PML | Plymouth Marine Laboratory

EUMETSAT

Ocean Colour Multi-Mission Algorithm Prototype System (OMAPS)

Algorithm Theoretical Baseline Document for AC-RR, Atmospheric Correction Round Robin

Ref: D2.1.3
Date:
Issue: 1.4
For: EUMETSAT
Ref: ID 995318

Table of Contents

Document Control

Version No.	Release Date	Author/Contributor	Reason for issue
1.0-draft	2019-12-01	Hajo Krasemann, Carsten Brockmann	internal review
1.1	2021-10-05	Dagmar Müller	update
1.2	2021-11-24	Dagmar Müller	Revision
1.3	2021-12-07	Dagmar Müller, Hajo Krasemann	Revision
1.4	2022-01-13	Dagmar Müller, Ewa Kwiatkowska	Revision

1 Purpose

A reliable and stable atmospheric correction (AC) is the basis for high quality ocean colour products. The development of new or improved ACs is still an ongoing issue and makes it necessary to compare them in an objective and repeatable manner.
This document describes the required inputs, the methods and the statistical results for the assessment of an optimal AC for a multi-channel ocean colour sensor.

The OMAPS Atmospheric Correction Round Robin (RR-AC) Module inherits its baseline from ESA's Ocean-Colour CCI project and the round robin procedure developed within that project. The main ideas of the Atmospheric Correction Round Robin for the Ocean-Colour CCI were published in a special issue of Remote Sensing of Environment in 2015 (Müller (2015)). More details can be found in OC-CCI's PVSAR (Müller (2015-2)).

To guarantee an objective selection from a set of different atmospheric correction processors, the common validation strategy of comparisons between in situ reference and satellite-derived water leaving reflectance spectra has been extended by a ranking system. In principle, the statistical parameters such as median absolute difference etc. and measures of goodness of fit e.g., spectral angle mapper (SAM), are analysed and used for evaluation. In addition, to get a better overview, they are transformed into relative scores, which evaluate the relationship of quality dependent on the algorithms under study. By reducing the manifold of statistical parameters to a single number, the score, it is possible to assess the sensitivity of these scores to the database of in situ reference measurements. A bootstrapping method varies the composition of the in-situ reference measurements and shows by that the uncertainty of the scoring results.

Together with the OMAPS Matchup Module, the RR-AC Module partly implements the recommendations of the S3 matchup protocols (EUMETSAT, 2021).

2 Datasets

2.1 Copernicus OCDB

For OMAPS, the Copernicus OCDB is (https://ocdb.eumetsat.int/) the required source for the in-situ measurements.
The in situ data gathered from OCDB follows protocols and quality standards as described in detail in section 'Preparing in situ data: other sources' in https://ocdb.readthedocs.io/en/latest/ocdb-MDB-user-manual.html#mdb-files-content.
This means that the content and available variables can be different for the given datasets.
Among the variables available from the OCDB are remote sensing reflectance, concentration of chlorophyll-a, inherent optical properties like phytoplankton absorption (a_{ph}), total absorption of detritus and gelbstoff (a_{dg}), backscattering of particles (bb_p) and the diffuse attenuation coefficient (k_d). Examples of datasets in the OCDB are shown in EUMETSAT, 2021d.
For the RR-AC described here, only remote sensing reflectance measurements are required.
The in situ data is bandshifted in the Matchup Module if differences between in-situ wavelength and central wavelength of the closest OLCI band are above 1 nm (see description in ATBD Matchup Module).
Water type specific grouping of in-situ data has to be done manually by the user, if needed.

2.2 OLCI L2 data - Atmospheric correction processors

The candidates for the atmospheric correction procedure have to be implemented in the online processor. There, the processing of potential matchup scenes, cut into minifiles, is undertaken. The Matchup Module as part of the offline processor will extract those macropixels, which fulfil the requirements in spatial availability of data and timeliness between in-situ measurement and overpass (see 'S3 Matchup protocol' in the Matchup Module ATBD).

All quality flags of the ACs have to be available for further processing, so that valid pixel expression can be applied. The AC's output should consist of remote sensing reflectances and quality flags.
If the RR-AC Module is applied to testing modifications in one AC (e.g., like varying the aerosol model), these modifications have to be implemented in the online processor. The RR-AC Module has no direct influence on the AC processing, but it will take any Matchup Module extractions as input.

The OMAPS RR-AC Module is equipped to handle matchup data extractions from the Matchup Module and by default four atmospheric corrections for S3-OLCI, i.e., POLYMER (Steinmetz (2011)), L2gen (Gordon and Wang (1994), Bailey et al. (2010)), SACSO (EUMETSAT 2021c) and the ground segment standard algorithm IPF (EUMETSAT 2021a, b). To incorporate other ACs, the initialisation file, which handles the processing parameters of the RR-AC Module, has to be adjusted and valid pixel expressions for other processors have to be added.

System vicarious calibration has to be applied during the generation of the OLCI L2 data, if needed, before creating matchup extractions with the Matchup Module.

Following the recommendations, a BRDF correction for S3-OLCI data is implemented in the Matchup Module as well. It is applied to IPF products if selected in the configuration of the Matchup Module (see ATBD Matchup Module).

3 AC Round Robin Protocol

Matchups are generated by extracting ocean colour data from a specific scene corresponding to appropriate in situ reference measurements, and by aggregating and filtering the data according to selected qualities.

The number of ACs (or versions of ACs) and their corresponding matchup extractions, which can be analysed in the RR-AC simultaneously, is not restricted. The minimum number is two, an upper limit of four or five is recommended so that diagnostic plots remain readable.

The default quality flagging is described in section 3.1 and the specific filters and the quality definition in section 3.2. The details of the available statistical parameters and the methodology behind the scoring are specified in section 3.3. An example of the RR-AC Module is given in section 3.4.

3.1 AC-processor quality flagging

The quality flags of each AC processor need to be complemented by a proper identification of cloudy and otherwise unusable pixels. It is important to note that the flagging can have a major impact on the final result of the RR-AC. It is recommended to use a common pixel screening for a given sensor and all ACs applied to the observations of that sensor. In the context of OMAPS, for the identification of pixels, Idepix (Kirches et al 2016, or SNAP S3 toolbox help) is used to identify land, cloud, snow/ice and mixed pixels. Idepix is available for most common ocean colour sensors beyond Sentinel 3 OLCI.

During the OLCI L2 processing with one of the ACs, the Idepix processor is also applied to the L1b data (TOA reflectances). The Idepix classification masks are included in the output of the Matchup Module, so that for each pixel the classification based on L1b data (top of atmosphere) and the quality description of processing results of the AC are returned for further application in the RR. The user can define valid pixel expressions from the Idepix classification and the AC quality flags. Idepix flags are available in the Polymer, sacso and l2gen products and extractions.

