Mueller, D.; Krasemann, H.; Brewin, R.J.W.; Brockmann, C.; Deschamps, P.-Y.; Doerffer, R.; Fomferra, N.; Franz, B.A.; Grant, M.G.; Groom, S.B.; Melin, F.; Platt, T.; Regner, P.; Sathyendranath, S.; Steinmetz, F.; Swinton, J.:

The Ocean Colour Climate Change Initiative: I. A methodology for assessing atmospheric correction processors based on in-situ measurements

In: Remote Sensing of Environment (2015) Elsevier

DOI: 10.1016/j.rse.2013.11.026

The Ocean Colour Climate Change Initiative: I. A methodology for assessing atmospheric correction processors based on in-situ measurements

Dagmar Müller[a,*], Hajo Krasemann[a], Robert J.W. Brewin[b], Carsten Brockmann[c], Pierre-Yves Deschamps[d], Roland Doerffer[a], Norman Fomferra[c], Bryan A. Franz[e], Mike G. Grant[b], Steve B. Groom[b], Frédéric Mélin[f], Trevor Platt[b], Peter Regner[g], Shubha Sathyendranath[b], François Steinmetz[d], John Swinton[h]

[a] *Helmholtz-Zentrum Geesthacht, Germany*
[b] *Plymouth Marine Laboratory, UK*
[c] *Brockmann-Consult, Germany*
[d] *HYGEOS, France*
[e] *NASA, Ocean Biology Processing Group, USA*
[f] *European Commission - Joint Research Centre, Italy*
[g] *European Space Agency, Italy*
[h] *Telespazio VEGA UK*

Abstract

The Ocean Colour Climate Change Initiative intends to provide a long-term time series of ocean colour data and investigate the detectable climate impact. A reliable and stable atmospheric correction procedure is the basis for ocean colour products of the necessary high quality. In order to guarantee an objective selection from a set of four atmospheric correction processors, the common validation strategy of comparisons between in-situ and satellite-derived water leaving reflectance spectra, is extended by a ranking system. In principle, the statistical parameters such as root mean square error, bias, etc. and measures of goodness of fit, are transformed into relative scores, which evaluate the relationship of quality dependent on the algorithms under study. The sensitivity of these scores to the selected database has been assessed by a bootstrapping exercise, which allows identification of the uncertainty in the scoring results. Although the presented methodology is intended to be used in an algorithm selection process, this paper focusses on the scope of the methodology rather than the properties of the individual processors. [1]

Keywords: OC-CCI, CCI, Ocean-Colour, Climate Change, atmospheric correction, algorithm comparison, in-situ comparison, spectral comparison, boot strap

1. Introduction

Ocean-colour is recognised as an Essential Climate Variable (ECV) by the Global Climate Observation System GCOS-154 (2011). Many geophysical and bio-optical variables retrieved from

*Corresponding Author
Email address: dagmar.mueller@hzg.de (Dagmar Müller)

[1] This article is published in Remote Sensing of Environment (2015), http://dx.doi.org/10.1016/j.rse.2013.11.026.
Please cite as: Müller, D., et al., The Ocean Colour Climate Change Initiative: I. A methodology for assessing atmospheric correction processors based on in-situ measurements

ocean-colour data from satellites, such as chlorophyll concentration and inherent optical properties of the ocean, are relevant to climate research. All these products are derived from spectrally-resolved water-leaving radiances or reflectances, which are extracted from top-of-the-atmosphere radiance values measured by satellites using atmospheric-correction algorithms. Given that the atmospheric signal is typically 80% or more of the total signal at the top of the atmosphere, accurate Atmospheric Correction (AC) is key to a successful implementation of all in-water algorithms in routine use today.

Currently, several algorithms and approaches are used by space agencies for atmospheric correction of ocean-colour data. For example, National Aeronautics and Space Administration (NASA) uses the SeaDAS (SeaWiFS Data Analysis System, current version 6.3, SeaWIFS: Sea Wide Field-of-view Sensor) processor, based on the algorithm of Gordon and Wang (1994) with several subsequent modifications and improvements (IOCCG (2010)). Initially developed for processing data from NASA sensors such as the Coastal Zone Color Scanner (CZCS), SeaWiFS, and the Moderate resolution Imaging Spectroradiometer (MODIS), the SeaDAS processor has now also been extended to incorporate additional sensors. For the Medium Resolution Imaging Spectrometer (MERIS), the European Space Agency (ESA) uses the "MERIS Instrument Processing Facility" (IPF), whose latest version is 6.04; equivalent to the MEGS-8 (MERIS Ground Segment data processing prototype version 8.0). The implementation of the algorithm based on Antoine and Morel (1999) will be further noted as MEGS (see Bourg (2012)). Both the MEGS and the SeaDAS processors rely on the satellite signal in the near-infrared wavelengths to infer the optical properties of atmospheric aerosol, which are then extrapolated into the visible domain to implement the atmospheric correction in those wavelengths.

Alternate algorithms have also emerged that use both visible and near infrared wavebands for atmospheric correction, using techniques such as neural networks (Schiller and Doerffer (1999)), spectral optimisation methods (Chomko and Gordon (1998); Chomko and Gordon (2001); Chomko et al. (2003); Steinmetz et al. (2011)) and spectral matching methods (Gordon et al. (1997)).

The performance of the atmospheric correction algorithms is evaluated in a point-by-point comparison of normalised water leaving reflectances derived from satellite with in-situ measurements close in time and space, so called "match-ups". The analysis presented here, is confined to MERIS satellite data and match-up data from the "MERis MAtchup In-situ Database" (MERMAID). In order to define an objective selection process which identifies the most suitable AC processor, a methodology for in-situ comparisons is developed, which converts statistical parameters and their confidence intervals as representations of product quality into a relative score per processor. The influence of the match-up data selection on the scoring results is investigated. The stability and error of the scoring system is tested with the help of the bootstrap method and the results are discussed.

2. Preparation of in-situ data and satellite data with candidate processors

2.1. In-situ site selection

MERMAID has been created to allow an easy access to match-up data which combines normalised water leaving reflectances measured in-situ and derived from MERIS satellite data. The water leaving reflectance ρ is defined as (Eq. 1, Antoine and Morel (1998))

$$\rho\left(\lambda,\theta_v,\theta_s,\Delta\phi\right)=\pi L_w\left(\lambda,\theta_v,\theta_s,\Delta\phi\right)/E_s\left(\lambda\right)\cos\left(\theta_s\right) \tag{1}$$

with wavelength λ, sun zenith angle θ_s, viewing angle θ_v, azimuth angle difference $\Delta\phi$, water leaving radiance L_w and irradiance $E_s\left(\lambda\right)$. By normalisation the radiometric instances are

Table 1: In-situ dataset bands in comparison to MERIS central bands (extract from MERIS Optical Measurement Protocols, Issue 2, Table 2-2). Shown as italic are wavelengths of in-situ measurements which differ from the MERIS bands by 5 nm or more and which undergo a band shift correction.

DATASET	CENTRE BANDS								
MERMAID (MERIS)	412.5	442.5	490	510	560	620	665	681	709
AAOT (band-shifted)	413	443	490	–	560	–	665	–	–
Gustav-Dalén Tower (band-shifted)	412	439	490	–	554	–	668	–	–
Helsinki-Lighthouse (band-shifted)	413	441	491	–	555	–	668	–	–
BOUSSOLE	412	443	490	510	560	–	665	683	–
East English Channel	412	443	490	510	559	619	664	–	–
Plumes and Blooms	412	443	490	510	*555*	–	665	–	–
MOBY	412.5	442.5	490	510	560	620	665	681.25	708.75
NOMAD	411	443	489	510	*555*/560	619	665	683	–
SIMBADA	410	443	490	510	560	620	*670*	–	–

converted into a state which is independent of the observation geometry, i.e. the sun position is at the zenith and the viewing direction is in the nadir.

Specific stations of the AERONET-OC ("AErosol-RObotic-NETwork-Ocean-Color" component, Zibordi et al. (2009), Zibordi et al. (2010)) (AAOT [Aqua Alta Oceanographic Tower], Helsinki Lighthouse, Gustav Dalén Tower) are selected as well as the two major buoys located in deep open-ocean waters; the Marine Optical Buoy (MOBY, Clark et al. (1997)), and the buoy for the acquisition of long-term optical times series BOUSSOLE (Bouée pour l'acquisition de Séries Optiques à Long Terme, Antoine et al. (2008)). In addition, the data ensembles from different cruises (Plumes and Blooms, NOMAD (Werdell and Bailey (2005)), SIMBADA (Deschamp et al. (2004))) are considered. These stations are supposed to comprise chlorophyll dominated optical water types (case 1 water) which cover most of the open oceans. In a stricter definition of case 1 the data set is restricted to spectra with reflectances at 560 nm smaller than 0.01.

For the AERONET sites, MERMAID provides the data with site specific band-shift correction as not all in-situ radiometers share the same spectral bands as MERIS (Zibordi et al. (2009)). To NOMAD and SIMBADA data, an empirical band-shift is applied at 555 nm to 560 nm and 670 nm to 665 nm respectively, where necessary. The empirical band-shift utilises in-situ data from NOMAD, where in-situ measurements at 555 and 560 nm or 665 and 670 nm have been taken simultaneously. Their dependence can be described by a linear relationship, if bias-corrected logarithmic reflectances ρ are considered. The linear fit assumes errors for both variables. To correct the spectral mismatch at 555 nm to 560 nm the following empirical relationship (Eq. 2) is applied:

$$\log_{10}\rho\,(560) = bias + a + b \cdot \log_{10}\rho\,(555) \tag{2}$$

with $bias = 0.0172$, $a = 0.1735$ and $b = 1.0768$. The band-shift from 670 to 665 nm uses a $bias = 0.0751$, $a = -0.0198$ and $b = 1.035$. Even though this empirical approach is not ideal, it serves the purpose of the analysis to increase the number of exhaustive spectra. (Table 1)

2.2. AC processor selection

The candidates for the atmospheric correction procedure is the standard processor for MERIS, here noted as MEGS, the SeaDAS 6.3, the POLYMER processor in the algorithm's version 2.4.1 (Steinmetz et al. (2011)), and an implementation of the ForwardNN, which is a modification of the MERIS' standard processor for retrieval of case 2 water constituents. MERIS IPF-6, commonly referred to as MEGS8, has a NeuralNet-algorithm applied for atmospheric correction specific for the retrieval of case 2 water constituents (Doerffer (2011)).

The MEGS processor has been developed specifically for the MERIS sensor and has undergone continuous improvement and optimisation.

SeaDAS started with CZCS, was optimised and applied to SeaWIFS and MODIS and was recently extended to other sensors such as MERIS. Especially being applicable to many sensors makes it a prominent candidate for producing a multi-sensor long-term climate data record.

