ESA Ocean Colour Climate Change Initiative – Phase 3

Product User Guide for v6.0 Dataset

Ref: D4.2
Date: 23/04/2024

Document Properties

Project	:	**Ocean Colour Climate Change Initiative (OC_CCI) – Phase 3**
Document Title:		**Product User Guide for v6.0 Dataset**
Reference	:	**D4.2**
Date	:	**23 April 2024**
Issue	:	**6.5**
	:	**ESA / ESRIN**
Science Lead	:	[Dr Shubha Sathyendranath – PML]
Reviewed	:	[Steve Groom – PML]
ESA Technical Officer	:	[Dr Sarah Connors and Dr Roberto Sabia – ESA]

Address : Plymouth Marine Laboratory (PML)
Prospect Place
The Hoe
Plymouth, PL1 2DH
United Kingdom

Tel: +44 (0)1752 633100

Copyright : © 2024 Plymouth Marine Laboratory. This document is the property of Plymouth Marine Laboratory. It is supplied on the express terms that it be treated as confidential and may not be copied or disclosed to any third party, except as defined in the contract, or unless authorised by Plymouth Marine Laboratory in writing.

Document Revision History

Issue/ Revision	Date	Authors	Comment
6.0	28/10/2022	Thomas Jackson, Shubha Sathyendranath, Steve Groom, Ben Calton	Initial creation from v5.0 PUG with updates for pending v6.0 data release
6.1	04/11/2022	Thomas Jackson	Minor updates
6.2	01/12/2023	Andrei Chuprin	Updated ‘Known issues’ section adding volcano eruption data quality warning
6.3	07/12/2023	Peter Miller	Reviewed and minor updates
6.4	13/12/2023	Andrei Chuprin	Revised according to review
6.5	23/04/2024	Steve Groom, Andrei Chuprin	Revised following ESA review

Table of Contents

1 Acronyms 7

2 In brief 8

2.1 What are these data and what are they intended for? 8

2.2 Where to get the data / how to get more? 8

2.3 What are the key features of the v6.0 dataset compared with previous versions? 8

2.4 When are the next releases and what's in them? 9

2.5 What physical variables are in OC-CCI? 9

2.6 What changes are needed to previous-version programs, so they work when using the new v6.0 data? 11

2.7 Alternative resolutions and climatologies 11

2.8 How do these compare with other non-CCI ocean colour data sets? 12

2.9 Where can I get detailed information? 12

2.10 How to acknowledge the OC-CCI dataset 12

2.11 Where to get support? 13

3 Using the products 13

3.1 Applicability to different water types 13

3.2 Interpreting data values 14

3.3 Understanding the uncertainty estimates 14

3.4 Computation of unbiased data 17

3.5 Statistical Properties of the Chlorophyll Fields 17

3.6 Computation of the uncertainties of composite products 19

3.7 Creating composites of uncertainty variables 19

4 Tools and sample programs 20

4.1 Code snippets in various languages 21

4.1.1 Python 21

4.1.2 R 21

4.1.3 IDL 22

5 Known issues up to time of writing (April 2024) 22

5.1 Major errors 22

5.2 Data errors 23

5.3 Non-errors, but care required by users. 23

5.4 Trivial issues 25

5.5 Informational only 25

5.6 Noteworthy changes from v2.0/v3.0/v3.1/4.2 format ... 26
5.7 Noteworthy changes from v1.0 format (applies to v2.0 onwards) ... 27
6 The products: scientific overview ... 28
6.1 Comparison of OC-CCI v6.0 and v5.0 ... 28
6.2 Product overview ... 33
6.2.1 Chlorophyll-a concentration (mg m^{-3}) ... 33
6.2.2 Remote Sensing Reflectance (sr^{-1}) ... 34
6.2.3 Kd490: the attenuation coefficient for downwelling irradiance (m^{-1}) ... 35
6.2.4 Inherent Optical Properties (IOP): total absorption and backscattering coefficients and their components (a_{tot}, a_{ph}, a_{dg}, b_{bp}) (m^{-1}) ... 35
6.2.5 Uncertainty characterisation ... 37
6.2.6 Optical water classes ... 37
6.2.7 The data-day approach ... 38
7 The products: technical overview ... 39
7.1 General format description ... 39
7.2 Filename convention ... 39
7.3 Example filename ... 40
7.4 Grid format, map projection and coverage ... 41
7.5 Geographic grid format ... 41
7.6 Binned grid format ... 41
7.7 File structure ... 42
7.8 Specific elements of the sinusoidal products ... 42
7.9 Specific elements of the geographic products ... 43
7.10 Product dimensions ... 43
7.11 Flags ... 43
7.12 Geophysical variables ... 44
7.13 Data sources (number of observations) ... 46
7.14 High level metadata ... 47
8 How were the products made? ... 48
8.1 Input datasets ... 48
8.2 Level 2 processing and binning ... 48
8.3 Band shifting ... 48
8.4 Bias correction ... 48
8.5 Merging ... 49

8.6 Water class membership 49
8.7 Product generation 49
8.8 Uncertainty estimation 49
8.9 Reprojection 49
8.10 Additional/derived products 49
9 References 51
10 Earlier versions of OC-CCI dataset 52
V0 (September 2012) 52
V0.9 (May 2013) and v0.95 (July 2013) 52
V1.0rc1 (November 2013) 52
V1.0rc2 / V1.0 (December 2013) 53
V2.0 (April 2015) 53
V3.0 (August 2016) 53
V3.1 (May 2017) 53
V4.2 (May 2019) 53
V5.0 (Oct 2020) 54

1 Acronyms

Acronym	Meaning
ATBD	Algorithm Theoretical Basis Document
BEAM	open-source toolbox and development platform for viewing, analysing and processing of remote sensing raster data
BODC	British Oceanographic Data Centre vocabulary
CAR	Climate Assessment Report
CDR	climate data record
CF	Climate Forecast
CMUG	Climate Modelling User Group
DOI	Digital Object Identifier
ECV	Essential Climate Variables
EO	Earth Observation
ESA	European Space Agency
EUMETSAT	European Organisation for the Exploitation of Meteorological Satellites
FTP	File Transfer Protocol
GAC	Global Area Coverage
HTTP	Hypertext Transfer Protocol
IDEPIX	Identification of Pixel properties
IDL	Interactive Data Language
IOP	Inherent Optical Property
Kd	Diffuse attenuation coefficient
LAC	Local Area Coverage
MERIS	Medium Resolution Imaging Spectrometer
MODIS	Moderate Resolution Imaging Spectroradiometer
NASA	National Aeronautics and Space Administration
NetCDF	network Common Data Form
NOAA	National Oceanic and Atmospheric Administration
Obs4MIPs	Observations for Model Intercomparisons Project
OC-CCI	Ocean Colour Climate Change Initiative (CCI)
OCI, OCI2, OC2, and OCx	Ocean colour algorithms
OLCI	Ocean and Land Colour Instrument
OPeNDAP	Open-source Project for a Network Data Access Protocol
PVASR	Product Validation & Algorithm Selection Report
QAA	Quasi-Analytical algorithm
RMS	Root mean squared
RMSD	Root mean squared deviation
RMSE	Root mean square Error
Rrs	Remote sensing reflectance
SeaDAS	Sensor (SeaWiFS) Data Analysis System
SeaWiFS	Sea-viewing Wide Field-of-view Sensor
SNAP	Sentinel Application Platform
VIIRS	Visible Infrared Imaging Radiometer Suite
WMS/WCS	Web Mapping Service/Web Coverage Service

2 In brief

2.1 What are these data and what are they intended for?

The ESA Climate Change Initiative (CCI) programme is generating a set of validated, error-characterised, Essential Climate Variables (ECVs) from satellite observations. The programme consists of twenty-seven projects, each addressing a particular ECV, complemented by the ESA Climate Modelling User Group CMUG).

The Ocean Colour CCI (OC-CCI) began phase 1 in 2010 with 3 years of initial investigation, ramp up and production of first products, and continued in phase 2 with another 3 years of improvement and annual data releases until 2017. During 2018 and the first quarter of 2019 the OC-CCI was maintained in an 'interim mode' with reduced funding to extend existing datasets (v3.1), research scientific advancement with a view to a future phase and produce a final reprocessing version (v4.2). In 2019 the OC-CCI project moved into the OC-CCI+ phase. This is now focussing on scientific development around the merged multi-sensor record with a focus on OLCI data, product uncertainties, optical water types and algorithm optimisation.

The OC-CCI project provides ocean colour ECV data, with a focus on oceanic and shelf, so-called Case 1, waters where the optical properties are determined by phytoplankton with other optical constituents co-varying with phytoplankton, which can be used, for example, in climate change prediction and assessment models. OC-CCI aims to produce the highest quality data, not containing the very latest data, which may be adjusted in the light of recalibration or assessment. The basic input data are typically calibrated level 1 or level 2, so OC-CCI is ultimately dependent on the controlling agency for the quality of the radiometric and spectral calibration and does the best possible within that limitation.

The latest dataset (v6.0) is created by band-shifting and bias-correcting SeaWiFS, MODIS, VIIRS, Sentinel 3A and 3B OLCI data to match MERIS data, merging the datasets and computing per-pixel uncertainty estimates.

2.2 Where to get the data / how to get more?

All data are available by simple FTP and HTTP and additional, more advanced, data services such as Open Geospatial Consortium compliant WMS/WCS services and OPeNDAP are available. See the data product description page on the OC-CCI website for links to pages detailing these:

https://climate.esa.int/en/projects/ocean-colour/data/

If you wish to acquire data by other means, please contact us (see "Where to get support?" below).

2.3 What are the key features of the v6.0 dataset compared with previous versions?

The v6.0 data is a significant change from previous versions, in that it includes Sentinel 3B OLCI data and the MERIS 4th reprocessing data.

Additionally, several components or pieces of software within the processing chain have been updated. Important updates from v5.0 to v6.0 include:

- Inclusion of the MERIS 4th reprocessing (V5.0 used the MERIS 3rd reprocessing).

- Addition of data from the Ocean and Land Colour Instrument (OLCI) aboard Sentinel 3B starting from 2018-05-15.
- Upgraded Quasi-Analytical algorithm (QAA) used in the band shifting to QAAv6 (the V5.0 data used the QAAv5).
- Minor update to the inter-sensor bias correction.
- Updates to the product masking scheme.
- MODIS and VIIRS data have been dropped from the record after 2019 due to concerns about the continued quality of data from these ageing sensors.
- Temporal extension of the dataset into 2023.

2.4 When are the next releases and what's in them?

OC-CCI+ is now approaching the end of the current phase. The next phase should begin in 2023 and will aim to release a v7.0 dataset in 2025 or 2026. Updates currently envisioned for v7.0 possibly include:

- Inclusion of any updated baseline reprocessing from NASA, NOAA, ESA OR EUMETSAT that have occurred prior to June 2022.
- Updates to the inter sensor bias correction schemes for increased harmonisation of sensors across wavelengths which currently remain troublesome.
- Inclusion of NOAA20 VIIRS data into the merged data record.
- A move to S3A OLCI as the reference sensor.
- Improved sensor error characterisation and uncertainty propagation for better per-pixel uncertainty estimates.