The recommended valid pixel expression of the Idepix classification is: not (IDEPIX_INVALID or IDEPIX_LAND or IDEPIX_CLOUD or IDEPIX_SNOW_ICE or IDEPIX_CLOUD_BUFFER or IDEPIX_CLOUD_SHADOW).

All atmospheric corrections provide their own sets of quality flags, which reflect their individual approaches to cloud masking and their respective range of applicability.

The recommended valid pixel expressions are listed in Table 1 for the four AC types. The Idepix flags can be added to Polymer, SACSO and L2gen valid pixel expressions.

Table 1 Valid pixel expressions for four ACs. If any flags listed as "exclude" is raised, the pixel becomes invalid. At least one of the "include" flags has to be raised, so that the pixel becomes valid.

Processor	**Valid pixel expression**
AC Polymer	**Exclude**: LAND, CLOUD_BASE, L1_INVALID, NEGATIVE_BB, OUT_OF_BOUNDS, EXCEPTION, THICK_AEROSOL, HIGH_AIR_MASS, EXTERNAL_MASK, dust_mask, IDEPIX_INVALID, IDEPIX_LAND, IDEPIX_CLOUD, IDEPIX_SNOW_ICE, IDEPIX_CLOUD_BUFFER, IDEPIX_CLOUD_SHADOW
AC SACSO	**Exclude**: LAND, CLOUD_BASE, L1_INVALID, NEGATIVE_BB, OUT_OF_BOUNDS, EXCEPTION, THICK_AEROSOL, HIGH_AIR_MASS, EXTERNAL_MASK, dust_mask, IDEPIX_INVALID, IDEPIX_LAND, IDEPIX_CLOUD, IDEPIX_SNOW_ICE, IDEPIX_CLOUD_BUFFER, IDEPIX_CLOUD_SHADOW
AC IPF	**Exclude**: 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 **Include**: WATER, INLAND_WATER
AC L2gen	**Exclude**: ATMFAIL, LAND, HIGHGLINT, HILT, HISATZEN, STRAYLIGHT, CLDICE, COCCOLITH, HISOLZEN, LOWLW, CHLFAIL, NAVWARN, MAXAERITER, CHLWARN, ATMWARN, NAVFAIL, IDEPIX_INVALID, IDEPIX_LAND, IDEPIX_CLOUD, IDEPIX_SNOW_ICE, IDEPIX_CLOUD_BUFFER, IDEPIX_CLOUD_SHADOW
L1 Pixel Identification Idepix	**Exclude:** IDEPIX_INVALID, IDEPIX_LAND, IDEPIX_CLOUD, IDEPIX_SNOW_ICE, IDEPIX_CLOUD_BUFFER, IDEPIX_CLOUD_SHADOW

Valid pixel expressions can be altered by the user in the configuration file. The OMAPS RR-AC Module is capable of interpreting two lists of existing flags, one naming the flags, which have to be raised so that the pixel becomes valid, the other list gathering all flags, which make the pixel invalid, if a single one of them is raised.

3.2 Aggregation and Filtering

3.2.1 Aggregation of macropixels, good match-up criteria

Satellite data screening used for matchup generation is a bit different than the method defined in EUMETSAT's OLCI Matchup Protocols and it will be adjusted in the future to align with the OLCI Protocols (EUMETSAT 2021, S3 matchup protocol). Here, a matchup point of good quality is created and selected by the following method:

- To each pixel in the macro-pixel, the defined AC flags are applied and only valid pixels are considered in the next steps (see next section 3.2.2).
- An outlier filter in the form of a standard deviation (σ) is applied to the remaining pixels per wavelength in the macro-pixel. If the pixels are within $\mu_\lambda - f \cdot \sigma_\lambda \leq Rrs_n(\lambda) \leq \mu_\lambda + f \cdot \sigma_\lambda$, with factor f = 1.5 as default, they are kept as valid. This filter is applied to each wavelength independently, as noise can be wavelength dependent.

- The number of valid pixels has to be larger than half the size of the micropixel (13 pixels in case of 5x5, 5 pixels in case of 3x3).

- Spatial homogeneity: If the number of valid pixels remains larger than half of the size of the macro-pixel, $N_{valid} > N_{macropixel}/2$, the mean μ and standard deviation σ of the remaining pixels is calculated and used to test for spatial homogeneity. If $\sigma_\lambda / \mu_\lambda < 0.15$, the macro-pixel is considered to be spatially homogeneous, and the mean of the remaining valid pixels is a good representative of the entire macro-pixel measurement. This is done for each wavelength independently.

If necessary, the spatial homogeneity criterion can be switched off in the configuration file.

3.2.2 Pixelwise filtering: CBQ and IBQ

For AC processor comparison relying on selected data, it is a common practise to restrict the comparison to identical data for all processor candidates where all quality restrictions of all candidate processors are obeyed. This we call the common best quality (CBQ). However, this approach achieves an incomplete picture. Therefore, tests are performed not only for a common set of common best quality but also for sets where the individual processor controls the accepted pixels, the individual best quality (IBQ). The latter is much closer to practical applications of the processors while the first approach gives a comparison of potential qualities.

Although the CBQ flags are the combination of all processor flags and for each processor the same pixels are chosen, the aggregated data points still may differ between different processors. During filtering for homogeneity of the 3x3 or 5x5 macropixel, outliers may or may not exist for a certain processor, so that the number of valid pixels may decrease below the limit for one processor but not for another due to processor specific noise in the data. The homogeneity criterion can lead to a different amount of available good match-up points, which is dependent on the noise (Figure 1).

One could make the criterion for common best quality stricter by demanding all pixels in a macro-pixel being accepted identical by all processor candidates, producing a strict common best quality. In OC-CCI this was tested and found not to change the results significantly but only cutting substantial in the number of accepted matchups (p.18ff). Therefore, the OMAPS RR-AC considers CBQ and IBQ only.

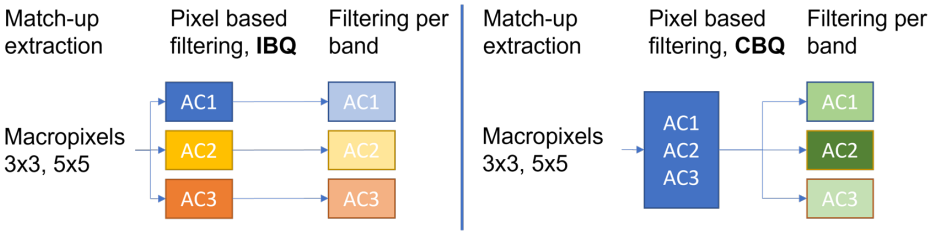

Figure 1: Overview of filtering steps in the RR-AC algorithm. After spatio-temporal filtering of match-up extraction, the macropixels are subjected to an AC dependent pixel-based filtering (IBQ, left), or the combined valid pixel expression of all ACs present are applied (CBQ, right). Afterwards outlier and spatial homogeneity tests on each macropixel is performed per band, which can lead to different numbers of valid matchups per band.