The processors MEGS and SeaDAS incorporate algorithms, which rest on the assumption that there is no signal coming from the water in the NIR. They are therefore by definition only valid in case 1 water, which holds this assumption. The atmospheric contribution is then extrapolated to the visible part of the spectrum. To further the application beyond case 1 waters, a bright pixel correction has been introduced in MEGS8.

Other algorithms that utilise both visible and near-infrared bands have been developed. The first algorithm of this type used operationally, had been included in the IPF-6. This neural net approach is optimised for case 2 waters and has been designed to work in sun glint conditions (Doerffer et al. (2008)). This algorithm has been recently modified to a combined forward-NN and an iterative optimisation method to allow usage with a flexible subset of a total 35 wavelength bands. A prototype version of this ForwardNN approach has been included in the analysis, which is known to suffer from an implementation error. The angular specifications are faulty which lead to a large loss of data to invalid products on the right hand side of each satellite orbit. Nevertheless this severe error does not strongly deteriorate the quality of the water leaving reflectance when compared to in-situ match-ups. As this paper focusses on the selection methodology, it has been decided to keep the uncorrected results of the ForwardNN. After a future revision of the processor results are expected to change for the better.

Another independent algorithm development of this type is the POLYMER processor. It also uses many wavelength bands in the visible and the near infrared region. Similarly to the NN-processor, it is capable of handling radiance data, which is strongly affected by sun glint and by successfully retrieving water leaving reflectances. A three-day composite map of chlorophyll derived from MERIS with the standard MEGS processor and the POLYMER processor, depicts the increase in spatial and temporal coverage vividly (Fig. 1).

2.3. Selection and preparation of match-up data

The selection of data points from the MERMAID database relies on several levels of combined quality information:

1. The satellite overpass has to be within a three hour interval before or after the in-situ measurement. All sky conditions are allowed, while the maximum wind speed is 9 m/s.
2. The quality of the in-situ measurement has to be approved by the principal investigator (PI) as specified by the MERMAID flags. A normalisation of the water reflectances also has to be applied either by the PI himself or by the MERMAID team.

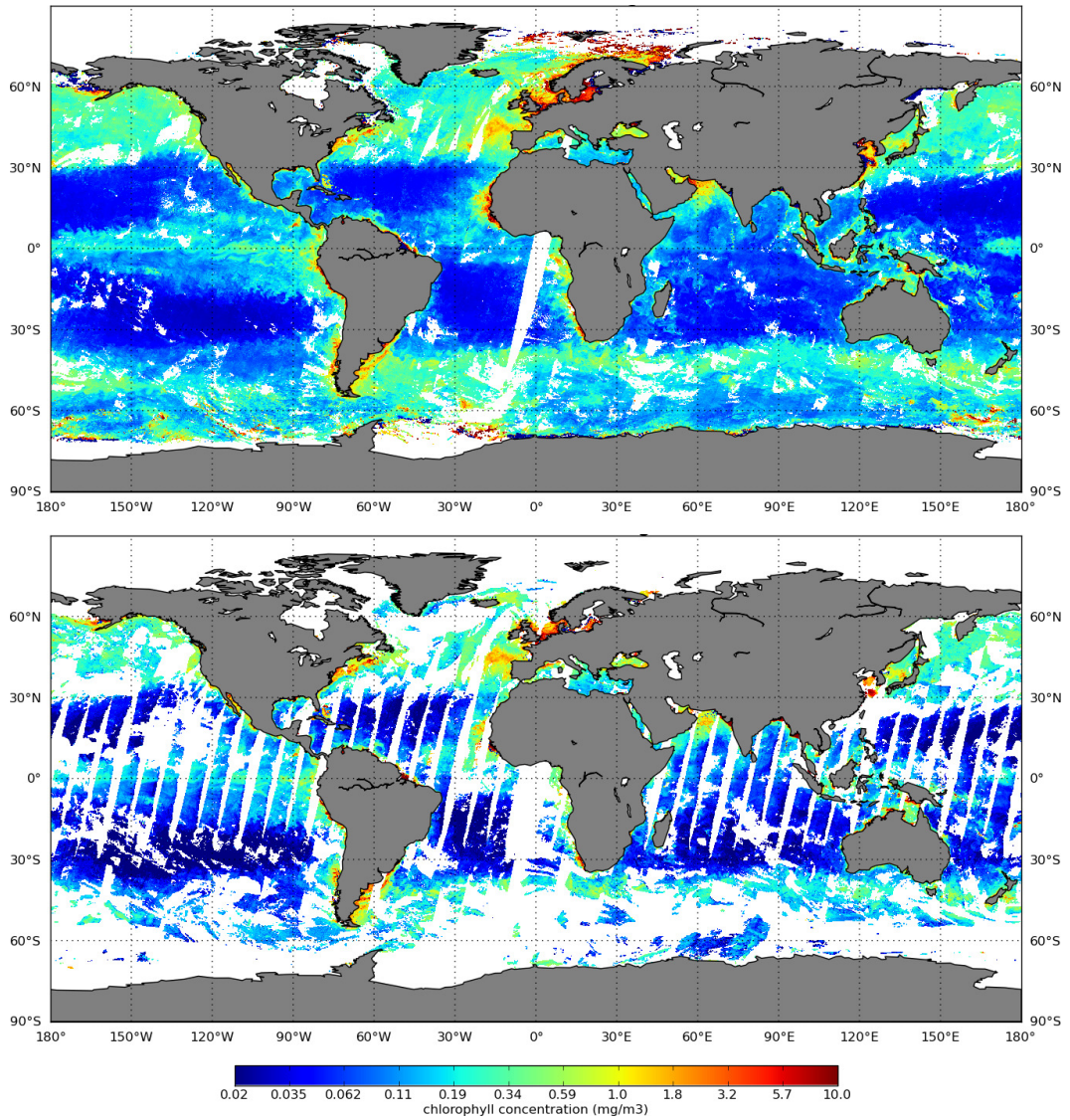


Figure 1: MERIS 3-day composite chlorophyll retrieval (March 20th to 22nd 2003) processed with POLYMER (top) and with the standard MEGS processor (bottom). The coverage for POLYMER is doubled (73% compared to 35.5%) mainly due to retrieval under sun glint conditions.

Table 2: The combination of quality flags, which defines the validity of level 2 products, is given for each atmospheric correction processor. Satellite data is considered in the analysis, only if it fulfills the following quality criteria.

Processor	Valid L2 product defined by combination of individual Quality Flags
MEGS 8.0	NOT (land OR cloud OR ice haze OR high glint OR uncertain normalised surface reflectance OR aerosol model outside database)
ForwardNN	sumsq $< 10^{-5}$ AND N.iter < 150
SeaDAS 6.3	NOT (land OR cloud OR sea ice OR high glint OR cloud shadow OR bright pixel OR aerosol max OR high solar zenith OR high sensor zenith OR navigation failure OR atmospheric correction warning OR atmospheric correction failure OR stray-light)
POLYMER	NOT (land OR cloud OR invalid L1 OR negative bb OR out of bounds)

3. To each of the 3 by 3 pixel of the macro-pixel extraction, the quality of the satellite product is given by the assorted flags of each processor individually.
4. After removing outliers from among the valid pixels with a 3-sigma filter, the quality of a macro-pixel is checked by statistical considerations. The macro-pixel has to be spatially homogeneous; a criterium which is met if the standard deviation of the valid pixels is smaller than 15%, relative to the median of those pixels. Overall more than five valid pixels have to remain (Cui et al. (2010)).
5. The median and standard deviation of the reflectances from each spatially homogeneous macro-pixel are afterwards compared directly with the in-situ measurement.

The quality flags, which define a valid satellite pixel after the implementation of each processor, are summarised in table 2.

Two different types of data sets are created from this selection:

- Individual Best Quality (IBQ): If, for at least one of the processors, more than half of the pixels in the macro-pixel are valid, it is considered to be of individual best quality.
- Common Best Quality (CBQ) is more restrictive: Each processor needs to provide five or more valid pixels to the macro-pixel. Afterwards, the homogeneity criterium is checked. If a single processor fails at a specific wavelength, all the match-up data for this wavelength is discarded for all processors. Therefore, exactly the same amount of data per wavelength and site is considered in the comparison.

In total, the analysis considers three datasets for each selected quality. Although the MOBY and BOUSSOLE data are used for vicarious calibration of the algorithms implemented in SeaDAS (MOBY, Franz et al. (2007), Mélin et al. (2011)) and MEGS (MOBY and BOUSSOLE, Lerebourg et al. (2011)), the global data-set (in the following referred to as *Global*) comprises all sites. The products of the two processors at these sites are expected to be less biassed and are therefore favoured in the comparison. As a compromise between losing a large number of well characterised and quality controlled in-situ measurements and risking preferential treatment of two processors, the global data-set and two subsets are studied; one reduced by the match-up points at MOBY (*Global-MOBY*); and the other data-set reduced by MOBY- and BOUSSOLE-data (*Global-MOBY-BOUSSOLE*).

Table 3: Possible number of match-up points for different sites and the individual and common best data selection based on the available in-situ data. Some wavebands at each site provide less data (wavelength marked with *) than the total amount of measurements. The number of match-up points with strict case 1 water type conditions are given in brackets.

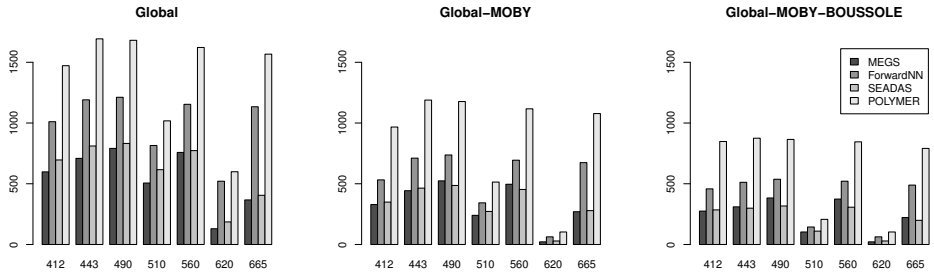
Site	IBQ total match up	CBQ total match up	In-situ wavelengths	Comment
AAOT	425 (60)	56 (7)	412, 443, 490, 560, 665	all spectra with all designated wavelengths
BOUSSOLE	343 (284)	73 (61)	412*, 443*, 490*, 510*, 560*, 665*	some spectra with missing wavelengths
MOBY	559 (559)	232 (232)	412, 443, 490, 510, 560, 620, 665	all spectra with all designated wavelengths, full MERIS set
East English Channel	7 (4)	2 (-)	412, 443, 490, 510, 560, 620, 665	all spectra with all designated wavelengths, full MERIS set
Gustav Dalen Tower	155 (120)	26 (21)	412*, 443, 490, 560, 665	412 missing occasionally
Helsinki Lighthouse	109 (96)	1 (1)	412*, 443, 490, 560, 665	412 missing occasionally
Plumes and Blooms	31 (30)	16 (14)	412, 443, 490, 510, 560, 665	all spectra with all designated wavelengths
NOMAD	113 (83)	32 (27)	412, 443, 490, 510*, 560, 620*, 665*	full MERIS set restricted by 620: IBQ: 18, CBQ: 1
SIMBADA	93 (67)	28 (24)	412*, 443*, 490*, 510, 560*, 620, 665	full MERIS set restricted by 412: IBQ: 63, CBQ: 24
Total	1835 (1303)	466 (387)		full MERIS band set: IBQ: 659, CBQ: 265

All data sets are screened for close-by match-ups in time and space. During a cruise, several in-situ measurements on the same day may refer to the same pixel or one in close proximity on a single scene. In order to avoid spatial dependencies, only the match-up point closest to satellite overpass remains in the data-set.