2.5 What physical variables are in OC-CCI?

Data variable	**Accompanying uncertainty variables**	**Notes and further references**
Rrs_412 Rrs_443 Rrs_490 Rrs_510 Rrs_560 Rrs_665	Rrs_412_rmsd Rrs_443_rmsd Rrs_490_rmsd Rrs_510_rmsd Rrs_560_rmsd Rrs_665_rmsd Rrs_412_bias Rrs_443_bias Rrs_490_bias Rrs_510_bias Rrs_560_bias Rrs_665_bias	Remote sensing reflectance at MERIS wavelengths POLYMER ATBD NASA SeaDAS/l2gen documentation (for L1C production)
chlor_a	chlor_a_log10_rmsd chlor_a_log10_bias	Chlorophyll-a estimated using a blended combination of OCI, OCI2, OC2, and OCx algorithms.

Data variable	Accompanying uncertainty variables	Notes and further references
		Blending ATBD In-water RR doc
atot_412 atot_443 atot_490 atot_510 atot_560 atot_665	*Not computed separately, as this is a convenience variable*	QAA total absorption ($a_{ph}+a_{dg}+a_w$, though QAA's decomposition method sometimes does not preserve this property)
adg_412 adg_443 adg_490 adg_510 adg_560 adg_665	adg_412_rmsd adg_443_rmsd adg_490_rmsd adg_510_rmsd adg_560_rmsd adg_665_rmsd adg_412_bias adg_443_bias adg_490_bias adg_510_bias adg_560_bias adg_665_bias	QAA absorption due to detrital and dissolved matter www.ioccg.org/groups/Software_OCA/QAA_v6_2014209.pdf
bbp_412 bbp_443 bbp_490 bbp_510 bbp_560 bbp_665	*Insufficient in-situ data to make a plausible estimate*	QAA backscatter due to particulate matter www.ioccg.org/groups/Software_OCA/QAA_v6_2014209.pdf
kd_490	kd_490_rmsd kd_490_bias	Attenuation coefficient (Lee algorithm with Zhang backscatter coefficients)
water_class1 water_class2 water_class3 water_class4 water_class5 water_class6 water_class7 water_class8 water_class9 water_class10 water_class11 water_class12 water_class13 water_class14	*n/a*	Water class memberships according to Jackson et al. (in prep.) and class definitions per the CCI derivations (broadly, classes range from open ocean to coastal waters as the class number increases) Water class ATBD

2.6 What changes are needed to previous-version programs, so they work when using the new v6.0 data?

For data manipulation such as processing or extraction, there should be no changes necessary, other than accommodating the time series extension and the altered version number. However, if algorithms are applied to the reflectance data, then it is important to consider the set of wavelengths available as this has changed relative to historical datasets (v4 and older).

A couple of small changes are necessary for programs that previously used v1.0 data:

To take account of a new time dimension for all variables: all data-carrying variables are now additionally dimensioned by time (i.e. [time,bin_index] for sinusoidal projection and [time,lat,lon] for geographic projection). As in v1.0, this dimension is of length 1, but may need to be accounted for in product loaders that previously expected a 1 or 2 dimensional product and will now find a 2 or 3 dimensional one. The reason for this change is to increase compatibility with common standards and tools, and to ease the use of languages and tools for aggregating multiple files into a single datacube. For a Python program that previously accessed the chlorophyll variable as:

```
print nc.variables["chlor_a"][:].mean()
```

It would now be:

```
print nc.variables["chlor_a"][0,:].mean()
```

Name changes for uncertainty variables: in v1.0, the names of all variables concerning uncertainty ended in *_bias_uncertainty* or *_rms_uncertainty*. The redundant "*_uncertainty*" component has been dropped and rms clarified to rmsd, meaning that, for example, the associated variables for *aph_412* are now *aph_412_rmsd* and *aph_412_bias*. The uncertainty variables for *chlor_a* are a special case as they are computed using the log10 values and are now *chlor_a_log10_rmsd* and *chlor_a_log10_bias* to provide maximum clarity.

Number of observations variables: the data type of the number-of-observations variables (*total_nobs, MERIS_nobs, MODIS_nobs, SeaWiFS_nobs* and the new *VIIRS_nobs*) has changed. Previously these were integers, reflecting a direct count of the number of observations falling into a cell. In v6.0, they are now floats, meaning that there may be "partial" observations from a sensor into a cell. This change is driven by the change of the binning algorithm to a supersampling one, allowing the contribution of a sensor observation falling across multiple cells to be properly accounted for in each cell.

2.7 Alternative resolutions and climatologies

OC CCI version 6 is provided at 4 km spatial resolution. As an ECV and climate data record (CDR) the OC-CCI dataset is of interest to a number of different users who might wish to work with the data at alternative resolutions. Previous OC-CCI data have already been re-gridded to a 0.5 degree resolution, suitable for comparison to climate models outputs for assessments such as those of the IPCC, as part of the Obs4MIPs project (https://pcmdi.github.io/obs4MIPs/ , https://www.wcrp-climate.org/modelling-wgcm-mip-catalogue/modelling-wgcm-mips/266-modelling-wgcm-catalogue-obs4mips). Version 6 has also now been re-gridded for Obs4MIPS. The coarser resolution data are made available on Obs4MIPs under the variable name 'chl'. We have also created climatologies of the OC-CCI data and re-gridded (for example to 9km) products are available on request from the OC-CCI team on help@esa-oceancolour-cci.org). A global 1km version has also been released, though this only contained a subset of the complete variable set (chlorophyll-a and remote sensing reflectances).

2.8 How do these compare with other non-CCI ocean colour data sets?

For a comparison of v6.0 data against earlier versions of OC-CCI, please see section 6.1.

Other related ocean-colour datasets include:

- GlobColour: merged and sensor products, with a near-real-time focus – http://globcolour.info
- MEaSURES: NASA-sponsored multi-sensor products from University of California, Santa Barbara - http://wiki.icess.ucsb.edu/measures
- Individual sensor products from the space agencies (e.g., MODIS, MERIS)

The primary focus of OC CCI is producing a full time series of consistent measurements for climate science purposes. For fuller comparisons, please see the Climate Assessment Report or the peer-reviewed publications linked on: https://climate.esa.int/en/projects/ocean-colour/key-documents/

2.9 Where can I get detailed information?

All project documentation and related publications can be found at the website: https://climate.esa.int/en/projects/ocean-colour - (there are menu items for documents and publications)

In addition to this Product User Guide, the most relevant documents are:

- Algorithm Theoretical Basis Documents (ATBDs) for the various major components, such as POLYMER, bias correction, band-shifting.
- System Prototype Specification, which describes the processing chain
- Input Output Data Definition, briefly overviewing data formats
- Product Validation and Algorithm Selection Report, which gives the evaluation and analysis leading to the selection of the algorithms used.

External documents that are particularly noteworthy are:

- The Climate Forecast (CF) NetCDF conventions (version 1.11) – http://cfconventions.org/
- Unidata Discovery Metadata Conventions - http://www.unidata.ucar.edu/software/thredds/current/netcdf-java/metadata/DataDiscoveryAttConvention.html (deprecated in favour of the broadly similar Attribute Convention for Data Discovery http://wiki.esipfed.org/index.php/Attribute_Convention_for_Data_Discovery)

2.10 How to acknowledge the OC-CCI dataset

When using the OC-CCI dataset within peer-reviewed papers or any other publications, we politely request the following text is included in the acknowledgements.

The European Space Agency (ESA) Climate Change Initiative (CCI) Ocean Colour data are developed and produced through funding and resources from ESA, Copernicus Climate Change Service (C3S) and the NERC EO Data Analysis and AI Service (NEODAAS).

The correct reference for OC-CCI products is:

Sathyendranath, S, Brewin, RJW, Brockmann, C, Brotas, V, Calton, B, Chuprin, A, Cipollini, P, Couto, AB, Dingle, J, Doerffer, R, Donlon, C, Dowell, M, Farman, A, Grant, M, Groom, S, Horseman, A, Jackson, T, Krasemann, H, Lavender, S, Martinez-Vicente, V, Mazeran, C, Mélin, F, Moore, TS, Müller, D, Regner, P, Roy, S, Steele, CJ, Steinmetz, F, Swinton, J, Taberner, M, Thompson, A, Valente, A, Zühlke, M, Brando, VE, Feng, H, Feldman, G, Franz, BA, Frouin, R, Gould, Jr., RW, Hooker, SB, Kahru, M, Kratzer, S, Mitchell, BG, Muller-Karger, F, Sosik, HM, Voss, KJ, Werdell, J, and Platt, T (2019) An ocean-colour time series for use in climate studies: the experience of the Ocean-Colour Climate Change Initiative (OC-CCI). Sensors: 19, 4285. doi:10.3390/s19194285

In addition, each version of OC-CCI products has its own DOI. Please provide reference to the appropriate DOI:

- v1.0: http://dx.doi.org/10.5285/E32FEB53-5DB1-44BC-8A09-A6275BA99407
- v2.0: http://dx.doi.org/10.5285/b0d6b9c5-14ba-499f-87c9-66416cd9a1dc
- v3.1: http://dx.doi.org/10.5285/9c334fbe6d424a708cf3c4cf0c6a53f5
- v4.2 http://dx.doi.org/10.5285/d62f7f801cb54c749d20e736d4a1039f
- v5.0 http://dx.doi.org/10.5285/1dbe7a109c0244aaad713e078fd3059a
- v6.0 http://dx.doi.org/10.5285/5011d22aae5a4671b0cbc7d05c56c4f0

We would also appreciate being notified so that we can list publications at:
https://climate.esa.int/en/projects/ocean-colour/publications/

2.11 Where to get support?

Feedback and questions regarding the use of the OC-CCI data are welcome – please email: help@esa-oceancolour-cci.org

Contact details for other purposes are at:
https://climate.esa.int/en/projects/ocean-colour/contacts/

3 Using the products

These are state-of-the-art products, and as such, not likely to be error free, even though the OC-CCI team has put in considerable effort to check the quality of the product and to eliminate problems as and when they were found. The OC-CCI team will continue to work on improving the products and data delivery, but it is recognised that wider community usage will provide valuable feedback to improve the products further. Please let us know what works, what does not work and if you find anything that looks like an error.

3.1 Applicability to different water types

The initial focus of OC-CCI was on primarily Case-1 waters; however, the in-situ data sets used in the round robin to choose the in-water algorithm did not exclude data from waters (sometimes called Case-2) where other optically active constituent do not co-vary with phytoplankton. Furthermore, the in-situ data used to compute the pixel uncertainties excluded only waters of depth < 10m. Hence, although the products were primarily designed for application in Case-1 waters, there was some applicability to Case-2. From v3.0

onwards, extra consideration has been given to Case-2 retrievals, flagging and algorithm choice (based on water type), so that the validity of the products will be enhanced.

The blended chlorophyll algorithm used in v3.0, to v6.0 attempts to weight the outputs of the best-performing algorithms based on the water types present, which improves performance in Case-2 waters compared to earlier versions that were mostly open-ocean focussed.

The optical classification of pixels provides some indication of whether the pixel is likely to belong to Case-1 or Case-2 waters: by inspecting the water class spectra, one can determine that some are clearly high-scattering Case-2 waters, and others Case-1 open ocean. Lower-numbered classes cover larger numbers of pixels and, as a rule of thumb, are more associated with open ocean, while higher-numbered classes tend to be more coastal.

