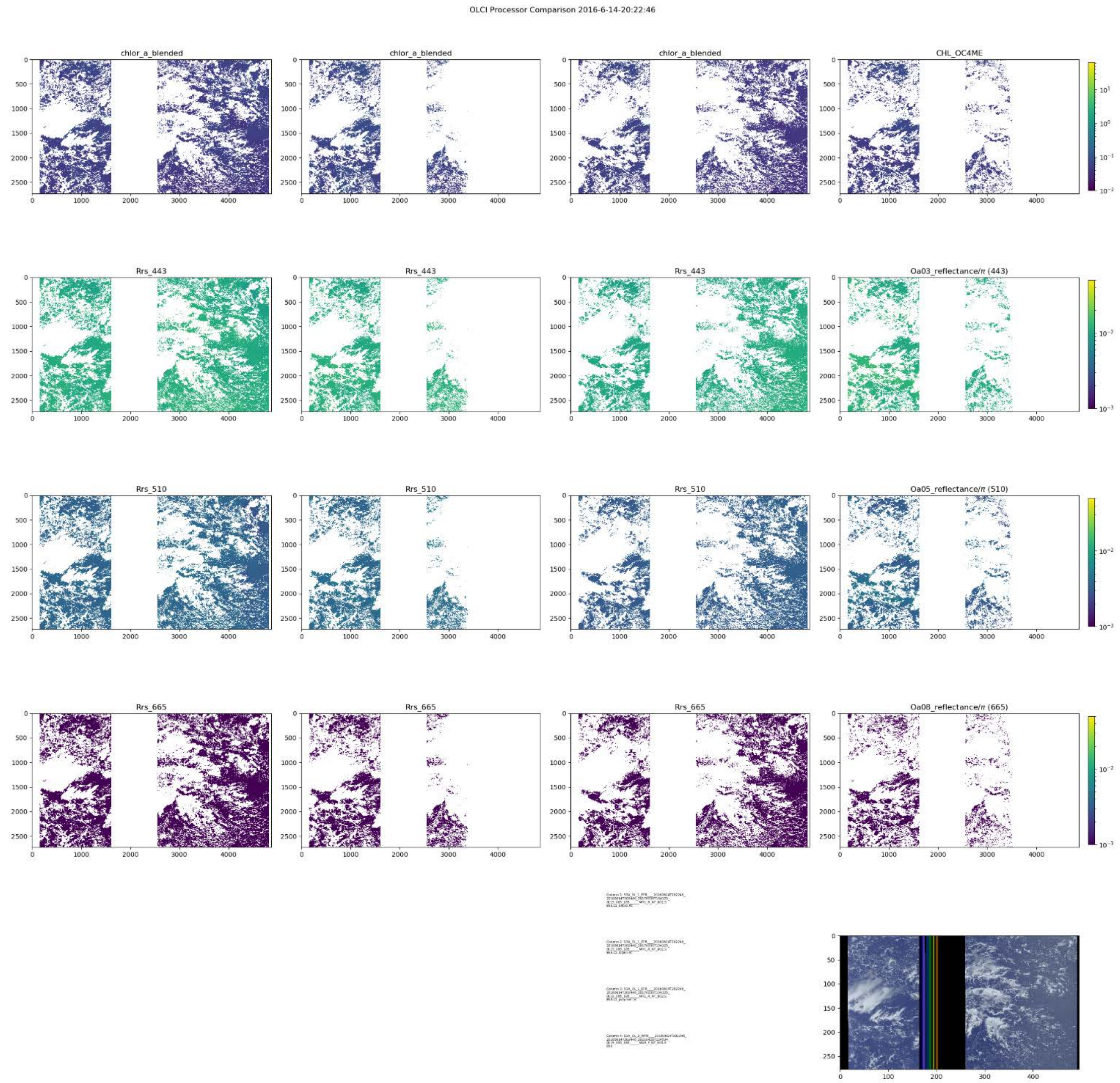


Figure 29: Comparison of (left to right) SACSO, L2GEN, POLYMER and EUMETSAT BASELINE Collection-3 for Granule S3A_OL_1_EFR____20160614T202246_20160614T202446_20170930T194125_0119_005_185 , containing an OLCI camera anomaly.