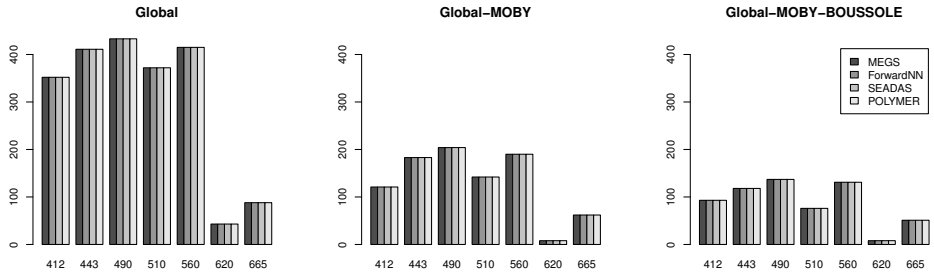
Although the OC-CCI project strives to employ an AC processor which handles several kinds of optical water types and not only the chlorophyll dominated open ocean (case 1), the restriction of the MEGS processor to these waters by definition needs to be addressed in the choice of the database. In order to check for the influence of non case 1 waters in the comparison, a data-set is fashioned, which is reduced to measurements in pure case 1 conditions. They are defined by the normalised water leaving reflectance at 560nm, which has to be smaller than 0.01.

A suitable processor should offer good results over the entire spectral range. The different in-situ sites provide different sets of wavelengths with only a few giving the full set of MERIS bands in the range of the visible spectrum (Tab. 3); considered to be 412, 443, 490, 510, 560, 620 and 665nm in this study. The common set of wavelengths consists of the five bands at 412, 443, 490, 560 and 665nm, which are employed to assess the goodness of spectral fit. In this case the number of spectra in the investigation is constricted by the availability of measurements at 665nm (Fig. 2). Confining the bands up to 665 nm is a reasonable choice for case 1 waters.

In the MERMAID database the full spectrum is only provided at some sites, like MOBY and the East English Channel, and with the collections of samples from cruises (NOMAD and SIMBADA). Not all of these spectra enclose the entire set as some are missing measurements at



(a) Individual Best Quality



(b) Common Best Quality

Figure 2: Number of match-up points per wavelength (nm) and atmospheric correction algorithm by the defined validity of individual flags and spatial homogeneity of macro-pixel. The number of full spectra is restricted by availability of in-situ measurements of the 620 nm band. Differences in IBQ reflect individual quality flags.

412, 510 or 620nm (see also number of available match-up points per wavelength, Fig. 2). The amount of data originating at different sites, which creates the overall global data set, has to be considered during interpretation. E.g. results for the 620nm band reflect the behaviour at MOBY predominantly. Data from MOBY in particular prevail any analysis which uses the full MERIS band-set or is based on the common best quality data set. Removing MOBY from the global data-set reduces the number of match-up points significantly.

3. Methods

A large variety of statistical parameters is used to cover important aspects in the quality assessment for a climatological data-set. Each measure and its associated information is described in detail. They are calculated from the normalised water leaving reflectance of single bands, denoted by λ, which are omitted in some of the formulae. The statistics are only calculated if 10 or more match-up points per wavelength are available. Each statistical parameter per wavelength is transformed into a negative oriented value, which suggests that the smaller the value, the better the result. Together with its respective confidence interval this is the prerequisite for the scoring scheme.

3.1. Selection of statistical parameters

The in-situ water leaving reflectances at 412nm up to 665nm are compared with the derived water leaving reflectances of four different atmospheric correction algorithms for each wavelength independently. The comparison at each in-situ site is based on a set of statistical measures for each wavelength and a test on the goodness of fit between in-situ and satellite spectra. In detail these measures are

- the absolute and the relative RMSE,
- the correlation coefficient r^2,
- the number of valid data points N relative to the total amount,
- the bias and the residual error in absolute terms (i.e., in the unit of ρ), and
- the intercept and slope of a linear regression, the latter of which assumes errors in satellite and in-situ data.
- As a test on the goodness of fit between in-situ and satellite spectra, a common set of five bands is used to calculate a mean chi-square value.

3.1.1. Absolute and relative Root mean square error (RMSE) and its confidence interval at 95%

The absolute and relative RMSE are defined as:

$$RMSE.abs = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(X_i^E(\lambda) - X_i^M(\lambda)\right)^2},\ RMSE.rel = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\frac{X_i^E(\lambda) - X_i^M(\lambda)}{X_i^M(\lambda)}\right)^2} \tag{3}$$

The superscript E denotes the estimated variable (i.e. the water leaving reflectance from a particular atmospheric correction algorithm) and the superscript M denotes the measured variable (i.e. the water leaving reflectance from an in-situ measurement).

The 95% confidence interval of the RMSE can be derived from the standard error

$$RMSE.abs_{st.error} = \frac{sd\left(X^E(\lambda) - X^M(\lambda)\right)}{\sqrt{N}}, \text{ or } RMSE.abs_{st.error} = \frac{1}{\sqrt{N}} sd\left(\frac{X^E(\lambda) - X^M(\lambda)}{X^M(\lambda)}\right) \tag{4}$$

multiplied by the t-value for $\alpha = 0.95$ and degree of freedom N-2.

To judge different water types and spectra equally, both errors are analysed. They come with different shortcomings: the relative errors emphasise the behaviour at small values, e.g., as may occur in the red part of the spectra and may even lead to unrealistic large errors, while the absolute errors are sensitive to outliers.

The errors are negatively oriented values. Errors closer to zero are assigned the higher scores.

3.1.2. Correlation coefficient r

The correlation coefficient r is calculated for the global data-sets, which exhibit enough dynamic range in each of the spectral bands to allow for a meaningful correlation coefficient and regression. The 95% confidence interval of the correlation coefficient is calculated.

$$r = \frac{\sum\left(X^M - \overline{X^M}\right) \cdot \left(X^E - \overline{X^E}\right)}{\sqrt{\sum\left(X^M - \overline{X^M}\right)^2 \cdot \sum\left(X^E - \overline{X^E}\right)^2}} \tag{5}$$

The correlation coefficient should be close to a value of one. It is therefore transformed into a negative oriented value by

$$\tilde{r} = 1 - r \tag{6}$$

3.1.3. Bias

The bias is defined as the mean sum of differences between estimated variable and measured variable.

$$Bias = \frac{1}{N}\sum_{i=1}^{N} X_i^E - X_i^M \tag{7}$$

The bias should be close to zero. The scores will be based on the absolute values of the bias, $\widetilde{Bias} = \|Bias\|$, given equal weight to over- or under-estimation. The 95% confidence interval is the same as for the absolute RMSE.

3.1.4. Residual error

The estimated values are corrected for the bias, $X^E_{bias} = X^E - Bias$, and again the absolute RMSE and its 95% confidence interval is calculated.

The residual error accounts for the random error in the data.

3.1.5. Regression: orthogonal distance least-squares fitting

As both satellite and in-situ data are not error-free, the simple linear regression, which uses the minimisation of vertical distances, is discarded and an orthogonal distance least-squares fitting used instead. To solve this problem the first principal component is calculated as the eigenvector of the largest eigenvalue of the covariance matrix. Whereas the satellite data provides a standard deviation per match-up point σ_i^E as error estimate, no such number is available for the in-situ measurements. The errors are taken into account as weights w during the computation of

weighted variances and covariances, whereas the in-situ measurements are considered with equal weights. Confidence intervals for slope and intercept are estimated by cross-validation, which implies leaving a single data point out and the repetition of solving the eigenvalue problem. This method is applied to each wavelength separately. With the definitions of the weights w from the standard deviation σ of N data points

$$w'\left(X_i^E\right) = 1 - \frac{\sigma_i^E}{\left|X_i^E\right|} \; ; \; w\left(X_i^E\right) = \frac{w'\left(X_i^E\right)}{\sum_{i=1}^{N} w'\left(X_i^E\right)} \; ; \; w\left(X_i^M\right) = \frac{1}{N}$$

the weighted variance var_w and covariance cov_w can be defined as

$$var_w\left(X^E\right) = \frac{1}{N-1}\sum_{i=1}^{N} w\left(X_i^E\right)^2 \cdot \left(X_i^E - \overline{X^E}\right)^2 \tag{8}$$

$$cov_w\left(X^E, X^M\right) = \frac{1}{N-1}\sum_{i=1}^{N} w\left(X_i^E\right) \cdot \left(X_i^E - \overline{X^E}\right) \cdot w\left(X_i^M\right) \cdot \left(X_i^M - \overline{X^M}\right) \tag{9}$$

The eigenvalues eig of the covariance matrix are

$$\begin{aligned} \text{eig} \quad = \quad & \frac{var_w\left(X^E\right) + var_w\left(X^M\right)}{2} \\ & \pm\sqrt{\left(\frac{var_w\left(X^E\right) + var_w\left(X^M\right)}{2}\right)^2 - \left(var_w\left(X^E\right) \cdot var_w\left(X^M\right) - cov_w^2\left(X^E, X^M\right)\right)} \end{aligned} \tag{10}$$

The larger eigenvalue eig_{max} leads to the linear regression equation with slope a and intercept b, which explains $\text{eig}_{max} \cdot 100\%$ of the variance in the data:

$$a = \frac{\text{eig}_{max} - var_w\left(X^E\right)}{cov_w\left(X^M\right)} \; ; \; b = -a\overline{X^E} + \overline{X^M} \tag{11}$$

The slope should be close to a value of one. For scoring purposes, the absolute difference of $1 - a$ is used: $\tilde{a} = \|1 - a\|$. The intercept should be close to zero.