Due to the complexity and variety of Case-2 waters and the need for a global solution, customised/regional algorithms such as available from Copernicus Marine or other data providers will still offer better performance, but global scale analyses will find improved results in v3.0 onwards.

3.2 Interpreting data values

Upper and lower limits to the products have been applied, in all waters, based on what we know to be realistic in Case-1 waters and the range in the values used for validation and uncertainty characterization.

The following filters have been applied:

- Chlorophyll: all values less than 0.001 have been set to 0.001 mg/m^3, and values greater than 100 have been set to 100 mg/m^3
- Inherent Optical Properties: all values greater than 10 m^{-1} have been discarded.
- Effect of high path-length of light through the atmosphere: we have used air mass (sum of the inverse of the cosines of satellite viewing angle the sun zenith angle) to filter data that might have been affected by high path-length of light through the atmosphere. Data corresponding to air mass greater than 5 have been eliminated from the products. This value was adopted as a compromise between having some data in high latitudes and reducing errors due to high air mass (see atmospheric correction ATBD for more details).
- A filter was applied to MERIS and MODIS that discriminated and removed pixels with spectral shapes that indicated the presence of high levels of aerosols (primarily Sahara dust). Details on these filters can be found in the SPS document.
- R_{rs}: all negative R_{rs} values have been discarded, except for 665 nm.

3.3 Understanding the uncertainty estimates

The user consultation undertaken at the beginning of the OC-CCI project revealed that the user community required uncertainty estimates that are based on validation of the products against matched in-situ observations. All products, except particle back-scattering coefficient, are therefore accompanied by uncertainty characteristics at every pixel. The uncertainties provided are the root-mean-square difference (RMSD, Δ) and bias (δ), computed on the basis of match-up in-situ data. They provide estimates of the extent to which the satellite observations are likely to differ from in-situ observations. They were estimated first for

each of the optical water classes identified, and then assigned to each pixel, also on the basis of optical water classes: using OC-CCI remote-sensing reflectance spectra (R_{rs}) for each pixel, the fuzzy membership of each optical class for that pixel at that time was calculated. Then, the uncertainties for that pixel were computed as weighted averages of the uncertainties of the classes for that pixel. The fuzzy membership was used as the weighting factor for each class. The fuzzy logic method used for optical classification and uncertainty assignment follows the work of Moore et al. (2009) and Jackson et al. (2017). The v6.0 optical classification includes 14 classes, generated specifically for the V5.0 products. The primary difference between v1.0 OC CCI and all subsequent versions (i.e., v2.0 onwards) is that in the later data sets the optical classification is based on OC-CCI satellite R_{rs} data as input to the classification, instead of in-situ R_{rs} data. For further details regarding optical classification, please see section 8.6.

The root-mean-square difference (Δ_p) for each product, for each pixel *p*, is given by:

$$\Delta_p = \sqrt{\frac{\sum_{k=1}^{K} w_{k,p}\Delta_k^2}{\sum_{k=1}^{K} w_{k,p}}} \qquad \text{(Equation 2.1)}$$

Where *k=1…K* are the optical classes, Δ_k is the root-mean-square difference for each class, and $w_{k,p}$ are the weighting factors, (fuzzy memberships) of each of the *K* optical water classes for that pixel.

The bias δ_p is computed similarly:

$$d_p = \frac{\sum_{k=1}^{K} w_{k,p} d_k}{\sum_{k=1}^{K} w_{k,p}} \qquad \text{(Equation 2.2)}$$

The unbiased, or centred, root-mean square difference (which is the same as the standard deviation), can be computed as:

$$\sigma_p = \sqrt{\left|\Delta_p^2 - \delta_p^2\right|} \qquad \text{(Equation 2.3)}$$

Note that these estimates are only as good as the quality and representativeness of the in-situ match up data sets that were available for uncertainty estimation. Geographical coverage and representation of different water types were best for chlorophyll (See Figure 1), followed by K_d and then by R_{rs}.

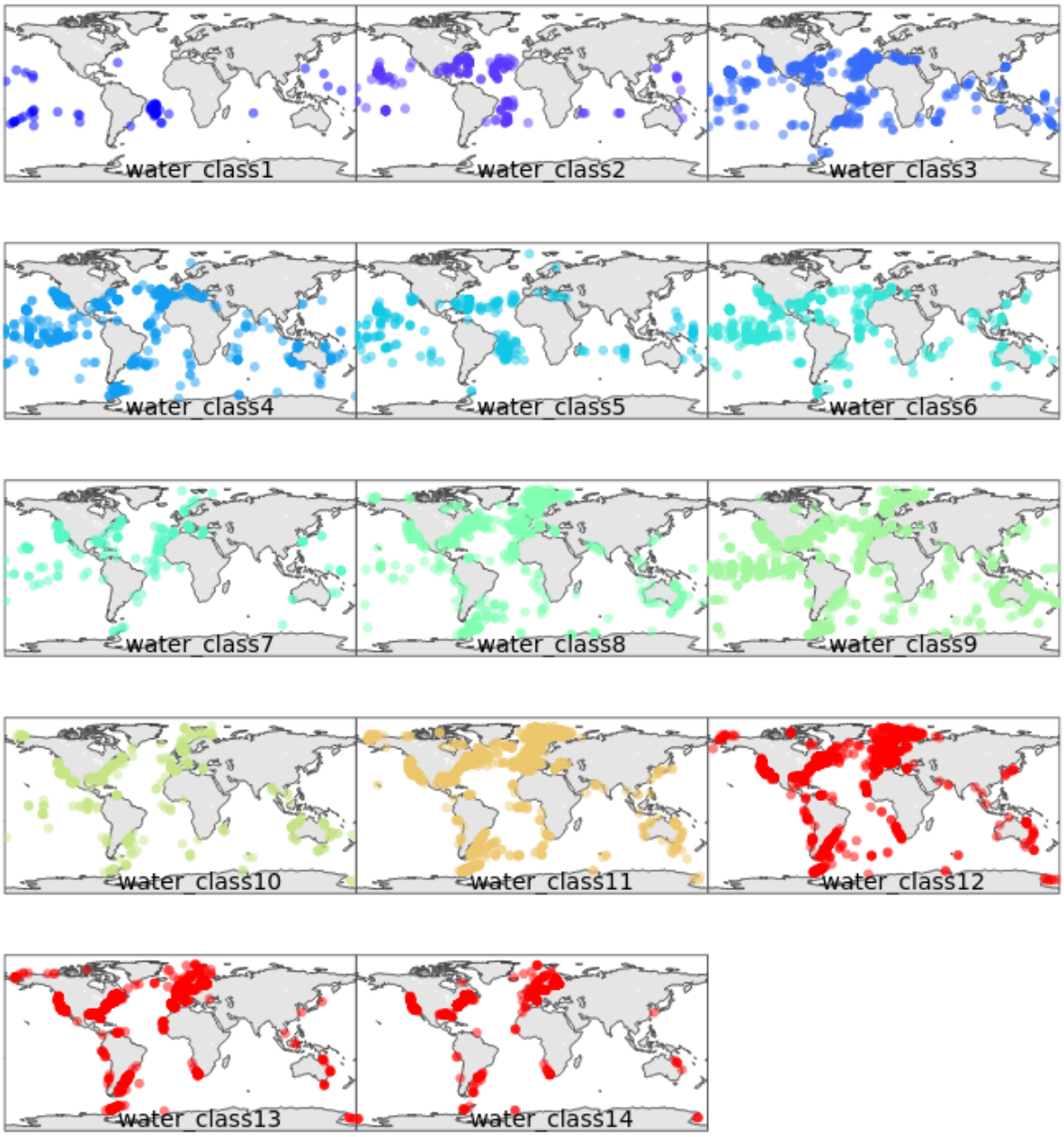


Figure 1: Geographical coverage of water types for Chl-a matchups: see section 8.6 and ATBD for description of water classes.

Table 1: Number of Chlorophyll-a matchups dominated by each waterclass membership. Locations for v6 matchups are shown in Figure 1.

Water Class	***1***	***2***	***3***	***4***	***5***	***6***	***7***	***8***	***9***	***10***	***11***	***12***	***13***	***14***	***All***
V5	*43*	*268*	*763*	*744*	*345*	*463*	*231*	*1376*	*2027*	*697*	*3807*	*3307*	*3029*	*802*	*17902*
V6	*54*	*333*	*787*	*748*	*403*	*458*	*234*	*1603*	*2021*	*532*	*4010*	*3855*	*3124*	*913*	*19075*

3.4 Computation of unbiased data

The uncertainty statistics provided with the v6.0 products allow the user to compute unbiased values of the products. Unbiased values $\hat{X}_p$ of any variable *X* can be computed as:

$$\hat{X}_p = X_p + \delta_p \quad \text{(Equation 2.4)}$$

In the case of chlorophyll data, δ_p is provided for $\log_{10}$(chlorophyll). So, for the particular case of chlorophyll, the unbiased chlorophyll at pixel *p*, say $\hat{C}_p$, can be found from

$$\hat{C}_p = 10^{(log_{10}(C_p)+\delta_p)} \quad \text{(Equation 2.5)}$$

where C_p is the value of the chlorophyll product at the pixel. Because the satellite observation was subtracted from the in-situ value to compute the bias δ_p, hence, $\delta_p < 0$ implies an overestimation by the satellite. Note, uncertainties can arise from residual inter-sensor differences or issues with the atmospheric correction, inter alia.

3.5 Statistical Properties of the Chlorophyll Fields

It is important to bear in mind that uncertainties in chlorophyll (Δ_p, δ_p, and, hence, σ_p) are reported for data that have been logarithmically transformed, given the well-known log-normal distribution of chlorophyll data. Transformation using common (base 10) logarithms was selected over natural logarithms, because it is the conventional practice in the field. If properties related to natural-log transformations are required, it is easy to convert the standard deviation σ_p included in the products, to its corresponding value σ_e for natural-log-transformed data:

$$\sigma_e = \ln(10)\sigma_p \quad \text{(Equation 2.6)}$$

A similar relationship exists between μ_p, the mean of the common-log-transformed distribution and the corresponding mean μ_e for the natural-log-transformed data:

$$m_e = \ln(10)m_p \quad \text{(Equation 2.7)}$$

The satellite chlorophyll products themselves are reported as untransformed values, and, therefore, caution must be exercised when combining the observations and the uncertainties.

If we assume that the chlorophyll product after bias correction, $\hat{C}_p$, represents the expected or mean value m_p of untransformed chlorophyll data that follow a log-normal distribution at that pixel at that time, then it is related to μ_e according to the equation below:

$$m_e = \ln(m_p) - \frac{1}{2}\sigma_e^2 \quad \text{(Equation 2.8)}$$

or,

$$m_p = \log_{10}(m_p) - \frac{1}{2}\ln(10)\,\sigma_p^2 \quad \text{(Equation 2.9)}$$

such that μ_e can be estimated from the quantities provided. This equation clearly shows that the mean of the log-transformed data (μ_p) is different from the logarithm of the mean of the untransformed data (m_p). The standard deviation s_p of the corresponding untransformed log-normal distribution is given by:

$$s_p = m_p\sqrt{e^{(\sigma_e^2)} - 1} = m_p\sqrt{e^{([\ln(10)]^2\sigma_p^2)} - 1} \quad \text{(Equation 2.10)}$$

The geometric mean, m_g, of a log-normal distribution is given by

$$m_g = e^{(\mu_e)} = 10^{\mu_p} \quad \text{(Equation 2.11)}$$

and m_g is also equal to the median of the untransformed variable.