3.3 Statistics, scoring, and bootstrapping

3.3.1 Statistics

The following statistics are implemented and can be applied in the comparisons (following EUMETSAT, 2021). For measuring dispersion and bias per band, these measures can be chosen:

- Median Absolute Difference (MdAD); dispersion: $MdAD(\lambda) = median(|Rrs_{insitu}(\lambda) - Rrs_{OLCI}(\lambda)|)$
- Median Difference (MdD); bias: $MdD(\lambda) = median\left(Rrs_{insitu}(\lambda) - Rrs_{OLCI}(\lambda)\right)$
- Median Absolute Percentage Difference (MdAPD); dispersion:

$$MdAPD(\lambda) = 100 * median\left(\frac{|Rrs_{insitu}(\lambda) - Rrs_{OLCI}(\lambda)|}{Rrs_{insitu}(\lambda)}\right)$$

- Median Percentage Difference (MdPD); bias: $MdPD(\lambda) = 100 * median\left(\frac{Rrs_{insitu}(\lambda) - Rrs_{OLCI}(\lambda)}{Rrs_{insitu}(\lambda)}\right)$

The selection of the respective parameters derived with mean instead of median is still possible, but not recommended:

- Mean Absolute Difference (MAD); dispersion: $MAD(\lambda) = \frac{\sum|Rrs_{insitu}(\lambda) - Rrs_{OLCI}(\lambda)|}{n}$
- Mean Difference (MD); bias: $MD(\lambda) = \frac{\sum Rrs_{insitu}(\lambda) - Rrs_{OLCI}(\lambda)}{n}$
- Mean Absolute Percentage Difference (MAPD); dispersion: $MAPD(\lambda) = \frac{100 * \sum\frac{|Rrs_{insitu}(\lambda) - Rrs_{OLCI}(\lambda)|}{Rrs_{insitu}(\lambda)}}{n}$
- Mean Percentage Difference (MPD); bias: $MPD(\lambda) = \frac{100 * \sum\frac{Rrs_{insitu}(\lambda) - Rrs_{OLCI}(\lambda)}{Rrs_{insitu}(\lambda)}}{n}$

For comparison of spectral shapes, two measures are implemented:

- Spectral angle mapper (SAM): $SAM = \frac{1}{N}\sum(\text{acos}\frac{<Rrs_{insitu}, Rrs_{OLCI}>}{|Rrs_{insitu}||Rrs_{OLCI}|})$
- χ2 value: $\chi 2 = \frac{1}{N}\sum\sum\left(\frac{(Y_{insitu} - Y_{OLCI})^2}{Y_{insitu}}\right)$ with $Y = Rrs\ (\lambda)/Rrs\ (560nm)$.

For OLCI, seven wavelengths are considered in the spectral measures (412, 443, 490, 510, 560, 620, and 665, all in nm) by default. As some wavelengths are very rarely measured in situ, the SAM and χ2 values are based on the wavelengths which occur most frequently. Although we are targeting for 11 bands in OLCI spectral measures, adding 400, 754, 779, 865, and 1020, this decision is adjusted on the availability of in situ reference measurements.
The default set of statistical parameters are four per band (MdAD, MdD, MdAPD, MdPD) and the two spectral measures SAM and χ^2.
The confidence intervals for all the mean difference values (or median difference values) are approximated by the following formula:

$$CI = mean \pm t.ppf_{N-1,1-\alpha/2} * \frac{std}{\sqrt{N}}$$

where *N* is the number of samples; α is the confidence level where the 5% level is used, i.e. $\alpha = 0.05$; *std* is the standard deviation of the sample; t.ppf is the percent point function of a Student's t-distribution (percent point function is the inverse of a cumulative distribution function) for degree of freedom df=*N*-1 and confidence level value for a two-tailed test 1- $\alpha/2$ = 1-0.05/2=0.975.
The median or mean of the differences takes the place of the 'mean' in this formula. N is replaced by the number of valid matchups per band (due to the homogeneity criterion, the numbers can be slightly different for each band). Although the number of valid matchup might not be selected as a statistical parameter in the scoring (user's choice), it is implicitly present in the scoring procedure as the confidence intervals per band are governed by the numbers of valid matchups per band.
There are no confidence intervals calculated for the spectral statistical measures SAM and the χ2 value.

3.3.2 Scoring algorithm

The results of all statistical properties per single wavelength and the spectral tests need to be added up into a single value for easy comparison. Simple conversions are made using the statistical value itself and its standard error or 95% confidence interval (*Figure 2*). The statistical values are converted in the following way, so that the ranking of the values becomes easy: the closer to zero the value is, the higher should be its score. The best statistical value is closest to zero and receives the highest score.

All statistical measures derived on absolute differences are used as they are (MAD, MAPD or the respective medians of absolute differences); statistics, which yield both positive and negative values, are interpreted as absolute values (MD and MPD are interpreted as $|MD|$ and |MPD|), so that a simple ranking of the statistics can be applied. Confidence intervals are calculated for all algorithms and statistical measures, which are based on mean or median values (of differences).
For each statistical measure, the best algorithm ($AC_j, band_x$) has the value the closest to zero, after taking the absolute if necessary, e.g., lowest bias or dispersion, lowest values in spectral angle or χ2. This algorithm receives 2 points as a score for this statistical measure and the variable under study.

- If the value corresponding to another algorithm (same variable) 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.) All scores of different statistical properties per atmospheric correction are summed up per wavelength, each of them with equal weight, e.g., for the default selection of statistical parameters the score is

$$score(AC_i, band_x) = score_{MdAD}(AC_i, band_x) + score_{Md}(AC_i, band_x) + score_{MdAPD}(AC_i, band_x) + score_{MdPD}(AC_i, band_x).$$

The sum of scores per wavelength $score(AC_i, band_x)$ is scaled so that the sum over all atmospheric corrections in the comparison equals their number N_{AC} :

$$\sum_{i=0}^{N_{AC}} score_{scaled}(AC_i, band_x) = N_{AC}$$

If the number of valid matchups per band $N(AC_i, band_x)$ is selected as a statistical parameter, it is transformed directly into scores by normalising them by the sum of matchups over all ACs for this one band. More valid matchups lead to a larger score.

The spectral statistics (SAM, χ^2) are translated into scores in the following manner: first, the χ^2-values of all ACs are normalized:

$$\chi^2_{norm}(AC_i) = \chi^2(AC_i) / \sum \chi^2(AC_j).$$

Secondly, the χ^2_{norm} values between 0 and 1 are transformed, so that value closer to zero is than closer to one, which can be interpreted as the score is getting higher, if the values are closer to zero:

$$score_{\chi^2}(AC_i) = 1 - \chi^2_{norm}(AC_i)$$

The measure of spectral shape, i.e., the χ2 value and SAM, receive the same weight as any single band. Therefore, the spectral scores are scaled so that their sum over the ACs equals the number of ACs:

$\sum_{i=1}^{N_{AC}} score_{SAM,scaled}(AC_i) = N_{AC}$ and $\sum_{i=1}^{N_{AC}} score_{\chi^2,scaled}(AC_i) = N_{AC}$.