3.1.6. Chi-square test of spectral shape

In the evaluation of an algorithm's performance it is of interest to assess not only results for single wavelengths but also how well the overall spectral shape is determined. The chi-square test is used to measure the goodness of fit between in-situ and satellite derived spectral distribution. Before the chi-square value is calculated for each match-up point, the spectral bias is removed by normalising both spectra (in-situ and satellite derived) to 560nm. This analysis is conducted on the most frequently measured set of wavelengths (412, 443, 449, 560, and 665nm). By normalising the spectra, only four wavelengths remain for the calculation. Two different values are derived and judged by the scoring; the mean chi-square, and the percentage of good fitting spectra.

For the Mean chi-square from all available spectra, the chi-square values are determined using four wavelengths (412, 443, 490, and 665nm). Due to normalisation the contribution of 560nm is zero. These values are averaged, disregarding outliers which are above the 95% confidence level.

$$\chi_j^2 = \sum_{i=1}^{4} \frac{\left(X_j^E\left(\lambda_i\right) - X_j^M\left(\lambda_i\right)\right)^2}{X_j^M\left(\lambda_i\right)}, \; \overline{\chi^2} = \frac{1}{N}\sum_{j=1}^{N} \chi_j^2 \tag{12}$$

In the individual best data-set these values refer to different numbers of match-up points. In addition, they vary due to the disregarded outliers.

The percentage of chi-square values ($N\chi^2$) lower than the 95% confidence level, which results in a χ^2 value of 3.8 for 1 degree-of-freedom, is calculated for the entire dataset. Spectra with higher chi-square values indicate cases in which the shape could not be well reconstructed. The percentage of good spectral matches is directly used in the scoring after normalisation.

3.1.7. Relative number of valid match-up points

The number of valid match-up points is related to the total number of possible match-ups (Tab. 3). This ratio is normalised with respect to the algorithms and the result directly taken as a score.

3.2. The scoring scheme

Most of the statistical properties come with a standard error or 95% confidence interval. All properties are transformed to negative orientated values, if necessary. To each property the evaluation scores are assigned by wavelength separately in the following manner:

- The best algorithm is the one with the smallest value in the statistical property and receives 2 points.
- If the value corresponding to another algorithm 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 doesn't overlap with the best algorithm, this algorithm receives 0 points.
- In order to weigh each wavelength equally the scores will be normalised, so that the sum of all points per wavelength and property over all algorithms equals 1.

All scores S are then summed up per wavelength and statistical property, which gives each of them equal weight. The measures of spectral shape, i.e. the mean χ^2 value and the number of spectra with a χ^2 lower than the 95% confidence level, receive the same weight as a single waveband. Their scores are therefore multiplied by eight (because there are eight statistical parameters considered per wavelength), when added up to a total score.

$$\begin{aligned} S_{\text{total}}\left(\text{Processor}\right) = & \left(\sum_{i=1}^{7} S_{\text{RMSE.abs}}\left(\lambda_i\right) + S_{\text{RMSE.rel}}\left(\lambda_i\right) + S_{\text{r}}\left(\lambda_i\right) + S_{\text{bias}}\left(\lambda_i\right)\right. \\ & \left. + S_{\text{res.error.abs}}\left(\lambda_i\right) + S_{\text{Slope}}\left(\lambda_i\right) + S_{\text{Intercept}}\left(\lambda_i\right) + S_{\text{N}}\left(\lambda_i\right)\right) \\ & + 8 \cdot \left(S_{\chi^2} + S_{N\chi^2}\right) \end{aligned} \tag{13}$$

In favouring the best algorithm strongly, this scoring system tends to behave in a non-linear way. This approach has been preferred over one of the relative scores, which considers all relationships to fixed limits per statistical parameters. It would have been necessary to define in absolutes of what is supposed to be a "good result". The choice would have been highly subjective. Another scoring approach which is applied in the comparison of in-water algorithms (Brewin et al. (2012)), uses the larger number of algorithms to its benefit. From over ten different retrieval algorithms for water constituents, the mean of the statistical value under study and the confidence interval, are calculated. All algorithms that perform within the confidence interval get the same score, while algorithms outside the interval receive a score of zero. With only four algorithms available this approach cannot be applied to the atmospheric correction comparison.

Table 4: Statistical parameters bias, correlation coefficient r and absolute RMSE with standard errors or confidence interval for normalised water leaving reflectance at 560nm, selection IBQ or CBQ, Global, and all water types.

Quality	Stat. param.	MEGS	ForwardNN	SeaDAS	POLYMER
IBQ	Bias	-7.6±1.3 10^{-4}	3.2±0.88 10^{-4}	-1.0±0.09 10^{-3}	-4.0±7.8 10^{-5}
	r	0.95 (0.95-0.96)	0.94 (0.93-0.94)	0.93 (0.92-0.94)	0.96 (0.95-0.96)
	abs. RMSE	2.2±0.13 10^{-3}	1.8±0.09 10^{-3}	1.9±0.09 10^{-3}	1.9±0.08 10^{-3}
CBQ	Bias	-5.9±1.1 10^{-4}	1.8±1.1 10^{-4}	-9.1±1.2 10^{-4}	5.0±8.7 10^{-5}
	r	0.91 (0.89-0.92)	0.88 (0.86-0.9)	0.90 (0.88-0.92)	0.92 (0.9-0.93)
	abs. RMSE	1.5±0.11 10^{-3}	1.4±0.11 10^{-3}	1.7±0.12 10^{-3}	1.0±0.09 10^{-3}

Table 5: Scores converted from statistical parameters bias, correlation coefficient r and absolute RMSE. Based on the example given in Table 4.

Quality	Stat. param.	MEGS	ForwardNN	SeaDAS	POLYMER
IBQ	Bias	0	0	0	1
	r	0.5	0	0	0.5
	abs. RMSE	0	0.33	0.33	0.33
CBQ	Bias	0	0.33	0	0.67
	r	0.33	0	0	0.67
	abs. RMSE	0	0	0	1

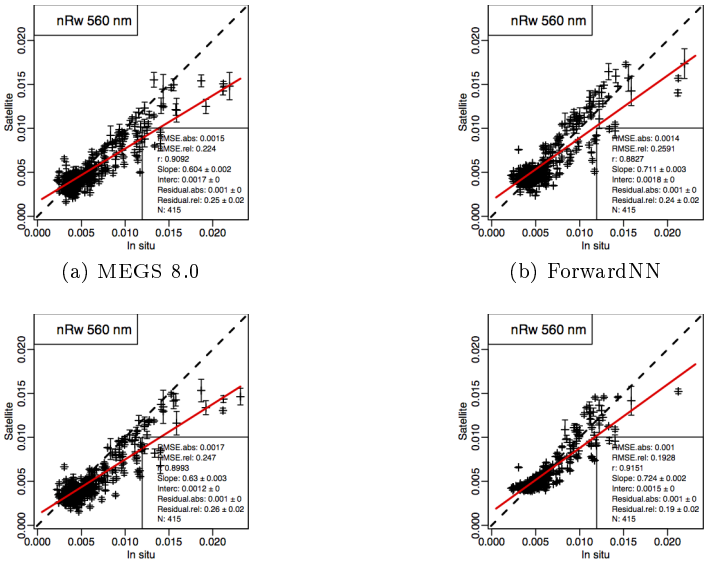
4. Results and Discussion

The datasets of in-situ and satellite data match-ups with their different qualities (IBQ or CBQ, all water types or case 1 only) and three combinations of sites, are analysed for the four atmospheric correction processors considered. All detailed plots and tabled results can be found in the ESA OC-CCI project report on product validation and algorithm selection ("Product Validation and Algorithm Selection Report" PVASR: Müller and Krasemann (2012)[1]). The results of the strict case 1 dataset are not discussed in detail, but can be found in Appendix C.

4.1. Comparison of normalised water leaving reflectances

A typical example of the comparison between normalised water leaving reflectances at 560nm of in-situ and satellite data origin is shown in Fig. 3. Data points from all sites which fulfil the common or the individual best quality criteria, are selected (Global, CBQ or IBQ; see definition section 2.3). Some statistical parameters for the comparison of each algorithm with the in-situ data, are given. The subset of statistical parameters (Tab. 4) demonstrates similar behaviour of the processors' results for both qualities. POLYMER exhibits the smallest bias: which is one order of magnitude lower compared to the other results, and the largest correlation coefficient. In terms of correlation coefficient and absolute RMSE, the processors perform very similarly. Together with the errors or confidence intervals, these values are converted into scores (Tab. 5) in order to acknowledge results from all wavelengths and all statistical parameters in a single number.

[1] http://www.esa-oceancolour-cci.org/, Resources, documents



(a) MEGS 8.0 (b) ForwardNN

(c) SeaDAS 6.3 (d) POLYMER 2.4.1

Figure 3: Comparison of water leaving reflectances of in-situ versus remotely sensed data at 560nm, CBQ, Global. Error bars visualise the standard deviation within the macro-pixel.

For this single wavelength of 560nm the scores of eight statistical parameters result in MEGS: 1.18 (CBQ: 0.87), ForwardNN: 1.6 (CBQ: 0.72), SeaDAS: 2.51 (CBQ: 1.54) and POLYMER: 2.71 (CBQ: 4.88).

Results are comparable and quite good for all wavelengths. As an example the comparison of all wavelengths is shown for the MEGS processing (Fig. 4).

4.2. Comparison of spectral shape

The goodness of fit of the satellite spectra is compared to the in-situ data by calculating their χ^2 values (see section 3.1.6). The distribution of χ^2 values should have its maximum close to zero and preferably a small half maximum width. The ForwardNN algorithm performs considerably better than POLYMER with respect to the reconstruction of the spectral shape, resulting in less than half the half maximum width and a sharp maximum closer to zero (Fig. 5). ForwardNN and MEGS perform comparably well (IBQ). The selection of CBQ consists of only 50 spectra, which are used in the analysis. In comparison with the IBQ selection, it can be seen that the results concerning spectral shape remain consistent in the rather sparse subset of CBQ, apart from SeaDAS. It seems necessary to base the interpretation on the IBQ dataset if the distribution half width should remain meaningful.

4.3. Converting the results to scores

After the statistical evaluation the results are converted into scores. The performance at all wavelengths and for reconstruction of spectral shape is summarised in this single number for each selection considering quality or combination of sites (Table 6). The interpretation is not straight-forward. The scores are always strongly interdependent as they indicate basically the performance of a processor relative to the best.