Confidence intervals on chlorophyll, calculated for the logarithmically transformed variables, will be symmetric. However, when expressed in terms of untransformed chlorophyll, the confidence limits will not be symmetric about the arithmetic mean chlorophyll. As an example, suppose the confidence interval is defined as two standard deviations on either side of the mean, that is $\mu_p \pm 2\sigma_p$ expressed in terms of $\log_{10}$-transformed chlorophyll. To back-transform, or write this interval in terms of untransformed chlorophyll, we proceed as follows. The upper confidence limit is $(\mu_p + 2\sigma_p)$ in the transformed units and, therefore, $10^{(\mu_p \pm 2\sigma_p)}$ in the untransformed units. Similarly, the lower confidence limit is $(\mu_p - 2\sigma_p)$ in the transformed units or $10^{(\mu_p \pm 2\sigma_p)}$ in the untransformed units. Confidence limits calculated in this fashion lie about the quantity 10^{μ_p}, which is the geometric mean (or the median) of the untransformed chlorophyll values.

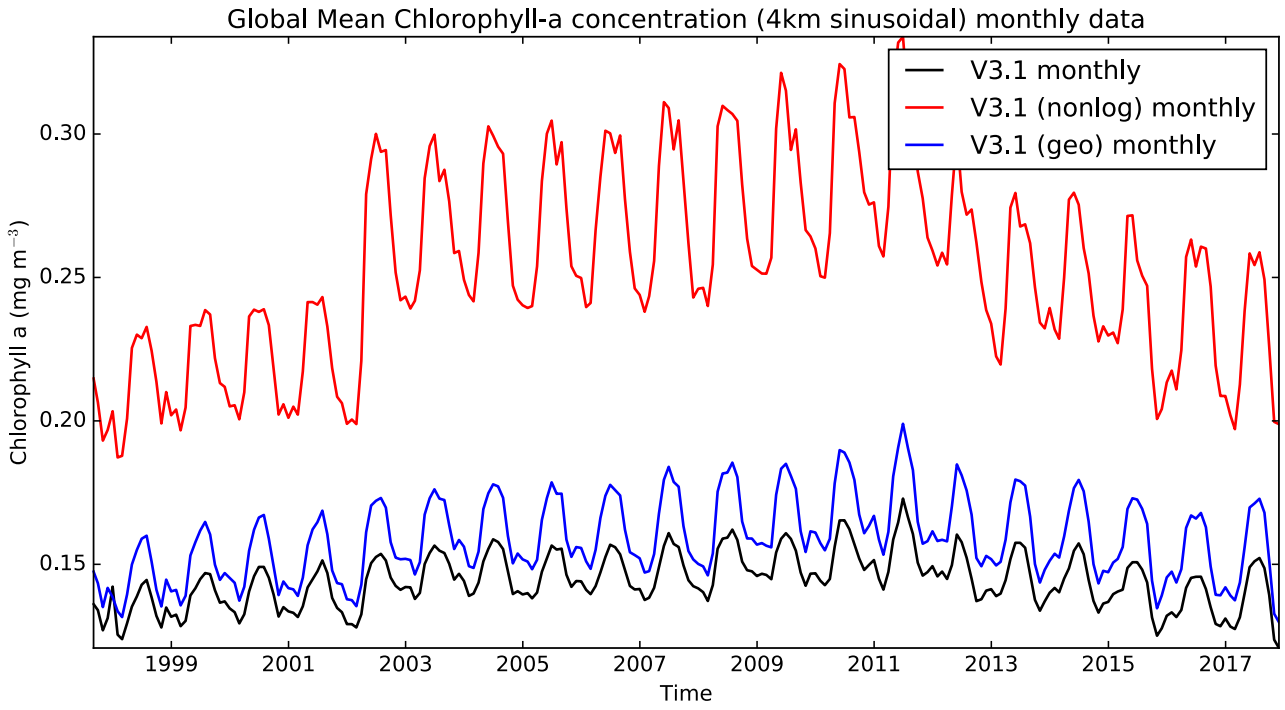


Figure 2: Comparison of Global mean Chlorophyll-a calculated using 1) the correct approach, log(chl) from sinusoidal projection files (black line), 2) chl-a from geographic projection files (red line) and 3) log(chl) from the geographic projection files.

It is also of note that the log-normal distribution of chlorophyll-a values means that users must take care when producing summary statistics for purposes such as time-series analysis or intercomparison purposes. For example, calculating a global mean chlorophyll-a should be done by logging the chlorophyll-a data prior to averaging and then unlogging. Additionally, it should be noted for all products that the geographic projection netcdf files will add additional weight to the high latitude regions compared to the sinusoidal projection data, due to the spatial distortion at high latitudes. Figure 2 shows an example of the impact of these two statistical distortions at the global scale. It is clear that the impact of not logging the data is more significant than the impact of using geographic rather than sinusoidal files. In the example given, the increase in coverage for coastal regions with the addition of POLYMER processed MERIS data appears as a significant jump in the mean chl-a record if you do not use the correct (log(chl)) approach. This is clearly the influence of

disproportionally small number of very high chlorophyll-a values as no such jump is seen in the log(chl) based mean estimate.

3.6 Computation of the uncertainties of composite products

The uncertainties of the data obtained by aggregating (compositing) the pixel values in space and time can be computed from the values provided by equations 2.1 - 2.3. The product value X_c, for a composite *c* of *N* pixels can be computed as:

$$X_c = \frac{\sum_{p=1}^{N} X_p}{N} \quad \text{(Equation 2.12)}$$

the composite root-mean-square deviation as:

$$\Delta_c = \sqrt{\frac{\sum_{p=1}^{N} \Delta_p^2}{N}} \quad \text{(Equation 2.13)}$$

and the bias as:

$$\delta_c = \frac{\sum_{p=1}^{N} \delta_p}{N} \quad \text{(Equation 2.14)}$$

By analogy with equation 2.3, the standard deviation of the composite value can be computed as:

$$\sigma_c = \sqrt{|\Delta_c^2 - \delta_c^2|} \quad \text{(Equation 2.15)}$$

With reference to chlorophyll product, the statistical properties can be computed by analogy with the case for the individual pixels given in previous sections.

Equations 2.12 - 2.15 were used to compute the eight-day and monthly composites in the data release. They can also be applied by the user for spatial or temporal re-gridding of the data to any user-required scale.

3.7 Creating composites of uncertainty variables

There are some statistical complexities involved in making a composite of the uncertainty variables – a simple average is not appropriate. Instead, please use the method described below.

When composites are generated, there will be a number N_v of valid pixels in each bin, each with errors characterised by RMSD Δ_p, bias m_p, standard deviation σ_p and water class membership W_p. Then the uncertainties in the composite product can be computed as:

$$\Delta_c = \sqrt{\frac{\sum_{i=1}^{N_v} \Delta_p^2}{N_v}}$$

$$m_c = \frac{\sum_{i=1}^{N_v} m_p}{N_v}$$

$$\sigma_c = \sqrt{\frac{\sum_{i=1}^{N_v} \sigma_p^2}{N_v}}$$

$$W_c = \frac{\sum_{i=1}^{N_v} W_p}{N_v}$$

4 Tools and sample programs

OC-CCI products are provided in NetCDF format, so can be ingested with all NetCDF compatible software packages. Note that the NetCDF library used must be version 4.0.0 or higher (first released 2008, so default installs on modern systems are likely to be sufficient) to support transparent internal compression and to read the products. Examples include the NetCDF operators, ncview, the Python netCDF4 library and R's netcdf package

The analysis package currently recommended is the SeNtinel Application Platform (SNAP) toolbox, specifically developed by ESA for the exploitation of Earth Observation data products. SNAP is open source and freely available from http://step.esa.int/main/toolboxes/snap/. Regarding the OC-CCI products, SNAP could be used for example to:

- view the images and metadata
- create regional subsets
- investigate the products by creating statistics, histograms, and scatter plots
- perform image analysis (e.g. clustering)
- validate ocean colour data by comparison with in-situ or any other kind of reference data
- analyse time series using the time series tool that is part of SNAP (see screenshot below)
- undertake band arithmetic using a fast expression language

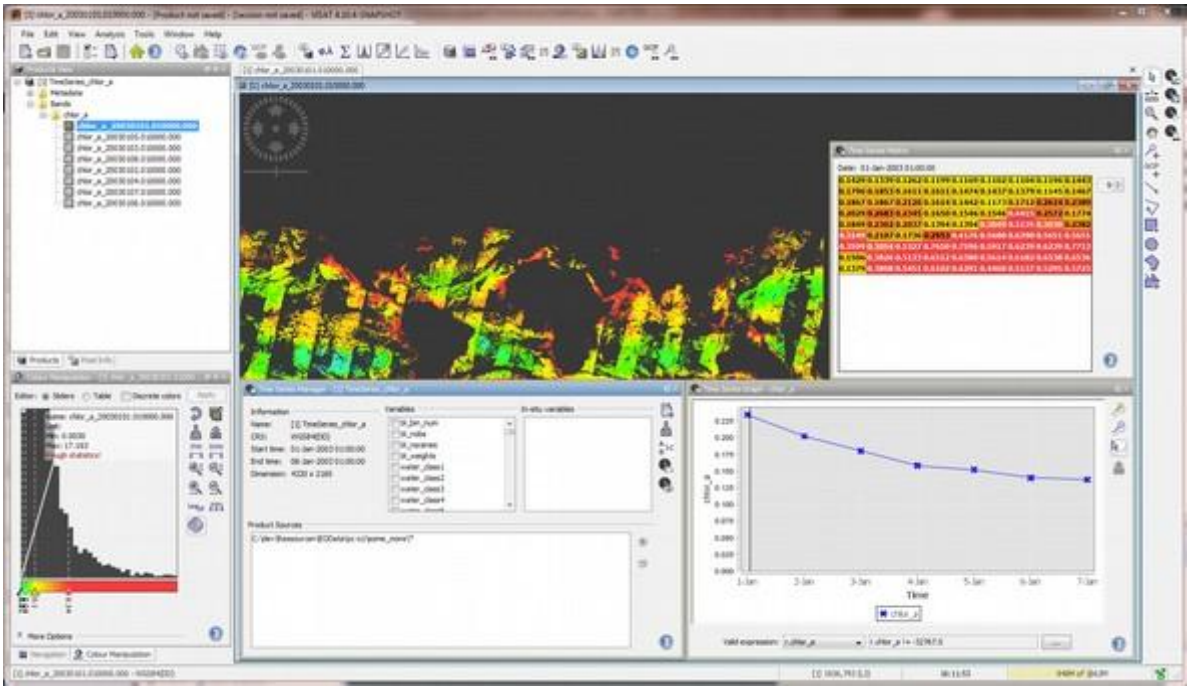


Figure 3: Viewing and manipulating data in SNAP

At the time of writing this section (April. 2024) the current released version of SNAP (9.0, with all updates applied) in its default configuration is capable of reading the v6.0 sinusoidal and geographic formats without difficulties.