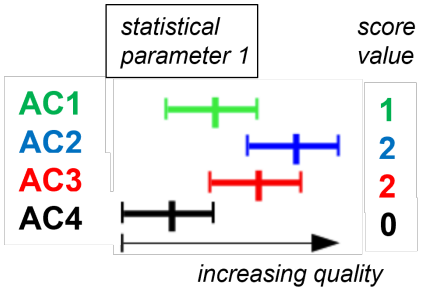

Figure 2: Example of the scoring scheme for one statistical parameter and four atmospheric corrections. The highest quality (blue) is followed by a not significantly lower (red), and significantly lower but with overlapping (green) and not overlapping error bars (black). This results in two points (blue), two points (red), one point (green), and no points for the least (black).

The total score for each atmospheric correction is added up from the scaled scores of each band and the scores of the spectral properties:

$$score_{total}(AC_i) = score_{SAM,scaled}(AC_i) + score_{\chi^2,scaled}(AC_i) + \sum_{x=1}^{N_{band}} score_{scaled}(AC_i, band_x)$$

If a single AC outperforms every other for each statistical parameter and each band and in the spectral goodness of fit, it will receive the maximum score value based on the default statistical parameters (two spectral parameters, combined scaled scores of N_{band}):

$$score_{total,MAX} = (2 + N_{band}) * N_{AC}$$

In favouring the best algorithm strongly, this scoring system tends to a non-linear behaviour.

3.3.3 Bootstrapping

The collection of in-situ data can impact the results of the statistics strongly. Maybe a particular optical water type is more often represented than another, and one AC might work better on data from oligotrophic than eutrophic conditions (probably by design). It is recommended to investigate the in-situ database thoroughly and pick a selection of in-situ spectra, which is suitable to the task.

The selection of in-situ spectra in the comparison yields one set of statistics for all ACs and the associated scores. This is the called later ‘single representation’. To analyse the influence of the selection on these results, the match-up data are resampled. Data points are randomly chosen from the population of match-up data, allowing for an unlimited number of repetitions of single data points. The resampled dataset has the same length as the original match-up dataset.

This bootstrapping constructs from the original ‘single representation’ a large number of slightly different representations and for these sets the statistics and subsequent scoring is performed correspondently resulting in a large set of slightly varying scoring results reflecting the variation due to the actual available match-up data.

The population of match-up data differs distinctively whether flag combinations of individual best quality (IBQ) or common best quality (CBQ) are used. In case of IBQ, all valid, aggregated match-up spectra based on the individual quality flagging of each AC, are combined. Some in-situ spectra might not have valid satellite match-up counterparts for all ACs, but they are still part of the IBQ population, which is used in the bootstrapping.

In case of CBQ, the aggregation of each macropixel is based on the same set of pixels for each AC, as all valid pixel expressions are combined. Still, the criterion of spatial homogeneity can filter out some bands or spectra for some of the ACs, so that the number of good matchups can vary between ACs even with the CBQ pixel selection. It is more likely, that all in-situ spectra have satellite match-up counterparts from all ACs.

Calculating statistics and scores for the resampled datasets provides information on the distribution of those statistical parameters and the scores. Histograms of these distributions are stored after the RR-AC processing. They allow for a more balanced analysis of the (total) scores. E.g., their histograms may show, that the score distributions of different ACs overlap strongly, so that the presumably "best" AC algorithm, which the single representation of the matchup data supports, might only be the best for some of the representations of the matchup data.

3.4 Atmospheric Correction Round Robin Module

With output files from the OMAPS Matchup Module and a configuration file, the RR-AC process can be started (Figure 3). The RR-AC Module has been coded in python.
After applying the flags, aggregating the valid pixels to their mean values per band and filtering them for spatial homogeneity (if asked for), the selected statistics are applied and converted into scores.

Two types of results are available:

- tabulated statistical measures, scores and scatter plots generated from the standard full set of matchups with no randomisation in the data selection, so no bootstrapping. This standard full set of matchups is also called 'single representation' of the matchup dataset because no duplication of data is present,
- distributions of statistical measures and scores generated as histograms and boxplots from the bootstrapping method. The bootstrapping uses randomised selection of reference data that results in duplication of some matchups.

In the bootstrapping, the statistics and scores are calculated on the resampled (randomised with repetition) dataset, the results are stored in separate files for each resampled representation of the matchup data. From these values, diagnostic plots like histograms of the scores and boxplots per band and AC of statistical parameters are generated. The figures from the bootstrapping evaluation and tables of statistics and scores from the entire data set (no randomisation) are collected automatically in the LaTeX documentation. It can be adjusted by hand to the needs of the user. Ideally, the histograms of scores from the bootstrapping show a normal distribution for each AC. These histograms can be interpreted by the user, but there is no automated evaluation of the score distribution in place. The mean score from the distribution can deviate from the tabulated score, which has been derived from the standard full unrandomized set of match-up data. It is recommended to interpret the standard full tabulated values always in connection with the bootstrapping distributions.

From the standard full unrandomized match-up data (referred to as 'single representation', because no duplication of data is present) the scatterplots per wavelength are created.

The following example is derived from a very small set of eleven match-up extracts and it is designed to illustrate the inputs and outputs of the RR-AC Module.

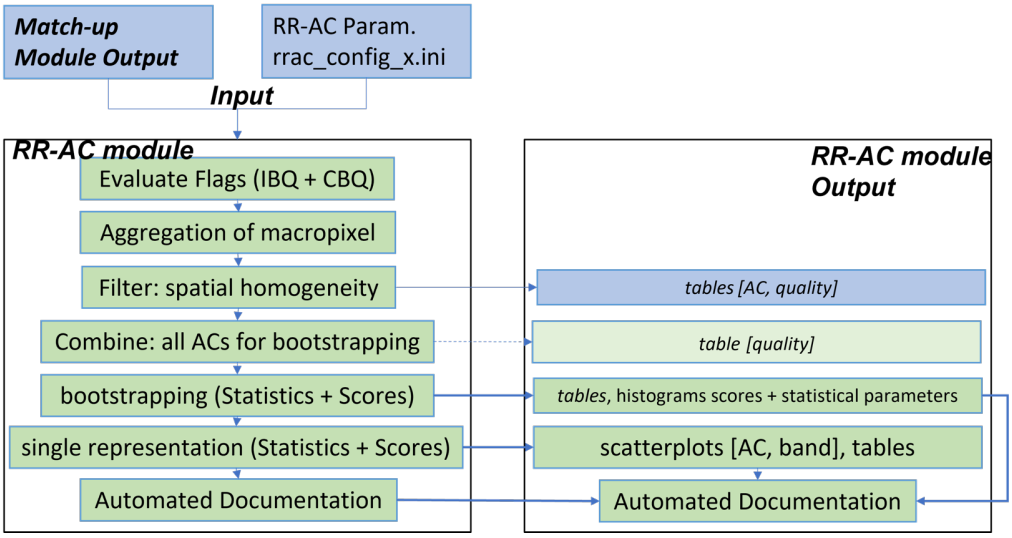

Figure 3: AC Round Robin Module overview.