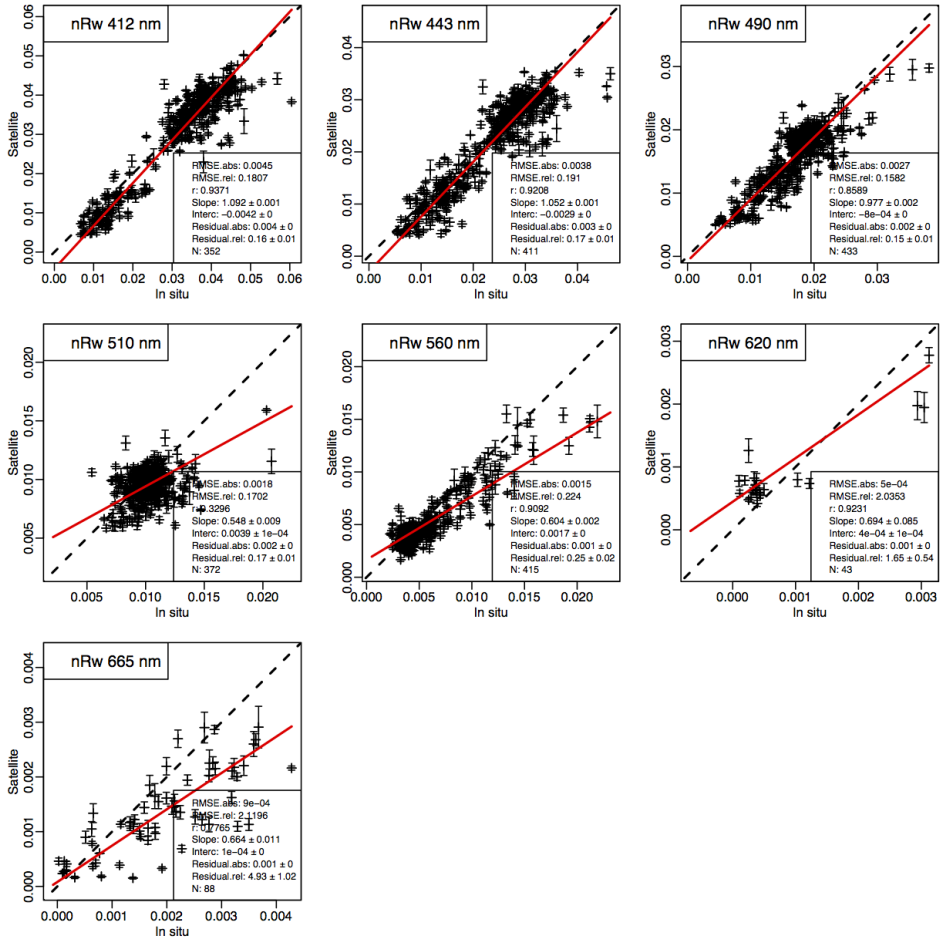
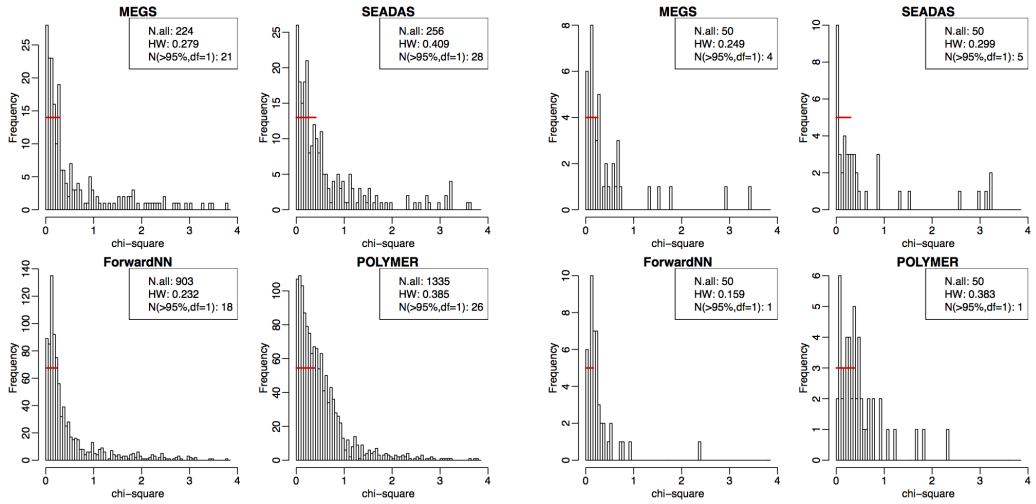


Figure 4: Comparison of water leaving reflectances of in-situ versus remotely sensed data for MEGS 8.0, CBQ Global.



(a) Individual best quality

(b) Common best quality

Figure 5: Distribution of χ^2 per spectrum. Total number of spectra, N.all. Half maximum width, HW. Global data set. The histogram classes are the same for all algorithms, their frequency of spectra per class is variable in order to highlight the distribution shape.

Table 6: Total scores for different global data sets and the two selections of quality (IBQ or CBQ).

Quality	Dataset	MEGS	ForwardNN	SeaDAS	POLYMER
IBQ	Global	8.41	15.61	20.72	27.26
	Global-MOBY	10.58	18.85	17.55	24.98
	Global-MOBY-BOUSSOLE	10.18	17.91	19.94	23.95
CBQ	Global	13.76	14.07	21.22	22.98
	Global-MOBY	11.9	19.8	20.04	20.28
	Global-MOBY-BOUSSOLE	13.39	19.05	19.13	20.49

If MOBY and BOUSSOLE data are removed (see Fig. 6, top row), the bias of MEGS and SeaDAS products increase expectedly, while there is little effect on the scores. The changes in the MEGS' scores from the global to the reduced data-sets seem counter-intuitive at a first glance. For example, the increase in bias would lead to exactly the same scores for all three data-sets because the relationship towards the products of the best processor does not change. Doubled scores do not correspond to doubled quality in performance, as rather small absolute differences in statistical parameters can be magnified strongly.

Whereas POLYMER is a clear winner for the data-sets of individual best quality, the reduction of data points in the CBQ data-set seems to favour the conclusion that all processors - with the exception of MEGS - perform equally well. To strengthen the reliability of the scores, they are investigated in a bootstrap exercise.

4.4. Sensitivity of the scores

The statistical properties derived are used to construct the scoring as described in section 3.2. The methodology behind the scores has to be tested on non-linearity, their dependencies on the

selection of the chosen statistical measures and the database. This is necessary if they are to become a sound foundation to select the best atmospheric correction processor. By using a bootstrap method (Efron (1979)) these influences are investigated. By selecting the bootstrap method, the experiment is slightly biassed as all wavelengths have to be present. This bias occurs when incorporating the few remaining sites where reflectances at 510 or 620nm have been measured. These measurements are found - especially after MOBY data have been removed - mainly in the "Plumes and Blooms" and the SIMBADA data ensembles. The statistical evaluation proceeds only if 10 or more data points at all wavelengths are available, otherwise the sampling is repeated.

The results from the random selections (with repetitions) are compared to the total scores associated with the original datasets (referred to as *single representation*s, as every data point occurs exactly and only once).

Each distribution is outlined with a normal distribution function, using the median of the score distribution as μ and the standard deviation σ, multiplied by the height of the classes' maximum; highlighting the deviation from an unskewed Gaussian distribution.

4.4.1. Sensitivity to selection of data points

Currently the choice of match-up points is restricted to the MERMAID database, however, ideally changes in the database with respect to the amount of data at different sites should not affect the results of the total scores.

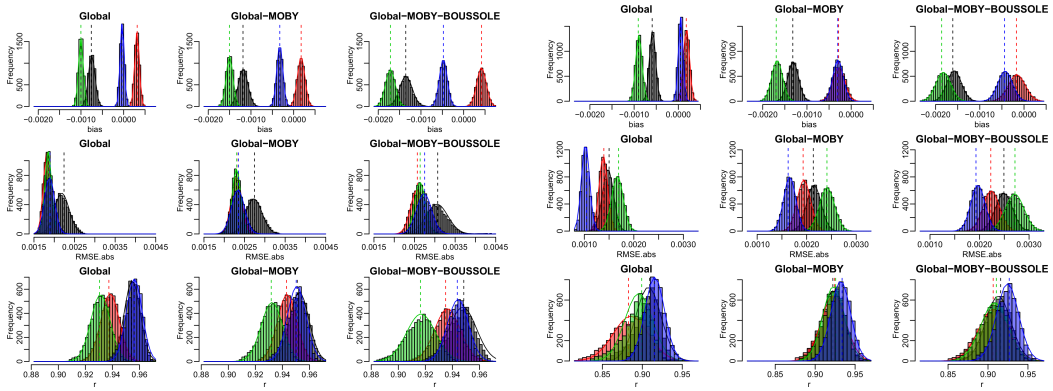
The IBQ and CBQ match-up points are analysed by using either the entire data-set or a reduced version which leaves out MOBY and/or BOUSSOLE data. These data-sets undergo resampling 5000 times. The resampled data-sets are then statistically evaluated and scored.

Statistical parameters. In the course of the bootstrap exercise, the statistical parameters derived from each sampling, are collected (Fig. 6). All parameters exhibit gaussian behaviour (more or less), with stronger skewness for the correlation coefficient r. In most cases, the value of the single representation (dashed lines, see Fig. 6a) is in agreement with the distribution maximum, unless it is strongly skewed. The width of the distribution supports the decisions derived from standard errors or confidence intervals during the conversion into scores (see Tab. 5). The smallest absolute bias is found for the POLYMER processor (Fig. 6a, top left) for which its distribution is clearly separated from all other processors; it is assigned a score of 1 for this reason (Tab. 5). The distribution of the correlation coefficient r of the two best performing processors in this respect coincide almost completely (Fig. 6a, bottom left). MEGS and POLYMER are therefore both given a score of 0.5. In the case of absolute RMSE, the distributions of the three best results coincide; therefore ForwardNN, SeaDAS and POLYMER are assigned a score of 0.33 each.

Total scores. The total scores largely show gaussian behaviour (Fig. 7) with the single representation values falling within the centre half of each distribution.