An alternative for working with the OC-CCI products is the SeaDAS Visualization tool. The core visualization package for SeaDAS 7 and up is the result of a collaboration between NASA and Brockmann Consult, and is

based on the BEAM/SNAP framework with extensions that provide the functionality of previous versions of SeaDAS (extra functions for NASA products). SeaDAS versions 7.3 and up handle both CCI projections.

Additionally, the Panoply data viewer is available from NASA free of charge at http://www.giss.nasa.gov/tools/panoply. However, no graphical display of the sinusoidal OC-CCI data products is possible since the tool does not support their geolocation encoding.

4.1 Code snippets in various languages

Please see the website for more examples.

4.1.1 Python

There are several NetCDF capable libraries, but PML most commonly uses “netCDF4” (available from https://github.com/Unidata/netcdf4-python or using “pip install netCDF4”), which interfaces well with numpy. A brief example of usage:

```
import netCDF4
nc = netCDF4.Dataset("/path/to/CCI/year/file.nc", "r")

# Display some global attributes
print nc.time_coverage_start
print nc.license

# Take the mean of a global variable
print nc.variables["chlor_a"][:].mean()
```

Another python package that is very useful for rapid access and investigation of the data files is the xarray package (http://xarray.pydata.org/en/stable/index.html). As the OC-CCI files are CF compliant they are compatible with the tools available through xarray.

4.1.2 R

As with Python there are several NetCDF packages in R but we recommend “ncdf4”, which can be added to your R build using install.packages(‘ncdf4’) and added to your session using library(‘ncdf4’). A brief example of using R to perform the same task as completed in the python example:

```
library('ncdf4')
nc=nc_open("/path/to/CCI/year/file.nc")

# Display a list of available variables
names(nc$var)

# Extract global chlorophyll-a data
v1<-ncvar_get(d1, d1$var$chlor_a)

# Close netcdf
nc_close(d1)

# Take the mean of the global chlorophyll-a variable
mean(v1, na.rm=T)
```

4.1.3 IDL

A brief example of using IDL to perform the same task as above:

```
; Open the file and assign it a file ID
fileID = ncdf_open("/path/to/CCI/year/file.nc", /read)

; Find the number of file attributes and variables in the netCDF
nc_struct=ncdf_inquire(fileID)
nvars = nc_struct.nvars
print, nvars

; List all variable names
for i=1,nvars-1 do print, NCDF_VARINQ(fileID,i)

; Find the variable id associated with a required variable
chlor = NCDF_VARID(fileID, 'chlor_a')

; Import the dateset for selected variable
varID=chlor
ncdf_varget,fileID,varID,variable

; When done with file, close it.
ncdf_close, fileID

; Replace all fill values with nan
i_nan = where(variable eq 9.96921e+36, /null)
variable[i_nan]='nan'

; Calculate the mean chlorophyll
print, mean(variable, /nan)
```

5 Known issues up to time of writing (April 2024)

This section lists all known issues with the data, as well as any characteristics commonly perceived as an issue, with notes on mitigations and impacts. Please note this list aims to be comprehensive and, thus, covers many minor issues.

In the event of minor correctable errors, errata will be made available for download. In the event of a major error being discovered, a new release would be made with the correction incorporated. In such circumstances please contact us at help@esa-oceancolour-cci.org.

5.1 Major errors

None found so far.

5.2 Data errors

None found so far

5.3 Non-errors, but care required by users.

Valid product pixels may have no matching uncertainty values: there are valid product pixels that have no matching uncertainty values. This is because the pixels are insufficiently well represented by any water type (typically below 1% for unusual waters) and thus any uncertainty computed based on class membership would be a very poor estimate. These pixels will be relatively uncommon / few through the time series, though they may include noteworthy pixels such as coccolithophore blooms.

Decay of MODIS calibration and r2018 reprocessing: NASA monitor the calibration of MODIS (Aqua) and regularly adjust it. The last processing was r2018.0, building on r2014.0.1, both working to correct some of the degradation noted. The recent OC-CCI products (v3.0 onwards) are less vulnerable to this degradation than v2.0, as they incorporate VIIRS (now stated by NASA to be more reliable than MODIS Aqua).

The r2018.0 reprocessing is, in fact, a multi-mission reprocessing to incorporate updates in instrument calibration and vicarious calibration for VIIRS on SNPP, MODIS on Aqua and Terra, and SeaWiFS.

OC-CCI data affected by strong natural events: often Earth observation data are severely affected by such natural events like solar eclipses, volcano eruptions, strong sandstorms, and similar conditions. The strongest though short-lasting effect cause solar eclipses; to mitigate their impact, whole granules and half-orbits have been removed. Volcano eruptions typically have a smaller spatial coverage, but their impact can be observed for days or weeks. Sandstorms and volcanos can both act to fertilise phytoplankton blooms but ocean colour data should be treated by users with caution: they show increased values of chlor-a and IOP variables such as b_{bp} but probably some additional investigation of data quality is required for such cases. One of the strongest eruptions of recent times was Hunga Tonga–Hunga Ha‘apai volcano eruption in 2021-2022: its impact should be taken into account in climate research in the South Pacific area. Chlor-a data for the volcano area (21.5, -19., -176.5, -174.0) for 8 consecutive days from 13th to 20th January 2022, is shown below, starting from top left to right, then bottom from left to right. A chlor-a anomaly is observed starting from January 15th 2022.

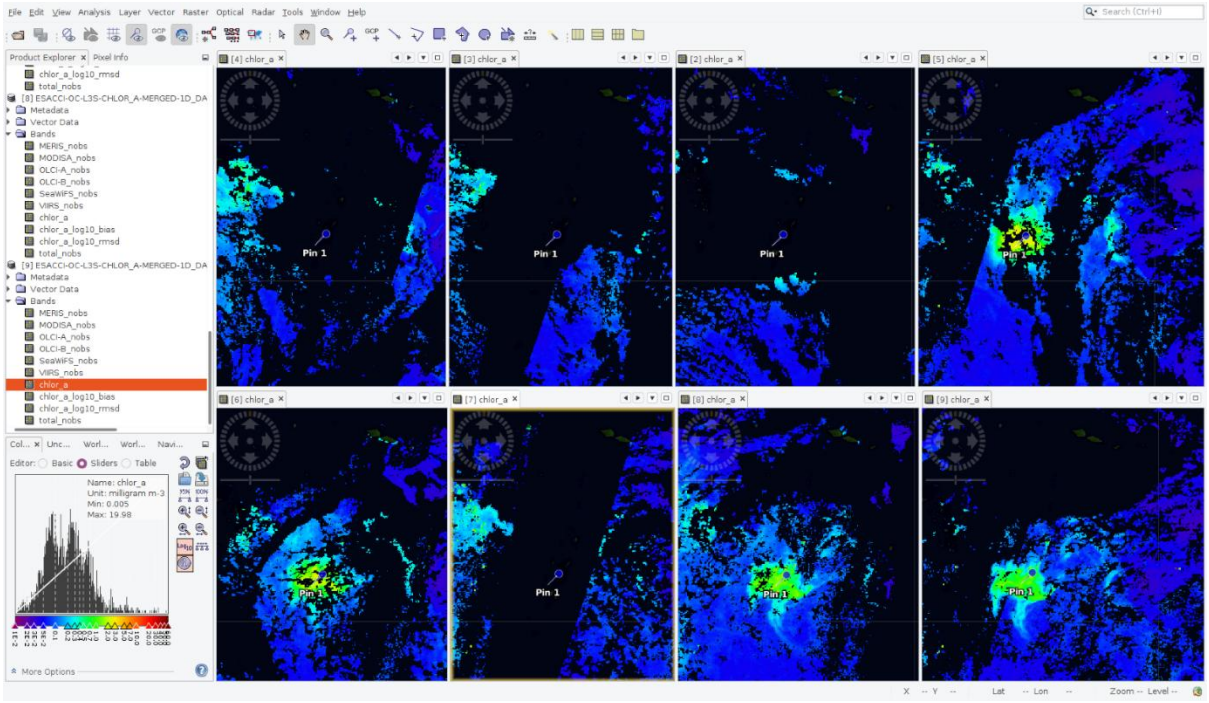


Figure 4 *Chlor-a in 5 day composite OC-CCI data in the area of Tonga-Hunga eruption.*

Number of Observations parameter affected for MERIS and OLCI sensors: MERIS L1B data sets and to lesser extent OLCI L1B data for both S3A and S3B contain files with fully or partially overlapping pixels. These duplicate files share the same orbit numbers and have very close, sometimes identical radiance values for overlapping pixels. If we compare these products processed to L2 and L3bin levels, with ones with removed duplicate files, observed difference in Rrs and chlor-a for corresponding pixels is typically within ±0.1%, in some places reaching 1% and more, usually at high latitudes. The only parameter changing significantly is number of observations: for overlapping pixels it doubles. MERIS dataset is more affected comparing to OLCI: about 15% of L3bin daily files contain such duplicate pixels, for OLCI the number is ~2%.

Rrs Validation Statistics for R2014 Reprocessing

Product Name	#	Mean Bias	Mean Absolute Error (MAE)	Random Error	Aqua Range	In situ Range
rrs412	4136	-0.00080	0.00128	0.00048	-0.00424 - 0.01645	-0.00000 - 0.01531
rrs443	4336	-0.00028	0.00083	0.00055	-0.00141 - 0.02080	0.00007 - 0.01871
rrs488	3926	-0.00067	0.00088	0.00021	-0.00021 - 0.02672	0.00039 - 0.02530
rrs531	2195	-0.00063	0.00083	0.00020	0.00093 - 0.02715	0.00113 - 0.02523
rrs547	3813	-0.00052	0.00079	0.00027	0.00086 - 0.02680	0.00117 - 0.02608
rrs555	3726	-0.00081	0.00096	0.00015	0.00061 - 0.02492	0.00102 - 0.02613
rrs667	3723	-0.00017	0.00030	0.00013	-0.00055 - 0.01334	0.00000 - 0.01416
rrs678	524	-0.00018	0.00035	0.00016	-0.00043 - 0.01007	0.00004 - 0.00904

Rrs Validation Statistics for R2018 Reprocessing

Product Name	#	Mean Bias	Mean Absolute Error (MAE)	Random Error	Aqua Range	In situ Range
rrs412	4171	0.00003	0.00101	0.00098	-0.00310 - 0.01744	-0.00000 - 0.01773
rrs443	4374	0.00006	0.00077	0.00070	-0.00126 - 0.02152	0.00007 - 0.01871
rrs488	3960	-0.00052	0.00078	0.00026	0.00001 - 0.02706	0.00039 - 0.02530
rrs531	2225	-0.00058	0.00079	0.00021	0.00092 - 0.02728	0.00113 - 0.02587
rrs547	3847	-0.00049	0.00077	0.00028	0.00086 - 0.02694	0.00117 - 0.02759
rrs555	3759	-0.00076	0.00091	0.00015	0.00058 - 0.02522	0.00102 - 0.02799
rrs667	3756	-0.00016	0.00029	0.00013	-0.00053 - 0.01339	0.00000 - 0.01416
rrs678	530	-0.00016	0.00034	0.00018	-0.00041 - 0.01008	0.00004 - 0.00904

Figure 5: Improvement in the validation statistics (vs in-situ data) for the r2018 NASA MODIS reprocessing versions. Further information available at https://oceancolor.gsfc.nasa.gov/reprocessing/r2018/aqua/

5.4 Trivial issues

IOP standard_name attributes in the NetCDFs are insufficiently descriptive (no distinction between a_{ph}, $a_{d}g$, etc). This is because improved names have not yet been accepted into the CF standard name list, despite apparent consensus having been reached a few years ago. This can only be resolved when suitable names are accepted into CF.