3.4.1 Inputs

Four match-up extraction files holding data from macropixels and AC quality flags are located in the same folder (see Figure 4: mdb_input_dir = path/to/matchup-data-folder, and the filenames of the AC matchup extractions mdb_input_files_ipf, mdb_input_files_l2gen, etc.). Each one is filtered by its respective flags and the valid pixels are aggregated.

The set of recommended valid pixel expressions is given in section 3.1, they can be set and changed according to the user's needs in the configuration file (Figure 4). E.g., for the POLYMER AC all flags which make a pixel valid when raised, have to be listed in polymer_bits_include, while all flags, which render a pixel invalid when raised, are listed in polymer_bits_exclude.[1]

The spatial homogeneity check of the valid pixels per macropixel (and per band) is switched off (configuration parameter check_homogen=False in Figure 4), because due to noise in the IPF data the bands starting at 620nm are excluded from the analysis otherwise.

The default set of statistical parameters is converted into scores (see Figure 4, statistical_parameters and score_parameters). The aggregated data of the four ACs is combined for the bootstrap part of the processing.

1 The valid pixel expressions in this example are outdated, but the example is kept as it serves its purpose of a simple illustration of generated outputs of the RR-AC Module. The valid pixel expressions are:

- POLYMER: !LAND & !CLOUD_BASE & !L1_INVALID & !NEGATIVE_BB & !OUT_OF_BOUNDS & !EXCEPTION & !THICK_AEROSOL & !HIGH_AIR_MASS & !EXTERNAL_MASK & !CASE2 & !INCONSISTENCY
- SACSO: !BAD
- IPF: !CLOUD & !CLOUD_AMBIGUOUS & !CLOUD_MARGIN & !INVALID & !COSMETIC & !SATURATED & !SUSPECT & !HISOLZEN & !HIGHGLINT & !SNOW_ICE & !AC_FAIL & !WHITECAPS & !ADJAC & !OC4ME_FAIL & !RWNEG_O2 & !RWNEG_O3 & !RWNEG_O4 & !RWNEG_O5 & !RWNEG_O6 & !RWNEG_O7 & !RWNEG_O8 & WATER
- L2GEN: !ATMFAIL & !LAND & !CLDICE & !SEAICE

```
[RR_AC]
; configuration for RoundRobin AC
mdb_input_dir = path/to/matchup-data-folder
mdb_input_files_ipf = MDB_S3A_OLCI_L2_OCDB_PML_L2_RRS_IPF_FULL_TEST.csv
mdb_input_files_l2gen = MDB_S3A_OLCI_L2_OCDB_PML_L2_RRS_L2GEN.csv
mdb_input_files_polymer = MDB_S3A_OLCI_L2_OCDB_PML_L2_POLY_TEST.csv
mdb_input_files_sacso = MDB_S3A_OLCI_L2_OCDB_PML_L2_SACSO_TEST.csv
rrac_output_dir_root = path/to/output-folder
processing_label = RRAC_4ACs_new
flag_type = IBQ
rrac_rho_type = rrs
insitu_rho_type = rrs
satellite_rho_type = rrs
check_homogen = False
write_bootstrap_data = True
n_bootstrap = 300
statistical_parameters = MdAD,MdD,MdAPD,MdPD,SAM,CHI2
chi2_insitu_bands = rrs_412,rrs_443,rrs_490,rrs_560.5796,rrs_665
chi2_insitu_band_norm = rrs_560.5796
score_parameters = MdAD,MdD,MdAPD,MdPD,SAM,CHI2
read_aggregated_data = False
polymer_bits_exclude = bitmask.LAND,bitmask.CLOUD_BASE,bitmask.L1_INVALI
polymer_bits_include =
sacso_bits_exclude = flags.BAD
sacso_bits_include =
ipf_bits_exclude = WQSF.CLOUD,WQSF.CLOUD_AMBIGUOUS,WQSF.CLOUD_MARGIN,WQS
ipf_bits_include = WQSF.WATER
l2gen_bits_exclude = l2_flags.ATMFAIL,l2_flags.LAND,l2_flags.CLDICE,l2_f
l2gen_bits_include =
```

Figure 4: Configuration file of the test example. The spatial homogeneity check of the valid pixels per macropixel (and per band) is switched off (against recommendations), because due to noise in the IPF data the bands starting at 620nm are excluded from the analysis. Valid pixel expressions are partially given and refer to an older stage of flag development. Recommended valid pixel expressions can be found in section 3.1.

3.4.2 Results

The results presented here are only a demonstration of the AC RR method and were created with a small dataset of 11 in situ reference points. The readers are referred to the OMAPS Product Validation Report for the detailed description of algorithm validations.

In the end, the RR-AC Module generates statistic tables and diagnostic plots to analyse the quality of the ACs.

The tables with statistics and scores are derived from the single representation of the matchup data, which consists of all data points, no repetition. Per band statistics can be found in Figure 5, the spectral statistics are summarized in Figure 6, and the associated scores of these statistics are collected in Figure 7, including the total score. All these results are gathered from the unrandomized data, so results of the bootstrap calculations are not shown here.

A collection of scatterplots is created showing remote sensing reflectances versus insitu reflectances per band for one AC (here IPF in Figure 8).

During bootstrap analysis the resampling allows the duplication of the data, though the total length of the dataset is constant. For each resampled representation of the matchup population the set of statistical parameters is calculated and stored, and they are transformed into scores as well. These scores are summarized as boxplots or histograms per reflectance band based statistical variable. The bootstrapping results in a distribution of the scores and there is no combination of bootstrapping results into a single score. An example is given for the median absolute difference (boxplot in Figure 9, histograms in Figure 10). In the same manner, histograms for the spectral statistical values (SAM and χ2) are generated (Figure 11). From all resampled matchup datasets, the statistics and their translation into scores are derived.

The distribution of scores is presented as a histogram plot (Figure 12). The scores depend on the selected bands, on the selected statistics, but also quite fundamentally on the valid pixel expressions. In this case, the maximum score value is

$$score_{total,MAX} = (2 + N_{band}) * N_{AC} = 9 * 4 = 36$$

With seven bands, two spectral statistics and four ACs.

An example of conversion of statistics into scores is given for the band at 412nm (Table 1). 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 2). 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_06.11 with 0.00153 is larger than this upper limit of the confidence interval of the best result. The lower limit of its confidence interval is 0.00153-0.000867=0.000661 (ipf_06.11), which falls below the upper limit of the best AC confidence interval. Therefore, ipf_06.11 receives 1 point in the scoring. The same reasoning applies to MdPD.