4.4 Time series analysis

Exact matchups between satellite overpass and in-situ measurements are a small subset of the available in-situ data available. Repeated measurements at a given location enable a time-series analysis to confirm whether the satellite products are correctly describing the phenology of the area, even if point to point comparisons are not possible. Suitable sites for such analysis are the Hawaii Ocean Time-series (HOTs) for chlorophyll-a and MOBY (Marine Optical BuoY) for Rrs, both representing highly oligotrophic water conditions.

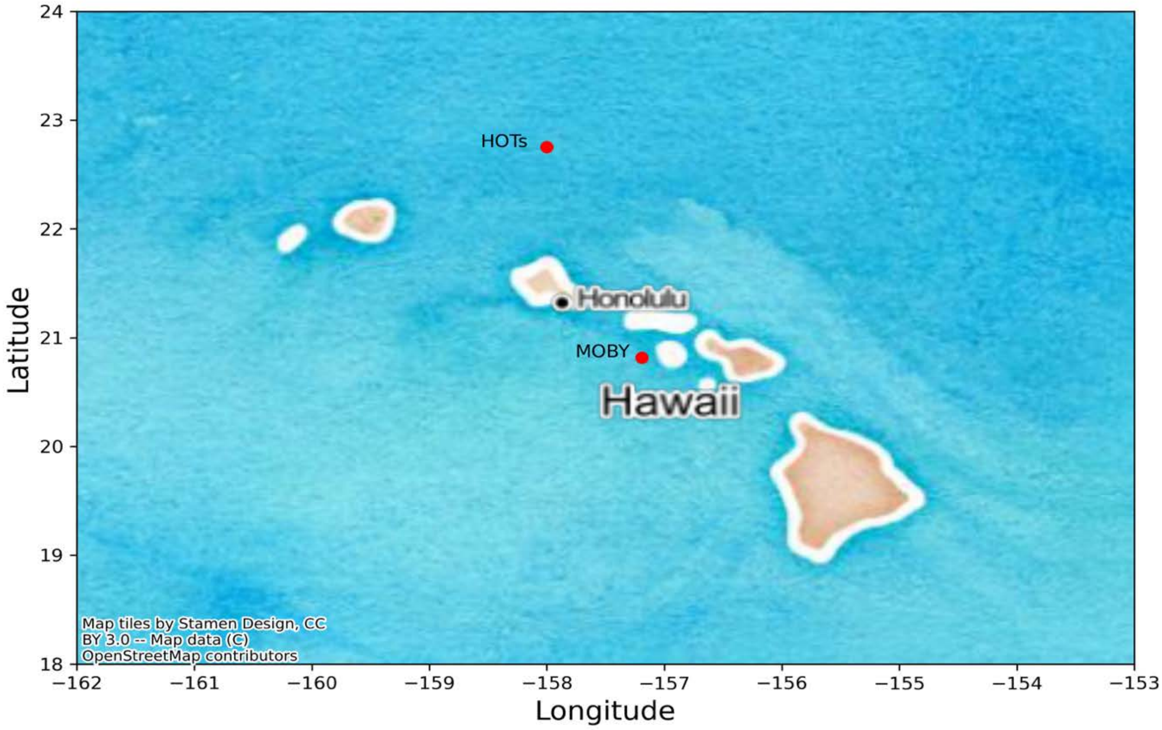


Figure 30: Map showing locations of HOTs and MOBY measurements.

A time series of granules containing the HOTs and MOBY locations were processed using the default configurations for all processors. A 5x5 pixel subsample was then extracted from the processed data centred on both the HOTs measurements and MOBY location. For each 5x5 pixel extraction the satellite chlorophyll-a and Rrs at bands 443nm, 510nm and 865nm were used to derive summary statistics (the value at the central grid pixel and mean, median, number of valid pixels, and standard deviation for the 5x5 grid).

The example comparative time series between satellite Chla and HOTs in-situ Chla are shown in Figure 31. The BASELINE Chla data is derived using the blended oci and OC4ME method (RD-4) and stored within the files as log10(Chla) so we have unlogged the mean and median values for comparison. The other 3 processing configurations (POLYMER, SACSO and L2GEN) all use a waterclass blending approach which blends the oci, oci2, oc2 and ocx algorithms (see section 4.2.1), though in these clear waters we are likely to only see significant contributions from oci and oci2 algorithms. All four processor configurations produce chlorophyll-a estimates that are of the right order of magnitude when compared with the values measured in-situ using both HPLC and Fluorometric methods. Although the seasonal cycle is relatively limited in magnitude at this location the timing of the peaks and troughs in the satellite estimates also show agreement with the phenology of the in-situ measurements. The SACSO chlorophyll-a is perhaps a little more noisy than the other three processors.