5.5 Informational only

Known holes in input data: not all days are fully covered due to periods of no data in the input datasets or uncorrectable errors when processing them. This is only of concern during the period when SeaWiFS was the sole instrument. So far, MODIS has not missed a day while it has been the only sensor available. These are the dates where no daily data exists at all:

- 1997-09-05
- 1997-09-07
- 1997-09-08
- 1997-09-11
- 1997-09-12
- 1997-09-13
- 1997-09-14

- 1997-09-17
- 1997-10-13
- 1997-10-14
- 1997-10-15
- 1997-10-16
- 1997-10-17
- 1997-10-18
- 1997-12-15
- 1998-07-10
- 1998-11-17
- 1998-11-18
- 1998-11-19
- 1998-11-20
- 1998-12-17
- 1999-01-25
- 1999-11-17
- 1999-11-18
- 2000-11-17
- 2001-11-18

There are also days with partial data (e.g., some missing MODIS granules). These typically occurred due to errors in the input data (e.g., a geolocation issue) or problems with one of the processing algorithms (e.g. the flagging or A/C algorithm may fail to generate a result in exceptional circumstances). In these cases, the granule was omitted, and a small gap may appear in the output data. Although there are not many of these instances, it is not worthwhile listing them here. Please contact us if you need a precise list.

Remaining bias: while every effort has been made to remove bias and minimize the difference between sensors, some inevitably remains. Users should be aware of the start and end times of the sensors used (SeaWiFS from September 1997 to December 2010, MERIS from April 2002 to April 2012, Aqua-MODIS from July 2002 to Dec 2019, S-VIIRS from 2012 to 2019, S3A OLCI from May 2016 and ongoing, and S3B OLCI from June 2018 and ongoing.

No uncertainty for a_{tot} and b_{bp}: a_{tot} is provided purely for convenience (being a combination of a_{ph}, a_{dg} and a fixed a_w) and we chose not to create unnecessary uncertainty variables that would only inflate file size: this is important as data volumes increase and the impact of storage of these files remains a contributor to carbon emissions.

There were insufficient in-situ data to provide more than a handful of matchups per water class for b_{bp}, so there are no RMSD and bias estimates; this will only be resolved by a larger in-situ database, requiring more cruises/collections in future.

Water classes don't sum to 1: this is an intentional feature of the water classification stage. Since a limited number of classes was used, they are not fully representative of all possible water types (meaning they may not reach a total membership of 1). Please see the section above on uncertainty for more detail.

5.6 Noteworthy changes from v2.0/v3.0/v3.1/4.2 format

v6.0 is fully compatible with v3.0/3.1/4.2/5.0. It is essentially fully compatible with v2.0 in terms of structure; the only change is to a non-core class of variables:

Number of observations variables: the data type of the number-of-observations variables (*total_nobs, MERIS_nobs, MODIS_nobs, SeaWiFS_nobs*, *VIIRS_nobs, OLCI-A_nobs,* and the new *OLCI-B_nobs*) has changed.

5.7 Noteworthy changes from v1.0 format (applies to v2.0 onwards)

For those moving directly from v1.0 to v6.0, a few small changes may be required to programs as detailed below. There are no format changes between v2.0, v3.0 and v3.1 and v6.0.

Addition of the time dimension to all variables: all data-carrying variables are now additionally dimensioned by time (i.e. [time,bin_index] for sinusoidal projection and [time,lat,lon] for geographic projection). As in v1.0, this dimension is of length 1, but may need to be accounted for in product loaders that previously expected a 1 (sinusoidal) or 2 (geographic) dimensional product and will now find a 2 or 3 dimensional one. The reason for this change is to increase compatibility with common standards and tools, and to ease the use of languages and tools for aggregating multiple files into a single datacube. For a Python program that previously accessed the chlorophyll variable as:

```
print nc.variables["chlor_a"][:].mean()
```

It would now be:

```
print nc.variables["chlor_a"][0,:].mean()
```

Name changes for uncertainty variables: in v1.0, the names of all variables dealing with uncertainty ended in *_bias_uncertainty* or *_rms_uncertainty*. The redundant _uncertainty component has been dropped and rms clarified to rmsd, meaning that, for example, the associated variables for *aph_412* are now *aph_412_rmsd* and *aph_412_bias*. The uncertainty variables for *chlor_a* are a special case as they are computed using the log10 values, and are now *chlor_a_log10_rmsd* and *chlor_a_log10_bias* to provide maximum clarity.

Number of observations variables: the data type of the number-of-observations variables (*total_nobs, MERIS_nobs, MODIS_nobs, SeaWiFS_nobs* and the new *VIIRS_nobs*) has changed. Previously, these were integers, reflecting a direct count of the number of observations falling into a cell. Since v3.0, they are now floats, meaning that there may be “partial” observations from a sensor into a cell. This change is driven by the change of the binning algorithm to a supersampling one, allowing the contribution of a sensor observation falling across multiple cells to be properly accounted for in each cell.

6 The products: scientific overview

The following sections provide a comparison with previous versions and an overview of the variables in the OC-CCI products. All information on the structure of the product files regarding dimensions, flags, or metadata is described on page 39.

The screenshots provided in these sections all follow the same pattern: the actual screenshot is always supported by a colour bar. A logarithmic scale is used for chlorophyll-a.

6.1 Comparison of OC-CCI v6.0 and v5.0

The OC-CCI data and each of its subsets comprise two parts, the direct product and the uncertainties associated with the product. We can use the in-situ database created in the project to compare the performance of the OC-CCI products between versions. An example of such a comparison is shown in Figure 6.

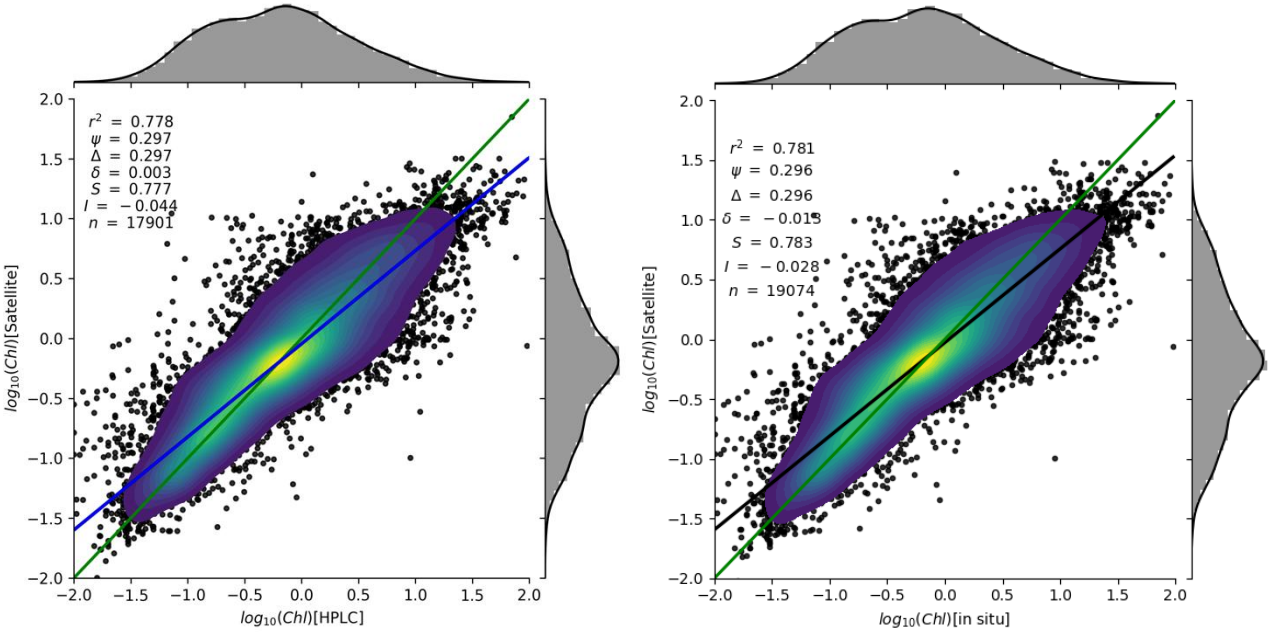


Figure 6: Comparison of v5.0 (left) and v6.0 (right) of the OC-CCI chl product when matched against the latest CCI database of in-situ chl-a measurements. Summary statistics shown are (from top to bottom) correlation coefficient, root-mean-square-difference (RMSD), un-biased RMSD, bias, slope of regression, intercept of regression and number of matchups. Green lines are 1:1 and the black/blue lines are the log-log linear regression.

It should be noted that the change in the performance statistics for a product, such as chlorophyll, between versions is a combination of factors (changes to the in-situ database, atmospheric processing algorithms, inter-sensor de-biasing, etc). Overall, we can see that the v6.0 chlorophyll product has a slightly larger absolute bias but reduced RMSE and improved correlation coefficient, slope and intercept compared to v5.0.

To compare the uncertainty variables between differing OC-CCI datasets one can 1) look at the large scale (global) range and distribution of uncertainty values as shown in Figure 7 and Figure 8, 2) look at the per waterclass uncertainty metrics as shown in Figure 9. Figure 7 shows a sign change in the $R_{rs}(490)$ bias from v5.0 to v6.0, a small negative to a small positive value. This slight positive shift is seen in most waterclasses

and wavelengths. The R_{rs} RMSD has been reduced across almost all waterclasses and wavelengths (Figure 9) and is most notable in coastal and high latitude regions.