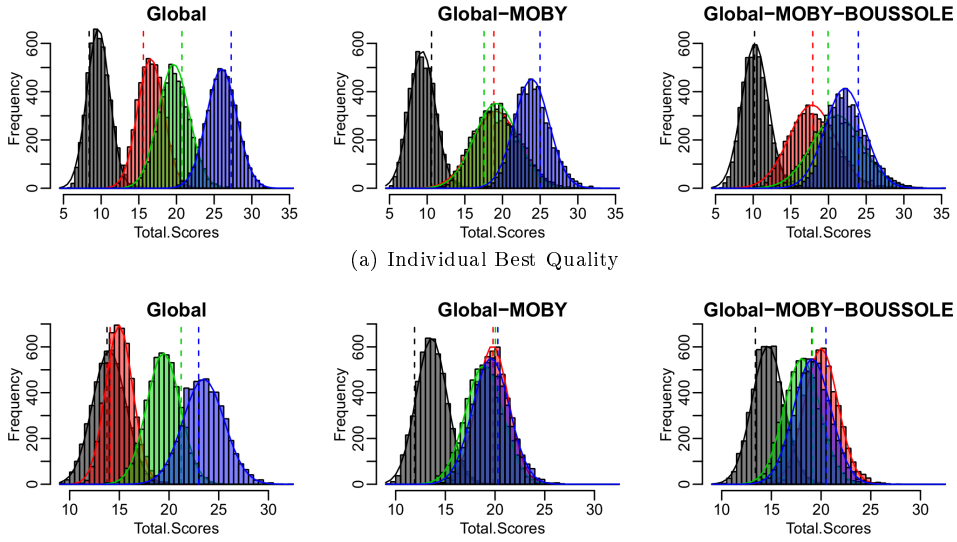
Comparing the median of the distribution with the results of the single representation reveals that simply judging based on the latter might lead to slight misinterpretations (Fig. 7, Tab. 7). In the IBQ selections, the total score for POLYMER is always higher in the single representation than in the maximum of the distribution. Instead of concluding that POLYMER performance is better than SeaDAS and ForwardNN (for Global-MOBY-BOUSSOLE), it is clear that within the overlap of the distributions all three processors perform equally well.

The width of the distribution of scores is rather large, which suggests strong influence from the selection of data points. Subsets of the data may lead to statistical parameters (and scores) that change the ranking of processor performance significantly if they are only investigated by the single representation.



(a) Individual Best Quality (b) Common Best Quality

Figure 6: Statistical parameters bias, RMSE and correlation coefficient in dependency to selection of data in bootstrap exercise for normalised water leaving reflectance at 560nm. Colours represent the atmospheric correction algorithms: POLYMER (blue), SeaDAS (green), ForwardNN (red) and MEGS (black). The dashed lines show the parameter value of the single representation.



(a) Individual Best Quality

(b) Common Best Quality

Figure 7: Distribution of total scores from bootstrapping the match-up points of IBQ (N=1835, without MOBY: N=1276, without MOBY and BOUSSOLE: N=933) or CBQ (N=466, without MOBY: N=234, without MOBY and BOUSSOLE: N=161). Bootstrapping repetition 5000 times. The dashed lines show the score of the single representation.

Table 7: Median and standard deviation of the total score distribution accessed by bootstrapping the match-up database, compared to the single representation, which is illustrated in Fig. 7.

(a) Individual Best Quality

Algorithm	Global	Global-MOBY	Global-MOBY-BOUSSOLE
MEGS 8.0	8.41 / 9.62±1.51	10.58 / 9.53±1.72	10.18 / 10.24±1.76
ForwardNN	15.61 / 16.58±1.74	18.85 / 19.13±3.06	17.91 / 17.8±3.06
POLYMER 2.4.1	27.26 / 26.00±2.01	24.98 / 23.87±2.29	23.95 / 22.17±2.63
SeaDAS 6.3	20.72 / 19.63±2.05	17.55 / 19.63±2.89	19.94 / 21.13±3.33

(b) Common Best Quality

Algorithm	Global	Global-MOBY	Global-MOBY-BOUSSOLE
MEGS 8.0	13.76 / 13.98±1.68	11.9 / 13.62±1.55	13.39 / 14.66±1.60
ForwardNN	14.07 / 14.98±1.36	19.8 / 19.75±1.62	19.05 / 20.01±1.68
POLYMER 2.4.1	22.98 / 23.48±2.07	20.28 / 19.59±1.86	20.46 / 19.11±1.87
SeaDAS 6.3	21.22 / 19.44±1.69	20.04 / 18.9±1.88	19.13 / 18.09±1.81

4.4.2. Sensitivity to selection of statistical parameters

The choice of statistics which are the foundation to the scores, should not affect the outcome of total scores. In a second bootstrap exercise this hypothesis has been tested.

The scores from the original datasets comprise 72 components (8 statistic parameters times 7 wavelengths plus 2 spectral parameters given the weight of a single wavelength) for each processor. These components are resampled 20,000 times, allowing for repetition. The resampled score components may randomly leave out one or more statistical measures at certain wavelengths, and double others. For each resampling the total scores are summed up for each processor respectively.

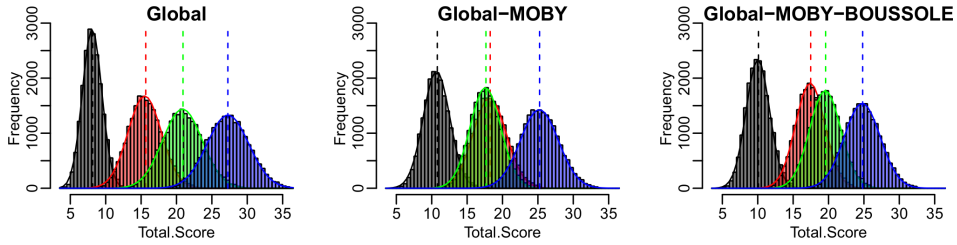
The total scores are normally distributed (Fig. 8). As expected, the median of the distribution is almost identical to the single representation of the total score (Tab. 8).

The choice of statistics does not affect the total scores in an unexpected way. The width of the distribution arises from the different performances of the algorithms at different wavelengths and for the tested properties.

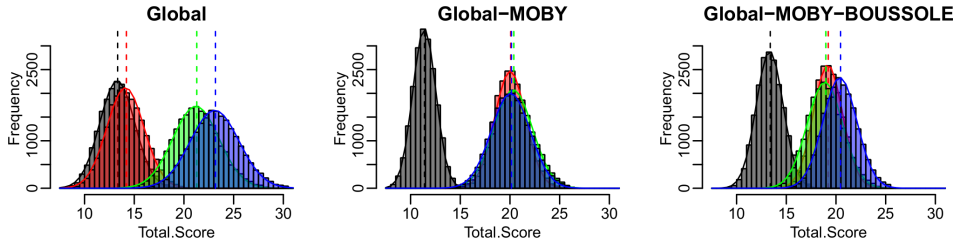
5. Conclusion

In order to implement an objective procedure to select the "best" atmospheric correction processor based on comparisons with in-situ data, a method to convert statistical properties and their confidence intervals into relative scores has been introduced. Although the conversion is straightforward, the interpretation of score values and the width of their associated distributions attained by bootstrap experiments need special care due to their non-linear characteristics. The scores are appointed for each wavelength and statistical parameter regardless of (hypothetical) absolute measures for a well performing processor. Twice the score value does not automatically indicate a doubling in product quality in absolute terms.

Although the scores for the single representation of all data points coincide quite well with the maximum of the scores' distribution obtained from the resampling of the bootstrap method, the width of the distributions is necessary for a conclusive interpretation of the scores. The narrower the distribution the more often a processor has been found in similar relative relationships with the others. Overlapping distributions result from data samplings which lead to interchanges in



(a) Individual Best Quality



(b) Common Best Quality

Figure 8: Distribution of total scores from bootstrapping 72 statistical measures (x 20,000). Colours represent the atmospheric correction algorithms: POLYMER (blue), SeaDAS (green), ForwardNN (red) and MEGS (black). The dotted lines show the score of the single representation.

Algorithm	Global	Global-MOBY	Global-MOBY-BOUSSOLE
MEGS 8.0	8.07 / 8.02±1.39	7.65 / 7.54±1.35	10.42 / 10.35±1.47
ForwardNN	15.31 / 15.24±2.36	19.64 / 19.59±2.42	19.13 / 19.06±2.06
POLYMER 2.4.1	27.50 / 27.43±2.97	25.11 / 25.06±2.59	22.84 / 22.80±2.43
SeaDAS 6.3	21.11 / 21.05±2.83	19.59 / 19.54±2.29	19.58 / 19.54±2.25

(a) Individual Best Quality

Algorithm	Global	Global-MOBY	Global-MOBY-BOUSSOLE
MEGS 8.0	16.19 / 16.08±2.12	13.39 / 13.32±1.54	15.58 / 15.50±1.93
ForwardNN	15.24 / 15.16±2.20	20.57 / 20.47±2.11	18.78 / 18.72±1.84
POLYMER 2.4.1	21.85 / 21.79±2.40	20.72 / 20.67±2.01	19.32 / 19.28±1.80
SeaDAS 6.3	18.71 / 18.63±2.45	17.41 / 17.35±1.71	18.36 / 18.29±1.79

(b) Common Best Quality

Table 8: Median and standard deviation of the total score distribution assessed by bootstrapping the statistical parameters, compared to the single representation.

best performance. Isolated distributions indicate that the processor performed well independent of the resampling.

From the distribution width (Fig. 7) it is evident that the selection of match-up points strongly influences the outcome of the scoring process. Changes in the database can rearrange the ranking of performances entirely. The evaluation of processors is therefore always dependent on the available in-situ data and it seems necessary to include resampling in order to avoid ambiguities in the interpretation with changing databases.

The method can easily be applied to less or more processor candidates, however score results are not directly comparable. Two different approaches are possible if, for example, a new version of an existing processor is introduced as a fifth candidate and is compared to the set of former candidates. If the new version replaces the older one in the analysis, while the other processors remain unchanged, changes in the relative relationship can be identified. On the other hand this experiment is not sufficient to answer the question of improvement between versions. As the entire system of relative connections may have changed, scores of two different runs cannot be compared directly. To identify the changes due to processor development, it is advisable to administer the same experiment on the dataset adding the new version as a fifth processor. In relation to the other processors, it is possible to distinguish differences in product quality between two different processor versions. Essentially, each experiment needs to be handled as an independent result.

If interpreted in this fashion the scoring statistics reveal a slight advantage of POLYMER, which becomes less pronounced if MOBY and/or BOUSSOLE data points are removed. Leaving both sites out, POLYMER remains only slightly better than the ForwardNN or SeaDAS in an IBQ selection, while all three processors show equal performance in those CBQ selections. Differences between IBQ and CBQ statistics arise from the larger dataset, especially in the POLYMER IBQ processing, which reduces uncertainties in the statistical parameters. Therefore, rather narrowbanded intervals have to be matched by confidence intervals of other processors' products during the scoring. Taking this reasoning into account the relationships portrayed in the IBQ and CBQ experiments are quite similar. Consistent with all experiments, the MEGS processor produces considerably less accurate results compared with the other participating processors.

The CBQ datasets are important in the decision process because the evaluation relies on exactly the same pixels. On the other hand, these results are the most affected by sparse statistics. By removing MOBY data, all processors, with the exception of MEGS, become similar in their performance.