In a second step, each set of scores for one statistical parameter is scaled by their column sum, so that each statistical parameter gains equal weight, when the scores are summed up (Table 3). 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 χ^2 (Figure 7). An example is given, how the χ^2 values are

directly translated into scores (Table 4). In this fashion, the spectral statistics receive as much weight in the total sum of scores of each band.

Table 1: 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_collec-tion_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 2: 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 3: 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 4: 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	N	MdAD	MdD	MdAPD	MdPD
polymer_4.17	rrs_412	12	0.0013	0.0008	10.295	6.0207
polymer_4.17	rrs_443	12	0.0008	0.0005	8.4636	4.8989
polymer_4.17	rrs_490	12	0.0003	0.0002	5.6421	2.4895
polymer_4.17	rrs_510	12	0.0006	0.0006	16.0204	16.0204
polymer_4.17	rrs_560	12	0.0001	-0.0	4.563	-1.9067
polymer_4.17	rrs_620	12	0.0	-0.0	24.2738	-24.2738
polymer_4.17	rrs_665	12	0.0	-0.0	14.6296	-7.6753
sacso_1.0	rrs_412	12	0.001	0.0002	6.3306	1.1287
sacso_1.0	rrs_443	12	0.0006	-0.0001	5.267	-0.7713
sacso_1.0	rrs_490	12	0.0008	-0.0008	11.6739	-11.6739
sacso_1.0	rrs_510	12	0.0002	-0.0002	6.0646	-4.5488
sacso_1.0	rrs_560	12	0.0003	-0.0003	24.4727	-24.4727
sacso_1.0	rrs_620	12	0.0001	-0.0001	77.8496	-77.8496
sacso_1.0	rrs_665	12	0.0001	-0.0001	143.4629	-143.4629
ipf_collection3	rrs_412	8	0.0015	0.0015	10.8419	10.8419
ipf_collection3	rrs_443	8	0.0007	0.0007	7.1677	7.1677
ipf_collection3	rrs_490	8	0.0003	-0.0001	4.8694	-1.6307
ipf_collection3	rrs_510	8	0.0003	0.0003	9.8445	9.8445
ipf_collection3	rrs_560	8	0.0001	0.0	5.2775	0.552
ipf_collection3	rrs_620	8	0.0	0.0	14.1563	9.6598
ipf_collection3	rrs_665	8	0.0	0.0	29.0479	29.0479
l2gen_9.5.1-V2021.2	rrs_412	10	0.0016	-0.001	9.5806	-6.4656
l2gen_9.5.1-V2021.2	rrs_443	10	0.0009	-0.0003	7.4702	-2.8587
l2gen_9.5.1-V2021.2	rrs_490	10	0.0005	-0.0005	7.5909	-6.9606
l2gen_9.5.1-V2021.2	rrs_510	10	0.0002	-0.0	6.7515	-1.1929
l2gen_9.5.1-V2021.2	rrs_560	10	0.0003	-0.0003	27.2831	-27.2831
l2gen_9.5.1-V2021.2	rrs_620	10	0.0002	-0.0002	100.8907	-100.8907
l2gen_9.5.1-V2021.2	rrs_665	10	0.0001	-0.0001	82.6084	-82.6084

Table A.1: Statistics: IBQ OLCI

Figure 5: 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	0.3576	0.0146
sacso_1.0	0.2551	0.0418
ipf_collection3	0.3996	0.0419
l2gen_9.5.1-V2021.2	0.3992	0.0225

Table A.2: Chi-square values: IBQ OLCI

Figure 6: 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	1.23	1.07	1.37	0.0	2.0	2.0	2.67	1.0	1.17	12.51
sacso_1.0	1.23	1.07	0.29	1.7	0.0	0.0	0.0	1.09	0.87	6.25
ipf_collection3	0.67	0.79	1.37	0.6	2.0	2.0	1.33	0.96	0.87	10.58
l2gen_9.5.1-V2021.2	0.87	1.07	0.97	1.7	0.0	0.0	0.0	0.96	1.08	6.65

Table A.3: Scores: IBQ OLCI

Figure 7: Automatically generated table in a LaTeX document of the scores assigned to the statistics for the single representation of the matchup data.

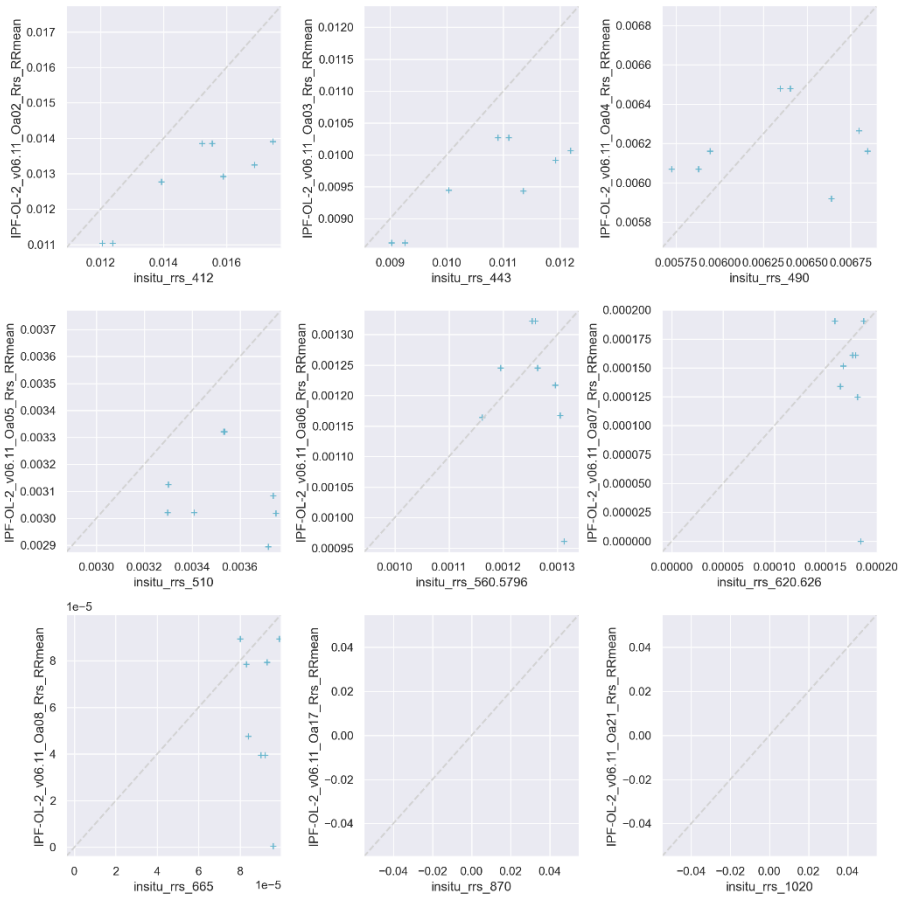

Figure 8 Example of scatterplots per band of aggregated satellite data versus insitu reflectances for one AC (here: IPF Collection 3).