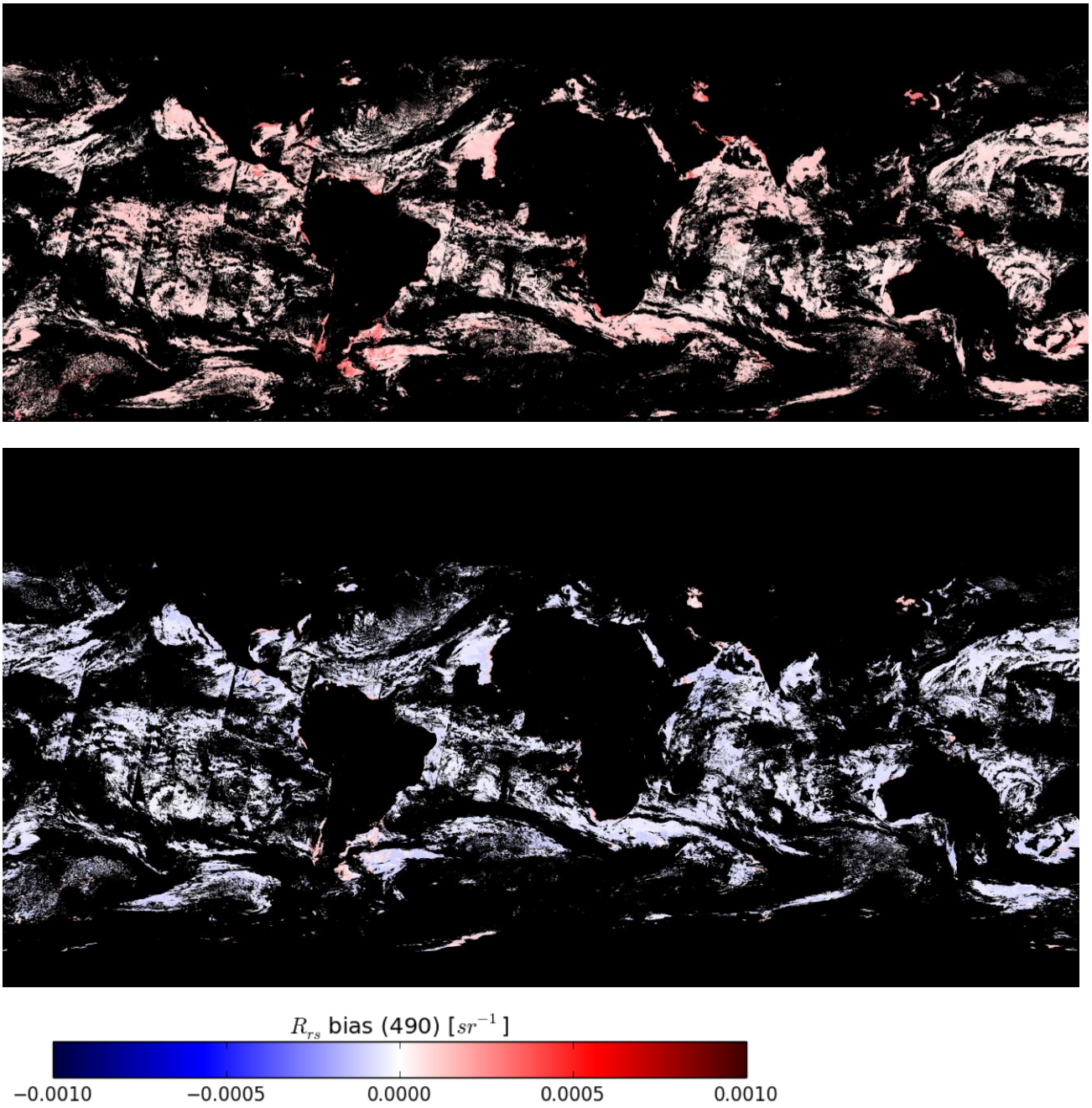


Figure 7: Comparison of the Rrs 490 bias uncertainty as given in v6.0 (top) and v5.0 (bottom).

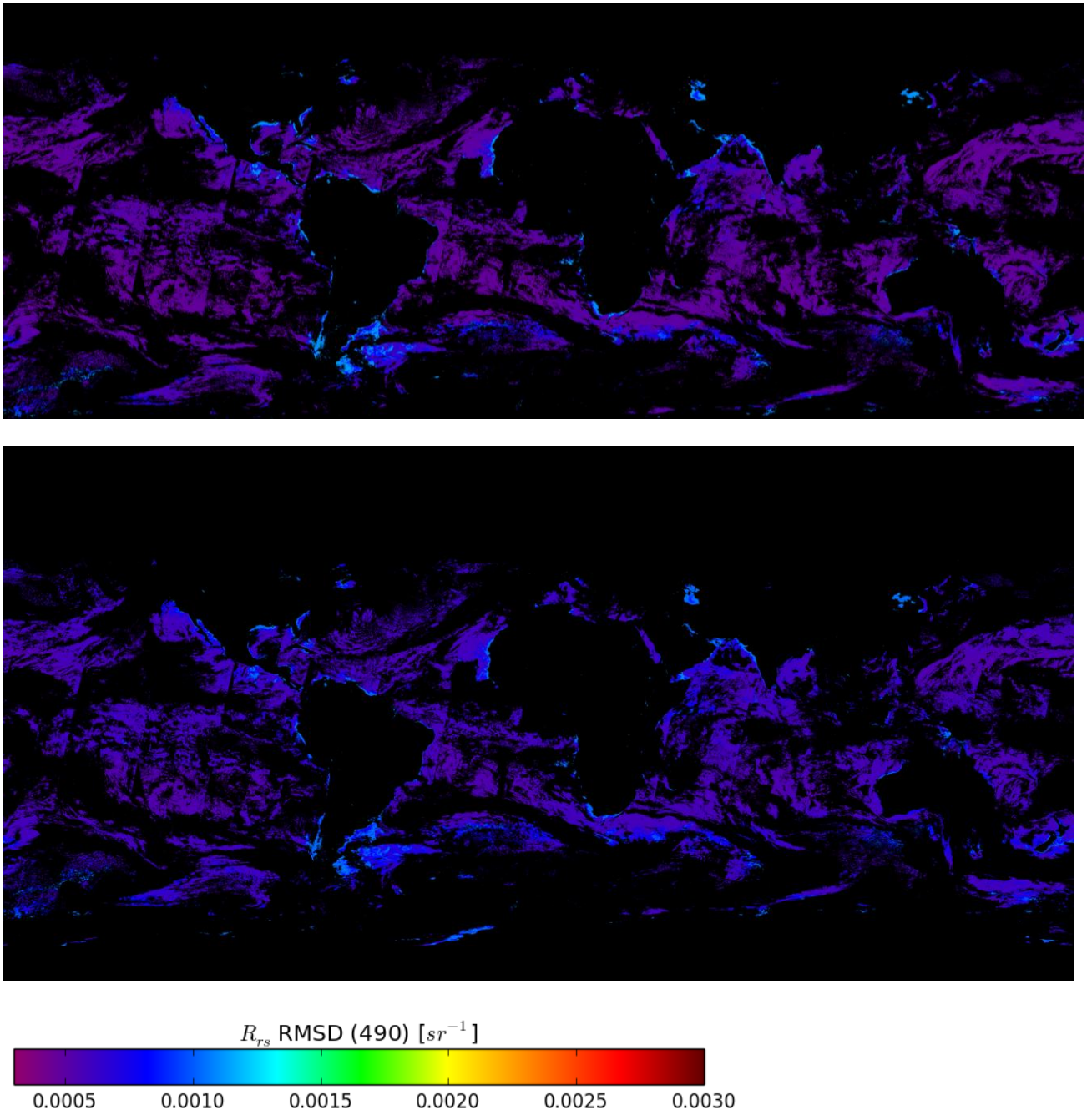


Figure 8: Comparison of the Rrs 490 rmsd uncertainty as given in v6.0 (top) and v5.0 (bottom).

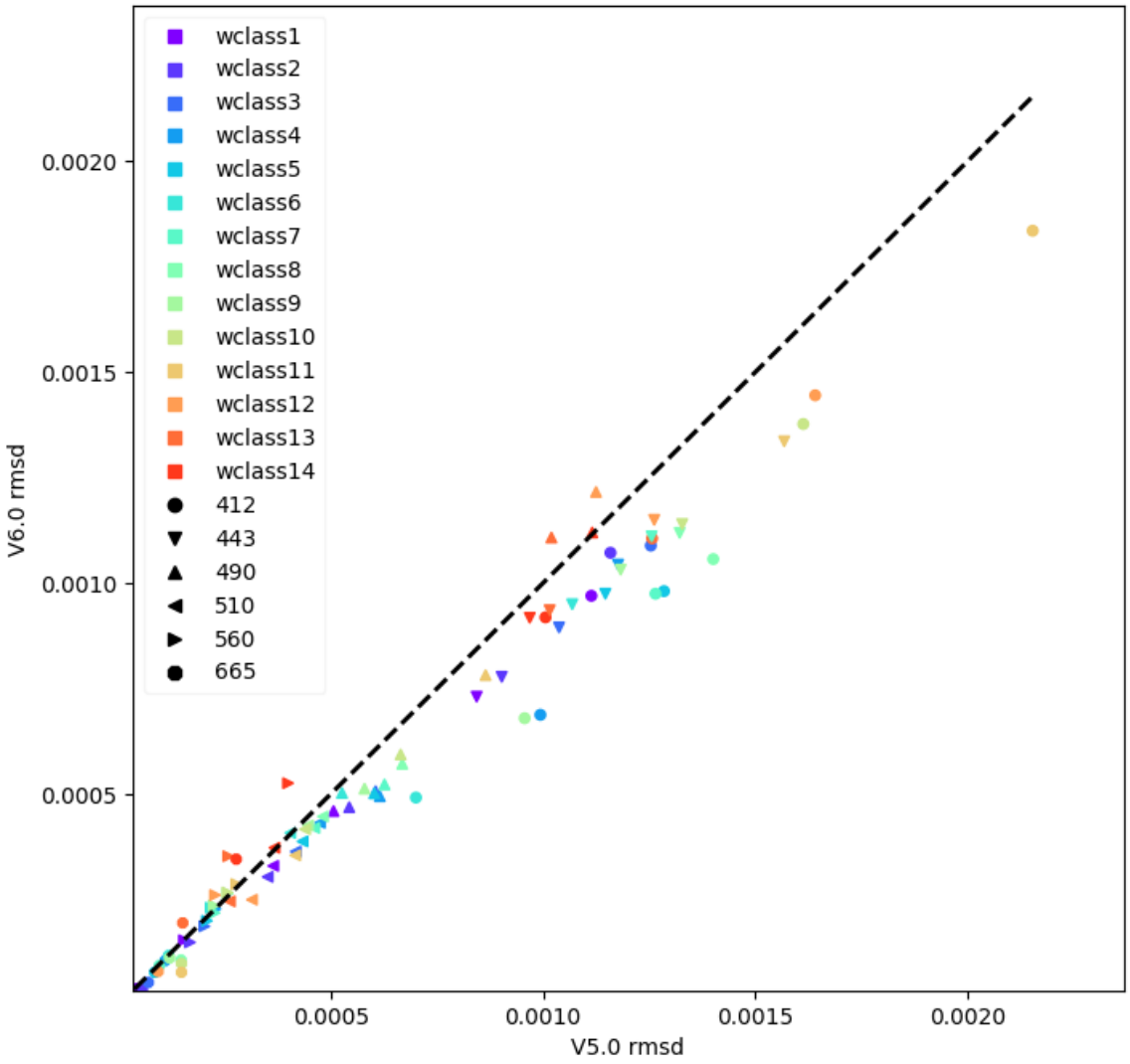


Figure 9: Comparison of the per-waterclass root mean square difference between V5.0 and V6.0 for all R_{rs} wavebands. For most waterclasses and wavebands the rmsd is reduced (below the 1:1 dashed line).

Overall, the OC-CCI v6.0 data has spatial coverage of data and product performance comparable with that of v5.0. The performance of the products in the most recent years of the product is better in v6.0 than 5.0 due to the addition of the S3B OLCI data stream and removal of the aged MODIS and VIIRS data streams.