The introduced scoring system including the bootstrap exercise is essential to estimate the variation introduced by the selected set of matchup points.

Nevertheless, even this extended validation methodology is not capable of capturing severe issues like strong angular dependencies, which should have been found out for the known issues of this ForwardNN processor version. In addition to match-up validation, we recommend further tests that involve along scan-line statistics.

Acknowledgements

This work is a contribution to the Ocean Colour Climate Change Initiative of the European Space Agency.

We thank ACRI-ST, ARGANS and ESA for access to the MERMAID system (http://hermes.acri.fr/mermaid) and especially we thank the scientist that have contributed their efforts to in-situ-data: David Antoine, Laboratoire d'Océanographie de Villefranche, France; Bob Arnone, Naval Research Laboratory, Stennis SpaceCenter, USA; William Balch, Bigelow Laboratory for Ocean Sciences, USA; Kendall Carder, University of South Florida, USA; Richard W. Gould, Naval Research Laboratory, Stennis Space Center, USA; Larry Harding, UMCES, USA; Stanford B. Hooker, NASA,

USA; Zhongping Lee, UMB, USA; Hubert Loisel, Université du Littoral-Côte d'Opale, France; Antonio Mannino, NASA Goddard Space Flight Center, USA; B. Gregory Mitchell, SCRIPPS, UCSD, USA; Ru Morrison, Woods Hole Oceanographic Institution, USA; Frank Muller-Karger, University of Maryland, USA; Norman Nelson, Earth Research Institute, UCSB, USA; David A. Siegel, Earth Research Institute, UCSB, USA; Dariusz Stramski, SCRIPPS, UCSD, USA; Ajit Subramaniam, University of Maryland, USA; Kenneth Voss, University of Miami, USA; and Guiseppe Zibordi, Joint Research Centre, Italy.

We would like to thank A. Northrop, VEGA, for her support with the manuscript.

Antoine, D., Guevel, P., Desté, J.-F., Bécu, G., Louis, F., Scott, A., Bardey, P., 2008. The BOUSSOLE buoy – a new transparent-to-swell taut mooring dedicated to marine optics: design, tests and performance at sea. Journal of Atmospheric and Oceanic Technology 25, 968–989.

Antoine, D., Morel, A., 1998. Relative importance of multiple scattering by air molecules and aerosols in forming the atmospheric path radiance in the visible and near infrared parts of the spectrum. Applied Optics 37, 2245–2259.

Antoine, D., Morel, A., 1999. A multiple scattering algorithm for atmospheric correction of remotely sensed ocean color (MERIS instrument): principle and implementation for atmospheres carrying various aerosols including absorbing ones. Int. J. Remote Sens. 20, 1875–1916.

Antoine, D., Morel, A., 2011. Atmospheric Correction of the MERIS observations Over Ocean Case 1 waters. Tech. rep., MERIS ATBD 2.7, Issue 5, revision 5.
URL `http://envisat.esa.int/instruments/meris/atbd/atbd_2.7.pdf`

Bourg, L., Feb. 2012. Evolution of the MERIS Instrument Processing Facility. Tech. Rep. 4, ESA.
URL `http://earth.eo.esa.int/pcs/envisat/meris/documentation/MERIS_IPF_evolution.pdf`

Brewin, R., Sathyendranath, S., Müller, D., Krasemann, H., Doerffer, R., Mélin, F., Brockmann, C., Fomferra, N., Peters, M., Grant, M., Steinmetz, F., Deschamps, P.-Y., Swinton, J., Smyth, T., Werdell, P., Franz, B. A., Maritorena, S., Devred, E., Lee, Z., Hu, C., Regner, P., 2012. The Ocean Colour Climate Change Initiative: III. A round-robin comparison on in-water bio-optical algorithms in open-ocean waters. Remote sens. Environ., submitted.

Chomko, R., Gordon, H., 1998. Atmospheric correction of ocean color imagery: use of the Junge power-law aerosol size distribution with variable refractive index to handle aerosol absorption. Appl. Opt. 37, 5560–5572.

Chomko, R., Gordon, H., 2001. Atmospheric correction of ocean color imagery: test of the spectral optimization algorithm with SeaWiFS. Appl. Opt. 40, 2973–2984.

Chomko, R., Gordon, H., Maritorena, S., Siegel, D. A., 2003. Simultaneous retrieval of oceanic and atmospheric parameters for ocean color imagery by spectral optimization: a validation. Remote sens. Environ. 84, 208–220.

Clark, D., Gordon, H., Voss, K., Ge, Y., Broenkow, W., Trees, C., 1997. Validation of atmospheric corrections over oceans. Journal of Geophysical Research 102, 17209–17217.

Cui, T., Zhang, J., Groom, S., Sun, L., Smyth, T., Sathyendranath, S., 2010. Validation of MERIS ocean-color products in the Bohai Sea: A case study for turbid coastal waters. Remote Sensing of Environment 114, 2326–2336.

Deschamp, P., Fougnie, B., Frouin, R., Lecomte, P., Verwaerde, C., 2004. SIMBAD: A field radiometer for satellite ocean-color validation. Appl. Opt. 43 (4055-4069).

Doerffer, R., 2011. Alternative Atmospheric Correction Procedure for Case 2 Water Remote Sensing using MERIS. ATBD 1.0, HZG.

Doerffer, R., Schiller, H., Fischer, J., Preusker, R., Bouvet, M., 2008. The impact of sun glint on the retrieval of water parameters and possibilities for the correction of MERIS scenes. In: 2nd MERIS / (A)ATSR User Workshop.

Efron, B., 1979. Bootstrap Methods: Another Look at the Jackknife. The Annals of Statistics 7 (1), 1–26.

Franz, B. A., 2012. Algorithm for Retrieval of Remote Sensing Reflectance from Satellite Ocean Color Sensors.
URL `http://oceancolor.gsfc.nasa.gov/WIKI/AtmoCor.html`

Franz, B. A., Bailey, S. W., Werdell, J., McClain, C. R., 2007. Sensor-independent approach to the vicarious calibration of satellite ocean color radiometry. Applied Optics 46, 5068–5082.

GCOS-154, Dec. 2011. Systmatic observation requirements for satellite-based data products for climate. Tech. rep., World Meteorological Organization (WMO), 7 bis, avenue de la Paix, CH-1211 Geneva 2, Switzerland.

Gordon, H., Du, T., Zhang, T., 1997. Remote sensing of ocean color and aerosol properties: resolving the issue of aerosol absorption. Appl. Opt. 36, 8670–8684.

Gordon, H., Wang, M., 1994. Retrieval of water-leaving radiance and aerosol optical thickness over the oceans with SeaWiFS: A preliminary algorithm. Appl. Opt. 33, 443–452.

IOCCG, 2010. Atmospheric Correction for Remotely-Sensed Ocean-Colour Products. Reports of the International Ocean-Colour Coordinating Group 10.

Lerebourg, C., Mazeran, C., Huot, J., Antoine, D., 2011. Vicarious adjustment of the MERIS Ocean Colour Radiometry. ATBD 2.24, ACRI.

Mélin, F., Vantrepotte, V., Clerici, M., D'Alimonte, D., Zibordi, G., Berthon, J.-F., Canuti, E., 2011. Multi-sensor satellite time series of optical properties and chlorophyll a concentration in the Adriatic Sea. Progress in Oceanography doi:10.1016/j.pocean.2010.12.001, in press.

Moore, G., Lavender, S., 2011. Case II.S Bright Pixel Atmospheric Correction. Tech. rep., MERIS ATBD 2.7, Issue 5, 2011.
URL `http://envisat.esa.int/instruments/meris/atbd/atbd_2.6.pdf`

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

Müller, D., Krasemann, H., 2012. Product validation and algorithm selection report, part 1 - atmospheric correction. Tech. Rep. AO-1/6207/09/I-LG D2.5, European Space Agency, ESRIN.

Nelder, J. A., Mead, R., 1965. A Simplex Method for Function Minimization. Computer Journal 7, 308–313.

Schiller, H., Doerffer, R., 1999. Neural network for emulation of an inverse model operational derivation of Case II water properties from MERIS data. Int. J. Remote Sensing 20 (9), 1735–1746.

Steinmetz, F., Deschamps, P.-Y., Ramon, D., 2011. Atmospheric correction in presence of sun glint: application to MERIS. Optics Express 19 (10), 9783–9800.

Werdell, P., Bailey, S., 2005. An improved in situ bio-optical data set for ocean color algorithm development and satellite data product validation. Remote sens. Environ. 98, 122–140.

Zibordi, G., Holben, B., Mélin, F., D'Alimonte, D., Berthon, J.-F., Slutsker, I., Giles, D., 2010. AERONET-OC: An overview. Canadian Journal of Remote Sensing 36 (5), 488–497.

Zibordi, G., Holben, B., Slutsker, I., Giles, D., D'Alimonte, D., Melin, F., Berthon, J., Vandemark, D., Feng, H., Schuster, G., Fabbri, B. E., Kaitala, S., Seppälä, J., 2009. AERONET-OC: A Network for the Validation of Ocean Color Primary Products. Journal of Atmospheric and Oceanic Technology 26, 1634–1651.

Appendix A. Short description of processors and their underlying algorithms

A short description of the tested processors and their underlying algorithms is given here with mainly the references of description elsewhere.

Appendix A.1. MEGS8 - The MERIS standard algorithm for atmospheric correction

This algorithm has been developed by Antoine and Morel (2011) for case 1 waters and has been extended to turbid waters by Moore and Lavender (2011).

The atmospheric correction for case 1 water is based on the assumption that the water leaving radiance in the near infrared spectral range > 700nm is very low due to the high absorption of water. The MERIS spectral bands at 708, 753, 778 and 865nm can then be used to determine the path radiance as well as its spectral shape. The path radiance is subtracted from the radiance at top of atmosphere (TOA) to get the water leaving radiance. The transmittance for the downward direction (sun zenith angle) and the upward direction (viewing zenith angle) is determined from the path radiance and used to determine the water leaving radiance.

The determination of the spectral shape of the path radiance is critical. This is determined by testing different aerosol types iteratively, with additional use of the spectral band at 560nm.

In case of turbid water, the loop of the atmospheric correction is extended by including the water leaving radiances in the near infrared bands, which are determined by suspended particles with a fixed spectral shape.

Appendix A.2. The ForwardNN algorithm

The algorithm for the determination of water leaving reflectances from top of atmosphere radiances ("atmospheric correction") and the retrieval of water optical properties and concentrations of water constituents is based on an iterative optimisation procedure. Within the loop, neural networks (NN) are used as forward models (s. Fig. A1). The inputs of the NNs are the parameters; the inherent optical properties or the aerosol optical thickness; the outputs are the water leaving radiance reflectance or the top of atmosphere reflectances. In the iteration loop the parameters of the forward models, i.e. the inputs of the NNs, are modified by an optimisation algorithm to achieve a best fit between the measured and computed spectrum with few iterations.