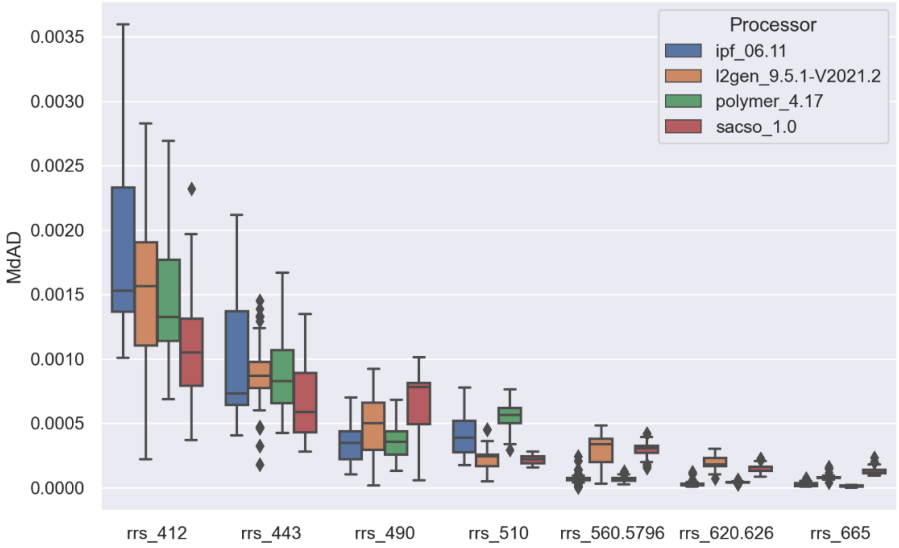

Figure 9 Boxplot of median absolute differences MdAD derived from all results of the bootstrap analysis. This visualisation of variance in the statistics depending on the composition of the matchup data is created for all statistical properties which have been selected in statistical parameters in the configuration file. (Note: ipf_06.11 translates into IPF Collection 3)

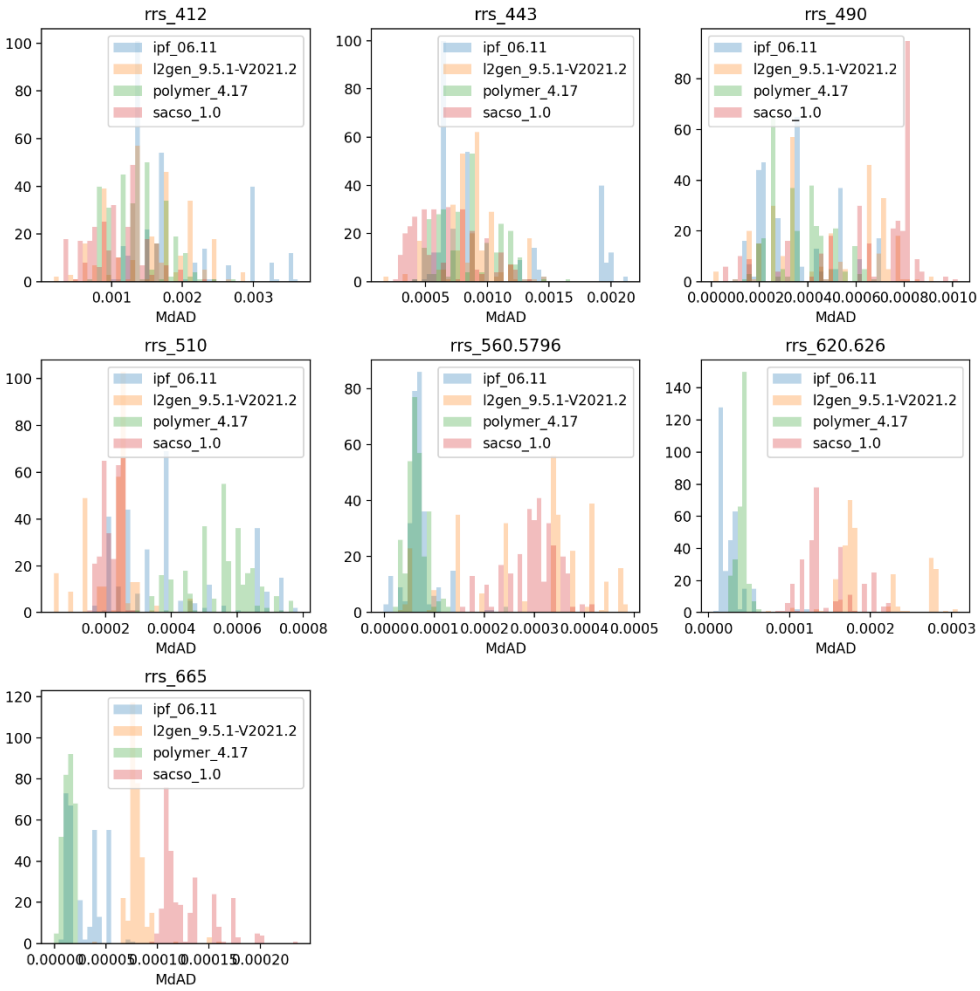

Figure 10: Distributions of Median Absolute Difference (MdAD) per band highlighting the variability of statistical parameters due to resampling of the matchup data in the bootstrap analysis. For each statistical parameter this kind of overview is created. (Note: ipf_06.11 translates into IPF Collection 3).

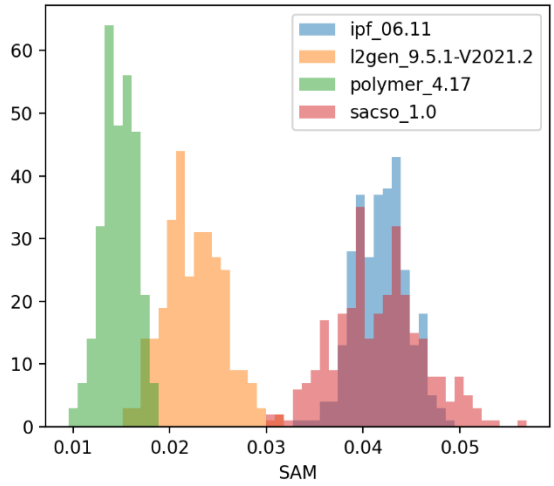

Figure 11 Distribution of spectral statistics SAM due to its variability caused by resampling the matchup data. (Also available for chi-square values). (Note: ipf_06.11 translates into IPF Collection 3)

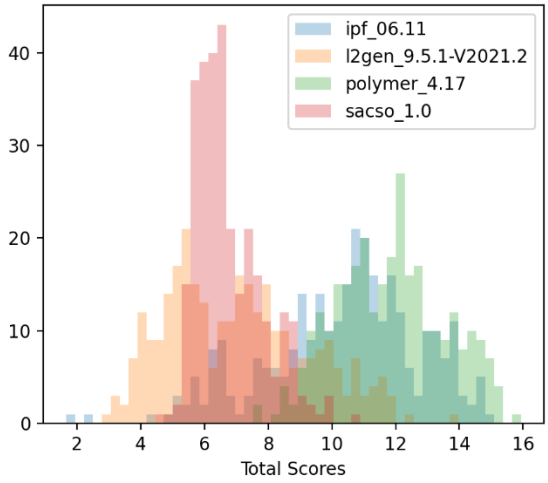

Figure 12 Distribution of total sum of scores derived from four statistical parameters per band and two spectral statistics. (Note: ipf_06.11 translates into IPF Collection 3)