Corresponding example comparisons between in-situ measurements and the remote sensing estimates can been observed for the Rrs time series with the MOBY measurements in Figure 32 and Figure 33. As example, we can see that for the 442.5 nm band the four processors perform similarly. The magnitude and seasonal variation are relatively well reproduced by all processors over the time series, as might be expected from the matchup scatterplots produced by the AC RR. It was mention in the AC RR assessment above that the absolute errors in the POLYMER processor seemed smaller than for the other processors, and averaging of positive and negative errors could give a higher weight to high accuracy with low precision. The time series information shows that the reduced dynamic range in the POLYMER 442.5nm band plot could be considered as reflecting a higher precision, while maintaining a similar mean accuracy to other processors.

The 510 nm band shows a more significant difference between the processors. In this case we can see that the BASELINE and POLYMER processors seem to do a good job of reproducing the 510 in-situ measurements but the L2GEN and SACSO both seem to consistently overestimate Rrs 510 in this region. Again, the POLYMER dynamic range in the plot is much smaller than for the other processors.

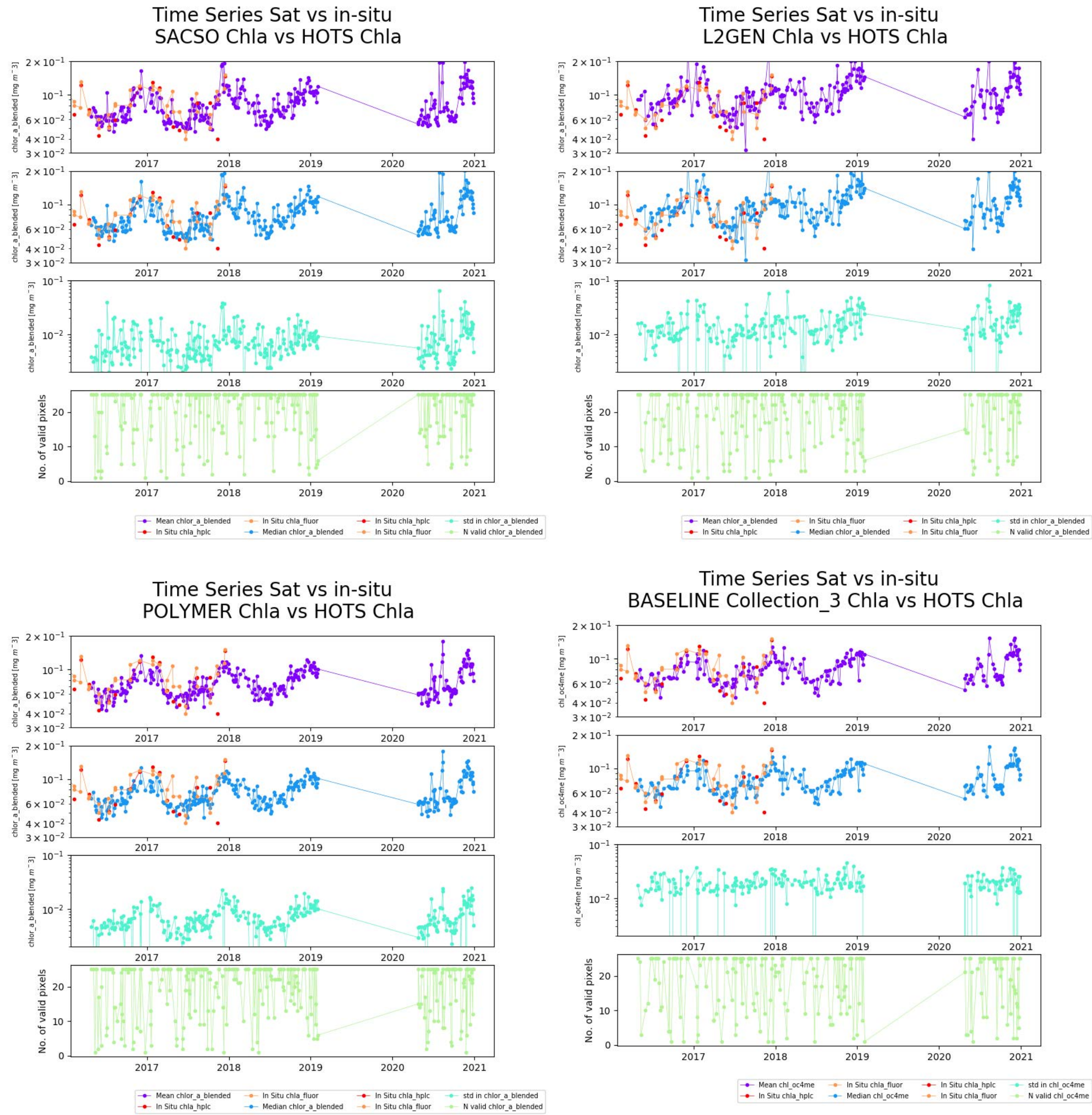


Figure 31: Example comparison of the extracted mean, median, standard deviation and number of valid pixels for chl-a for SACSO, L2GEN, POLYMER and BASELINE processors and the HOTS dataset.

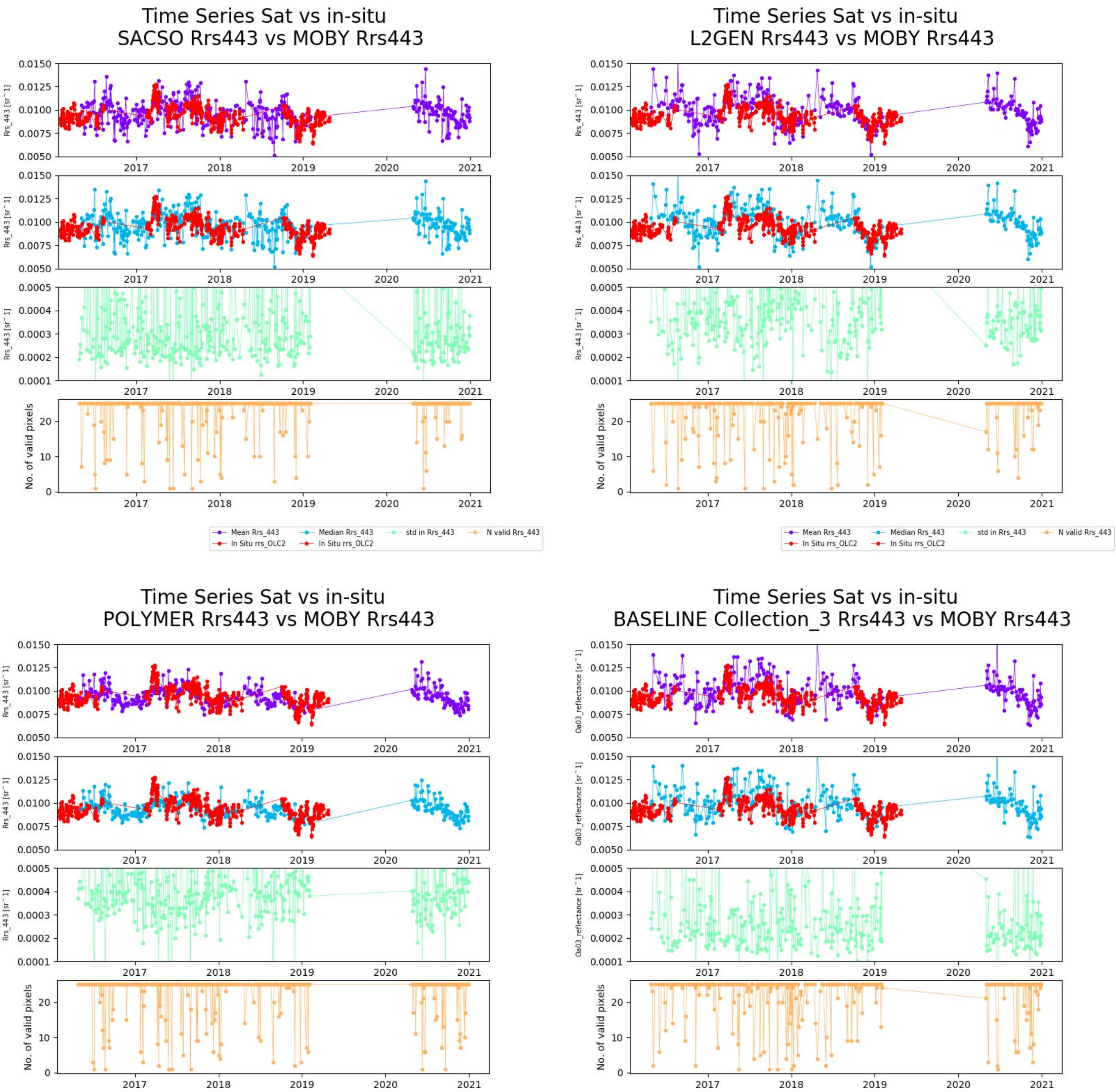


Figure 32: Example comparison of the extracted mean, median, standard deviation and number of valid pixels for Rrs 443 (or reflectance/pi for BASELINE) for SACSO, L2GEN, POLYMER and BASELINE processors. Data extracted from a 5x5 pixel area centred on the MOBY location.For the mean and median Rrs panels we also show the MOBY in-situ measurements of Rrs at the relevant OLCI bands for comparison.