The artificial neural networks are trained with a large simulated dataset of corresponding pairs of top of atmosphere (TOA) and water leaving radiance reflectances (Rw) or pairs of Rw and IOPs respectively, which cover most possible conditions of the atmosphere and water.

For this purpose, optical models were defined for atmosphere and water which cover different atmospheric properties; clear water of the open ocean with different phytoplankton pigment

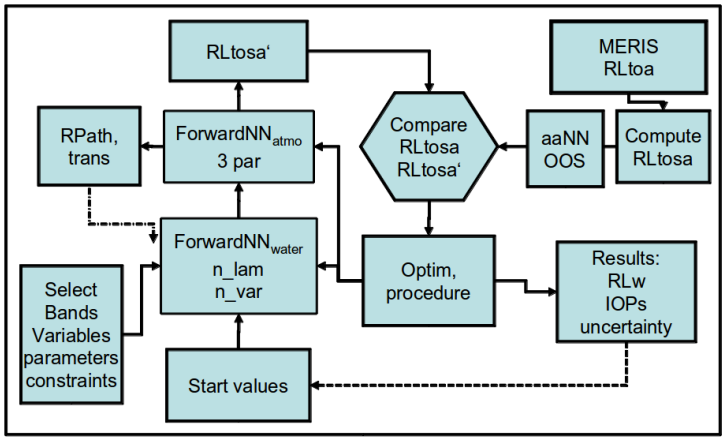


Figure A1: Scheme of ForwardNN algorithm for the atmospheric correction

concentrations; and coastal waters with high concentrations of dissolved and suspended water constituents. Thus, water reflectances can be retrieved from top of atmosphere reflectances over nearly all types of water. However, most critical for the successful applications of NNs are the underlying models of the optical properties of atmosphere and water and also the frequency distribution of the parameters.

The NNs are used for the OC-CCI project to simulate reflectances at 29 wavelengths. This covers the spectral range between 400 and 1020 nm and includes the spectral band sets of MERIS, MODIS, SeaWiFS and OLCI.

In the loop for fitting the observed spectra, the bands of interest can be selected according to the sensor and the importance of bands for a special type of water.

The version of the atmosphere model, which was used for generating the training dataset for the neural networks, included an error in the part for computing the transmittances (wrong angle). We did not wait to obtain a corrected version to demonstrate more clearly the strengths and limitations of processor comparisons.

Appendix A.3. SeaDAS

As SeaDAS we refer in this paper to the processor “l2gen” which is supplied with SeaDAS in version 6.3. “l2gen” is the Multi-Sensor Level 1 to Level 2 processing code of NASA’s Ocean Biology Processing Group (OBPG). The software is used by the OBPG for standard processing of all ocean products.

In the standard atmospheric correction algorithm employed for NASA ocean colour (atmocor2), the TOA radiance is modelled taking into account radiances from Rayleigh scattering by air molecules, the scattering by aerosols (including multiple scattering interactions with the air molecules), the contribution from surface whitecaps and foam, and diffuse and direct transmittances and polarisation. Further information can be found at Franz (2012) including references.

Appendix A.4. POLYMER

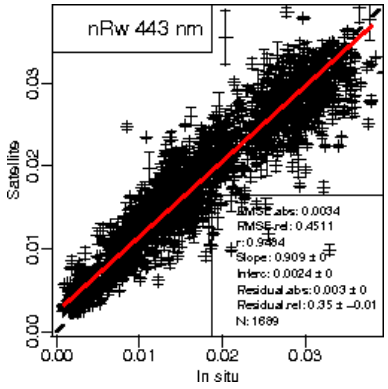
This algorithm has been developed by Steinmetz et al. (2011) for the atmospheric correction of MERIS imagery, and is being extended to other sensors, including MODIS. It has been particularly designed to work in presence of the specular reflection of the sun on the water surface; the sun glint. Atmospheric correction algorithms based on the estimation of the path radiance in near infrared bands usually do not work in these conditions, therefore Polymer leads to a

vastly improved spatial coverage of the oceans. The algorithm is a spectral matching method over the whole available sensor spectrum. It uses two decoupled models: the water reflectance is modelled using two parameters - the chlorophyll concentration and the particles backscattering coefficient, and is mainly based on a semi-analytical model by Morel and Maritorena (2001). The reflectance of the atmosphere, including aerosols and a contamination by the sun glint, is modelled using a simple analytical expression, close to a polynomial, which is the sum of three spectral components of variable amplitude and of fixed spectral dependencies, namely power laws with respective exponents of 0, -1 and -4. The resulting model of the top of atmosphere (TOA) reflectance is therefore described by five parameters, which are optimised to reproduce the measurement in an iterative process using the Nelder-Mead algorithm (Nelder and Mead (1965)). Finally the above-water reflectances are obtained by subtracting the estimated reflectance of the atmosphere and sun glint from the TOA reflectances.

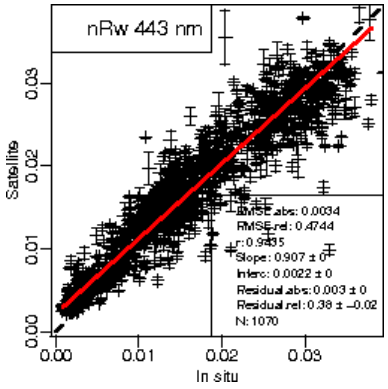
Appendix B. Quality of gained data points under glint condition

The major distinction between MEGS and SeaDAS on one hand and POLYMER and Forward-NN on the other hand is the ability of the latter to use pixels which are highly affected by glint and to gain about 40% more data. To analyse the quality of the gained pixels, specific subsets are selected having MEGS high glint flag raised (or off) while respecting all flags from the POLYMER algorithm. As an example water leaving reflectance at 443nm and as a measure for spectral agreement, the chi-square of spectral agreement to in-situ measurements at five bands is shown in Figure B1.

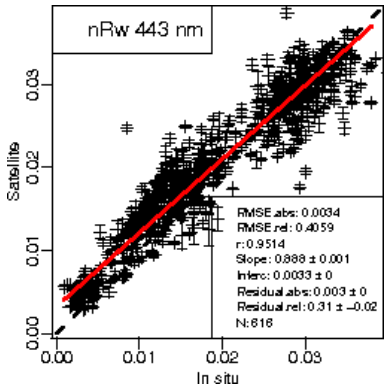
The figures show the correlation of all data points (Fig. B1a). In a second dataset, match-up points with any high-glint influence are excluded (Fig. B1b), while the third figure investigates only sun glint affected data (Fig. B1c). The results for ρ are only slightly deteriorated, some statistical parameters (correlation coefficient, relative residual error, relative RMSE) are even a little better under glint conditions. The spectral behaviour measured with the χ^2 distribution's half maximum width deteriorates slightly compared to those not using pixels under glint (Fig. B1d). All changes between glint and no-glint conditions are smaller than between the algorithms.



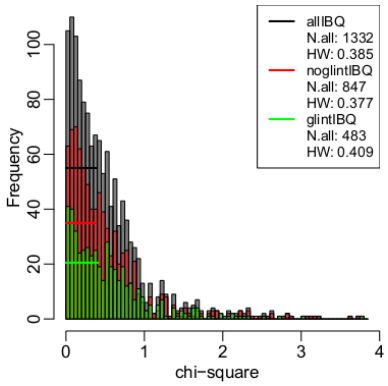
(a) All match-up points (IBQ).



(b) Match-up points without high glint conditions.



(c) Match-up points only in high glint conditions.



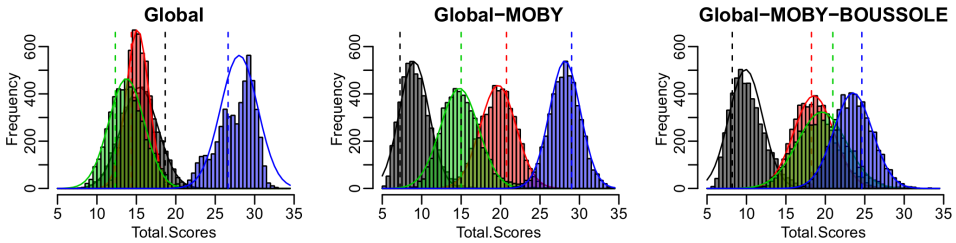
(d) Distribution of χ^2 for IBQ POLYMER 2.4.1 (black). Without high glint affected pixel (red) the half width decreases. Considering only high glint conditions (green) the half width increases.

Figure B1: Influence of sun glint conditions on water leaving reflectances at 443 nm or spectral shape derived with the POLYMER algorithm. Sun glint is identified by the highglint flag of the MEGS 8.0 processor.

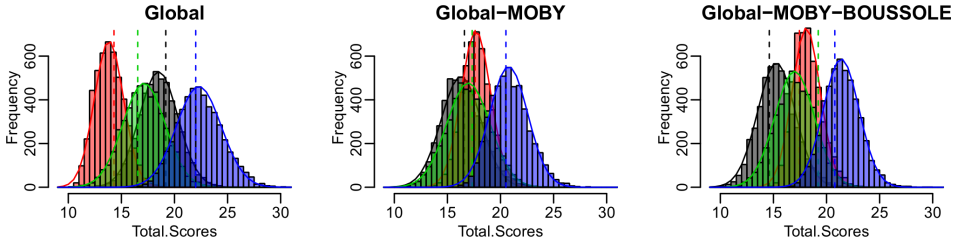
Appendix C. Example of selection with strict case 1 water condition

Considering that MEGS is only defined for case 1 water conditions, the experiment has been repeated with a selection of the database which takes only spectra, as long as the normalised water leaving reflectance of the in-situ measurement at 560 nm is smaller than 0.01.

For individual best quality the influence of vicarious calibration becomes obvious: while MEGS is equal in performance to SeaDAS and ForwardNN for the global dataset, the score lessens if MOBY or BOUSSOLE are removed (Fig. C1). This behaviour can also be recognised in the CBQ results and it reflects the expectation. Due to the width of the score distribution, no processor is significantly better than the others in the CBQ selection.



(a) Individual Best Quality



(b) Common Best Quality

Figure C1: Distribution of total scores from bootstrapping the match-up points with pure case 1 water type of IBQ (maximum N=1303, without MOBY: N=744, without MOBY and BOUSSOLE: N=460) or CBQ (N=387, without MOBY: N=155, without MOBY and BOUSSOLE: N=94). Bootstrapping repetition 5000 times. The dashed lines show the score of the single representation.