4 References

Bailey et al. (2010)
S.W. Bailey, B.A. Franz, P.J. Werdell. Estimation of near-infrared water-leaving reflectance for satellite ocean color data processing. Opt. Express, 18 (7) (2010), pp. 7521-7527

Brewin (2015)
R.J.W. Brewin, S. Sathyendranath, D. Müller, H. Krasemann, R. Doerffer, F. Mélin, C Brockmann, N. Fomferra, M. Peters, M. Grant, F. Steinmetz, P.-Y. Deschamps, J. Swinton, T. Smyth, P.J. Werdell, B. A. Franz, S. Maritorena, E. Devred, Z. Lee, Ch. Hu, and P. Regner. The ocean colour climate change initiative III: a round-robin comparison on in-water bio-optical algorithms. Remote sens. Environ., Volume 162, 1 June 2015, Pages 271–294

Brockmann (2011)
Carsten Brockmann, Ana Ruescas, Kerstin Stelzer. IDEPIX ATBD
https://esa-oceancolour-cci.org/sites/esa-oceancolour-cci.org/alfresco.php?file=00d1ee37-8259-4885-891f-7985ff354152&name=OC-CCI-PixelIdentification-ATBD-v1.0_signed.pdf

EUMETSAT (2021).
Recommendations for Sentinel-3 OLCI Ocean Colour product validations in comparison with in-situ measurements - Matchup Protocols EUM/SEN3/DOC/19/1092968. Version 7. https://www.eumetsat.int/media/44087

EUMETSAT (2021a)
Sentinel-3 Product Notice – OLCI Level-2 Ocean Colour. https://www.eumetsat.int/media/48139

EUMETSAT (2021b)
Sentinel-3 OLCI L2 report for baseline collection OL_L2M_003. https://www.eumetsat.int/media/47794

EUMETSAT (2021c)
Spectral matching Atmospheric Correction for Sentinel Ocean colour measurements (SACSO).
https://www.eumetsat.int/SACSO

EUMETSAT (2021d)
Ocean Colour Multi-Mission Algorithm Prototype System (OMAPS): Input Output Data Description. Issue 2.0, 15 October 2021

Gordon and Wang (1994)
H.R. Gordon, M. Wang. Retrieval of water-leaving radiance and aerosol optical thickness over the oceans with SeaWiFS: a preliminary algorithm. Appl. Opt., 33 (3) (1994), pp. 443-452

G. Kirches, J. Militzer, M. Böttcher, C. Brockmann, and P. Defourny (2016).
ESA CCI Land Cover (Phase 2): Algorithm Theoretical Basis Document – Part II: Preprocessing., Version 1.0, 2 May 2016.

Lee and Hu (2006)
Zhong Ping Lee and Chuanmin Hu. Global distribution of Case-1 waters: An analysis from SeaWiFS measurements. Remote Sensing of Environment, 101:270-276, 2006.

Lee, Z.-P., K.L. Carder, R.A. Arnone (2002)
Deriving inherent optical properties from water color: A multiband quasi-analytical algorithm for optically deep waters. Appl. Opt. 41(27): 5755-5772. https://doi.org/10.1364/AO.41.005755.

Lee, Z.-P., B. Lubac, P.J. Werdell, R.A. Arnone (2009)
An update of the Quasi-Analytical Algorithm (QAA v5). Tech. Rep. International Ocean-Colour Coordinating Group (http://www.ioccg.org/groups/software.html).

Morel and Gentili (1996)
A. Morel and B. Gentili. (1996) Diffuse reflectance of oceanic waters. iii. implication of bidirectionality for the remote-sensing problem. Appl. Opt. , 35:4850#4862.

Morel and Maritorena (2001)
André Morel and Stéphane Maritorena. Bio-optical properties of oceanic waters: A reappraisal. Journal of Geophysical Research , 106(C4):7163-7180, 2001.

Müller (2015)
Dagmar Müller, Hajo Krasemann, Robert J.W. Brewin, Carsten Brockmann, Pierre-Yves Deschamps, Roland Doerffer, Norman Fomferra, Bryan A. Franz, Mike G. Grant, Steve B. Groom, Frédéric Mélin, Trevor Platt, Peter Regner, Shubha Sathyendranath, François Steinmetz, John Swinton. The Ocean Colour Climate Change Initiative: I. A methodology for assessing atmospheric correction processors based on in-situ measurements. Remote sens. Environ., Volume 162, 1 June 2015, Pages 242–256

Müller (2015-1)
Dagmar Müller, Hajo Krasemann, Robert J.W. Brewin, Carsten Brockmann, Pierre-Yves Deschamps, Roland Doerffer, Norman Fomferra, Bryan A. Franz, Mike G. Grant, Steve B. Groom, Frédéric Mélin, Trevor Platt, Peter Regner, Shubha Sathyendranath, François Steinmetz, John Swinton. The Ocean Colour Climate Change Initiative: II. Spatial and temporal homogeneity of satellite data retrieval due to systematic effects in atmospheric correction processors. Remote sens. Environ., Volume 162, 1 June 2015, Pages 257–270

Müller (2015-2)
Dagmar Müller, Hajo Krasemann. Product Validation and Algorithm Selection Report, Part 1 -AC, v-3.0
https://esa-oceancolour-cci.org/sites/esa-oceancolour-cci.org/alfresco.php?file=d311392c-d6a5-4a34-9de1-353e90b3f943&name=PVASR_Part1-AC-v3_20151202.pdf

Park and Ruddick (2005)
Young-Je Park and Kevin Ruddick. Model of remote-sensing reflectance including bidirectional effects for case 1 and case 2 waters. APPLIED OPTICS , 44(7), 2005.

The Sentinel Application Platform (SNAP). http://step.esa.int/main/toolboxes/snap/

Steinmetz (2011)
François Steinmetz, Pierre-Yves Deschamps, and Didier Ramon. Atmospheric correction in presence of sun glint: application to MERIS. Optics Express Vol. 19, Issue 10, pp. 9783-9800 (2011) •https://doi.org/10.1364/OE.19.009783

Valente (2015/2016)
A. Valente et al. (2015): A compilation of global bio-optical in situ data for ocean-colour satellite applications. doi:10.1594/PANGAEA.854832

Valente, A. et al. (2016): A compilation of global bio-optical in situ data for ocean-colour satellite applications, Earth Syst. Sci. Data, 8, 235-252, doi:10.5194/essd-8-235-2016, 2016.