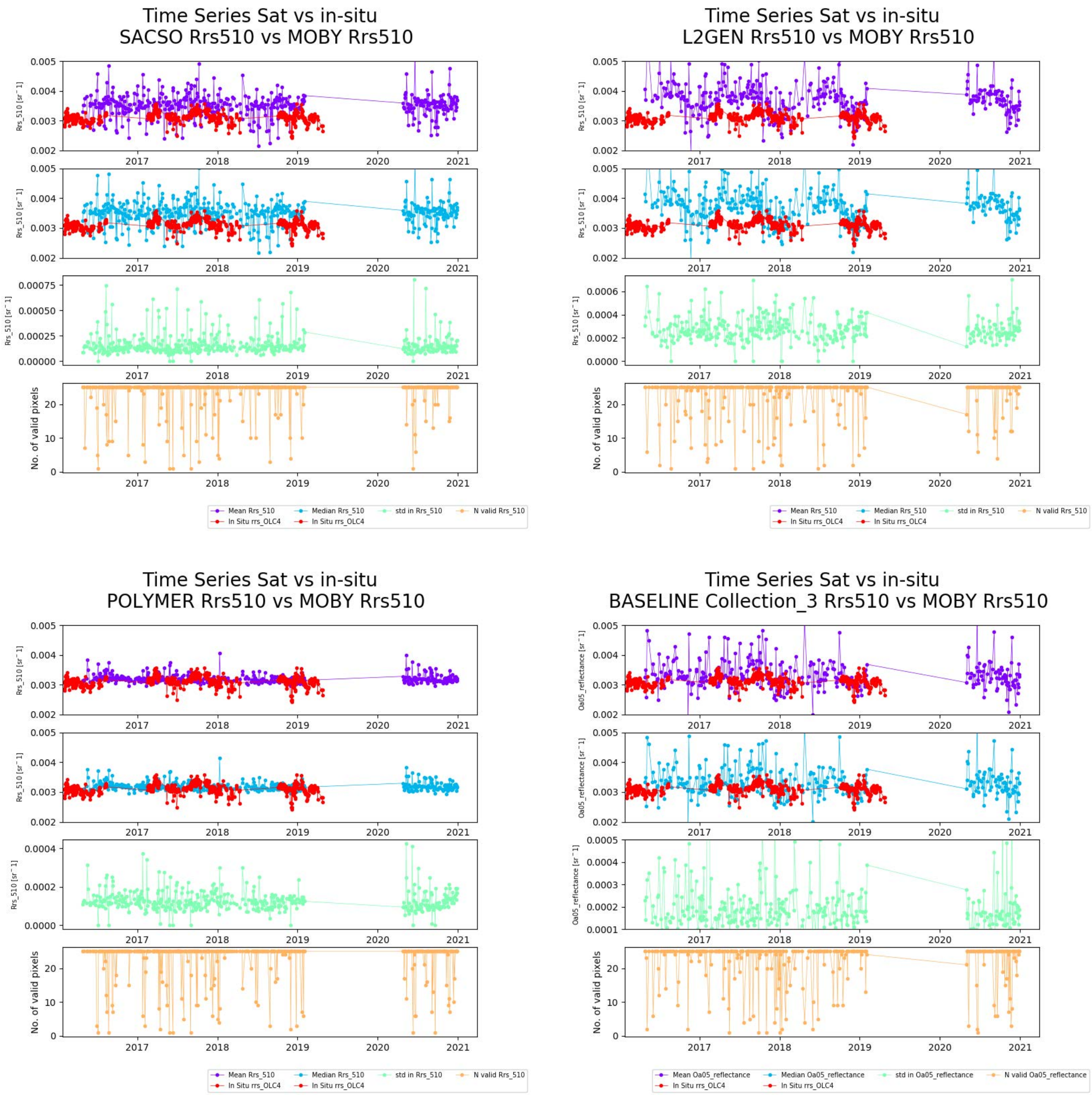


Figure 33: Example comparison of the extracted mean, median, standard deviation and number of valid pixels for Rrs 510 (or reflectance/pi for BASELINE) for SACSO, L2GEN, POLYMER and BASELINE processors. Data extracted from a 5x5 pixel area centred on the MOBY location. For the mean and median Rrs panels we also show the MOBY in-situ measurements of Rrs at the relevant OLCI bands for comparison.

5 Summary and recommendations

The example results presented above are a demonstration of the product assessment tools available withing the OMAPS processor. These analyses were performed with

- very limited in situ and diagnostic datasets,
- the scoring method that requires further development,
- tentative configurations for some processors, e.g. L2GEN,
- in situ radiometric datasets that can be further adjusted to OLCI nominal bands,

Therefore, the result interpretations should be considered as extremely preliminary and contextual, with the primary purpose being a demonstration of the OMAPS processor inter-comparison capabilities. The results will change when processors' configurations are updated following new information from relevant parties, such as a new L2GEN configuration from NASA for OLCI, and when validation datasets are extended to more representative global datasets.

Below, we allow a short discussion of the results presented in this document to give an example of the analysis when comprehensive and accurate comparisons are done in the future.

The results can suggest that there is no single optimal processor configuration as differing algorithms perform better under certain conditions. The POLYMER processor performs well in both the round robin intercomparisons and the time series analysis, but both of these are heavily biased towards clear ocean waters. Also, POLYMER did not score the highest in all round robin metrics, so there are clearly situations and metrics in which other processors could be considered superior. The diagnostic granule examination seems to support this conclusion, with POLYMER mostly showing 'worse' or anomalous performance relative to the other processors in extreme optical conditions. However, SACSO and BASELINE processors appear to provide more consistent retrieval fields on the diagnostic granules in complex atmospheric conditions and complex coastal waters, respectively. There are clear cases where the BASELINE and SACSO processors produce more coverage, or less suspicious data, than the POLYMER processor but there are also clear examples where retrievals are not sufficiently masked in SACSO/BASELINE configurations. The SACSO processor shows a lot of promise, and it may soon overtake POLYMER, but further development and assessment is needed. It would also be of great use if the SACSO processor could be made more time efficient as it currently takes approximately four times as long to process a granule as the other configuration options within the OMAPS system. L2GEN seems to have a very conservative masking scheme and rarely produces erroneous data but this comes at a significant sacrifice of 'good' data compared to that provided by the other processors. The comparison to time-series data near Hawaii further suggests that POLYMER or BASELINE processors may be performing the most optimal atmospheric processing. Yet, the details of the processing and the data match needs to be further investigated. The additional glint region coverage provided by POLYMER and SACSO is an additional benefit of these two processors as this represents a large increase in coverage when taken at the global level.

The in-water round robin suffers from a small amount of in-situ matchup data, meaning that drawing concrete conclusions is difficult. However, the results produced seem to mirror those found for a MERIS referenced OC-CCI dataset (V5.0) which included POLYMER processed OLCI and MERIS data. We might therefore conclude that using an OC-CCI style approach for OLCI processed with POLYMER, the blending of OC2, OCx, OCI and OCI2 algorithms could be the most optimal for the default set of optical water types.

In-Water Chlorophyll-a Algorithm	Optical water class memberships contributing to weighting
OCI	1
OCI2	2,3,4,5
OC2	7,8,9,10,11
OCx	6,12,13,14

Acknowledgements

This assessment was made using the OMAPS processing system and publicly available data. In addition to the support of the OMAPS team we would like to thank the following teams and organizations for their continued efforts in services such as data collection and provision that made this work possible.

- NOAA and the MOBY team for MOBY data.
- AERONET-OC teams at JRC and NASA for provision of the AERONET-OC data.
- NASA Ocean Color Project for SeaBASS in situ archive.
- PML team for Atlantic Meridional Transect cruise data and development of the OMAPS Online processor system.
- SOLVO for development and validation of OLCI OC-SVC gains.
- Brockmann Consult, GmbH, for OMAPS Offline processor module development and integration.
- HYGEOS and the SACSO team for their support in updating and integrating the POLYMER and SACSO processing configurations.
- FC.ID for their work in merging in-situ datasets.
- EUMETSAT Sentinel 3 team for their feedback and input on processor development and result interpretation.