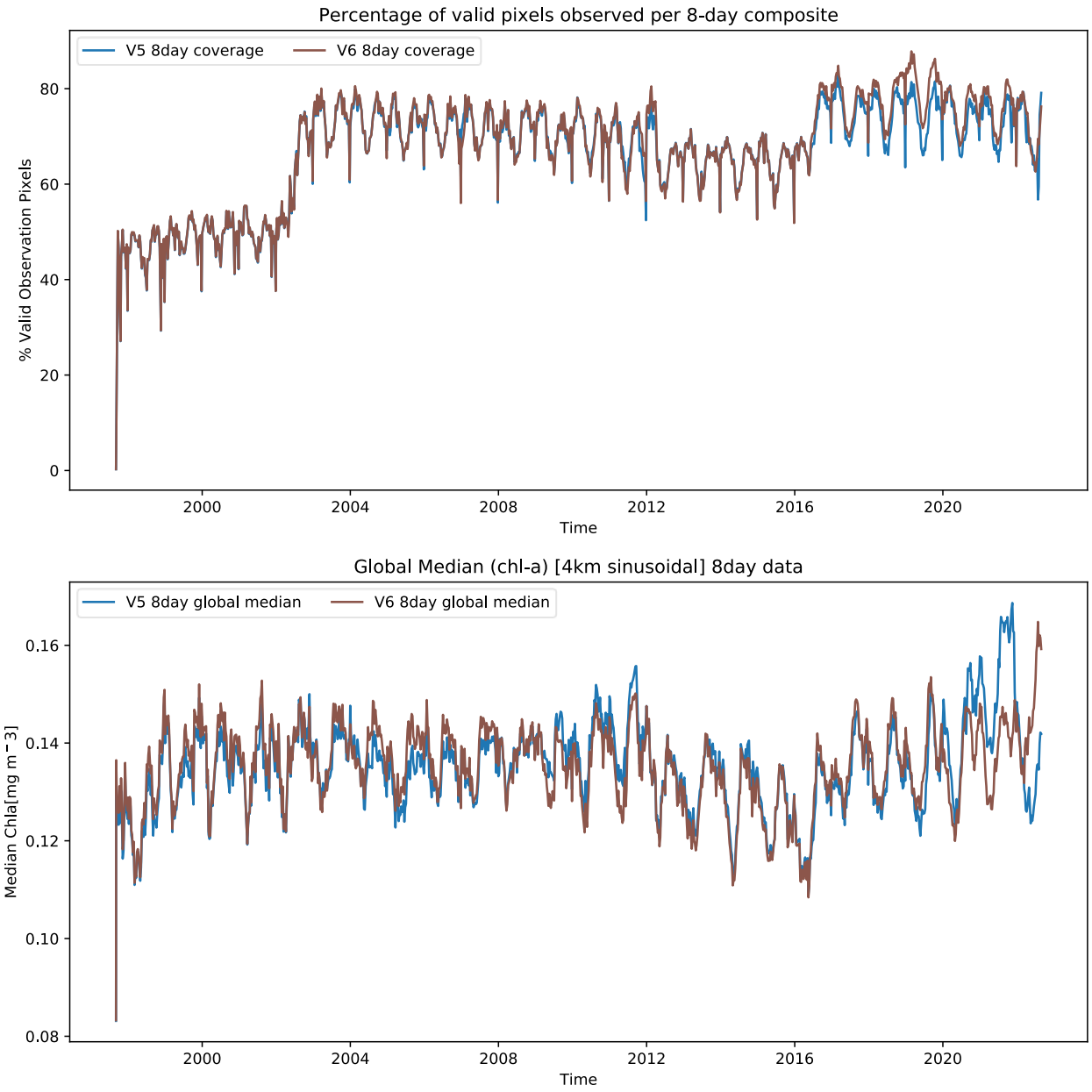


Figure 10: Comparison of the coverage and global median chlorophyll a concentration through time for v5 and v6 using the monthly and 8-day composites as input.

Comparing the v6.0 and v5.0 datasets using the global median chlorophyll-a and ocean coverage (Figure 10) we can see that the median chlorophyll-a appears largely consistent between versions 5 and 6 (showing the same seasonal cycle and similar values for most of the record) but differs in the most recent few years. This is due to the differing sensor inputs for 2020 onwards and we would consider the v5 less reliable in this period. The coverage of the v6.0 is very slightly increased compared to v5.0 in during the SeaWiFS years due to an updated masking scheme used in the v6.0 production. The enhanced coverage during MERIS and MODIS years is likely due to updates to POLYMER and IDEPIX. Additional coverage from OLCI 3B is roughly balanced by the loss of MODIS and VIIRS post 2019 but OLCI A and B combined give some of the best coverage in the record.

6.2 Product overview

6.2.1 Chlorophyll-a concentration (mg m^{-3})

The chlorophyll-a concentration (chl-a) is recognized as an Essential Climate Variable, and was identified as a key variable in the CCI-user survey, required by both modellers and EO scientists (see [AD 1]). Chlorophyll-a in the OC-CCI products has units of mg m^{-3} and is provided as daily products with a horizontal resolution of ~4 km/pixel. Furthermore, the root-mean-square (RMS) uncertainty and the bias in the log10 chlorophyll-a concentration are provided, based on comparison with match-up in-situ data. The chlorophyll-a values are calculated by blending algorithms based on the water-type as documented in the ATBD document. For v6.0, this involved the blending of the OCI algorithm (as implemented by NASA, itself a combination of CI and OC4), the OCI2 algorithm (an updated OCI parameterisation), the OC2 algorithm and the OCx algorithm. Each algorithm utilises the same OC-CCI merged R_{rs} products described below.

Figures 10-11 show, respectively, example of daily chl-a product, and the corresponding RMS uncertainty and bias.

Please note that while the chlorophyll values are provided in normal units, the uncertainty is based on log10 values due to the underlying natural distribution.

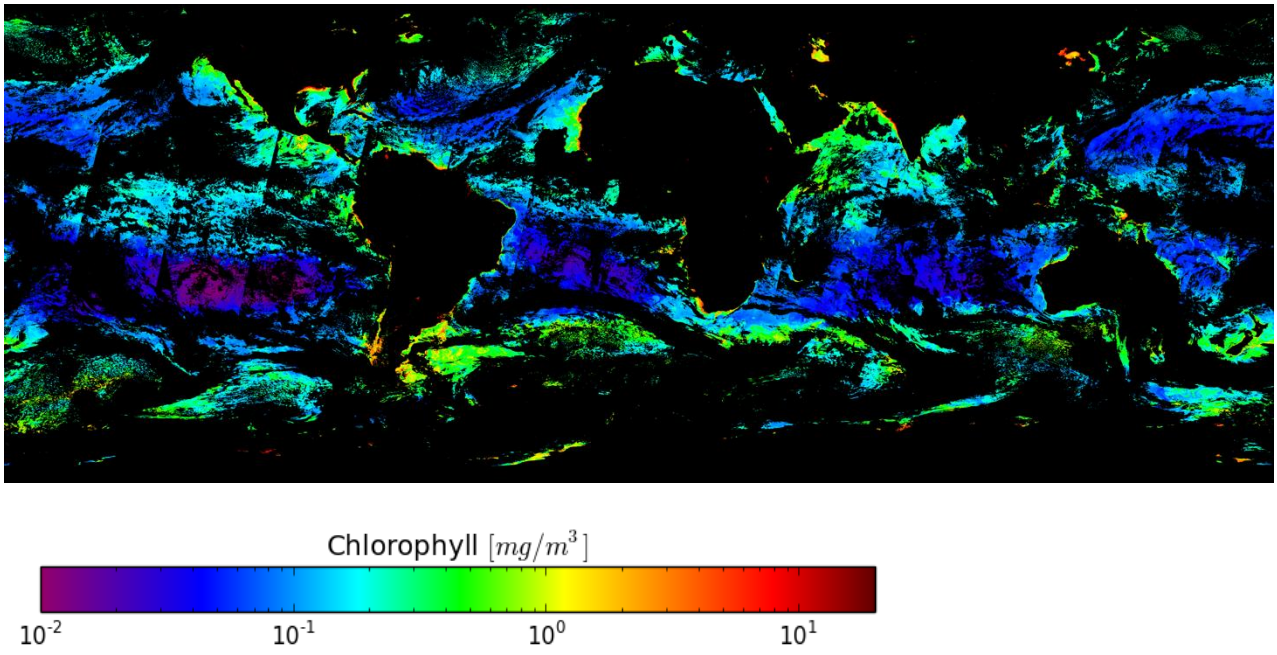


Figure 11: Chlorophyll-a concentration (1st Jan 2003)

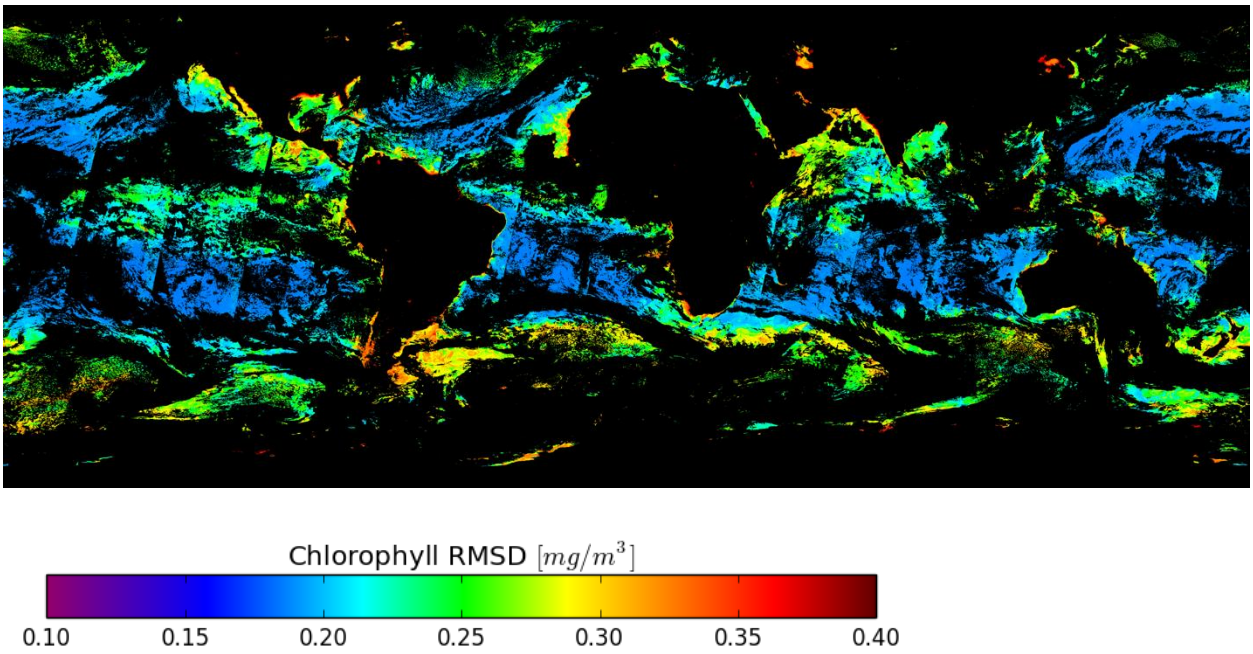


Figure 12: Root mean square difference of chlorophyll-a concentration (1st Jan 2003)

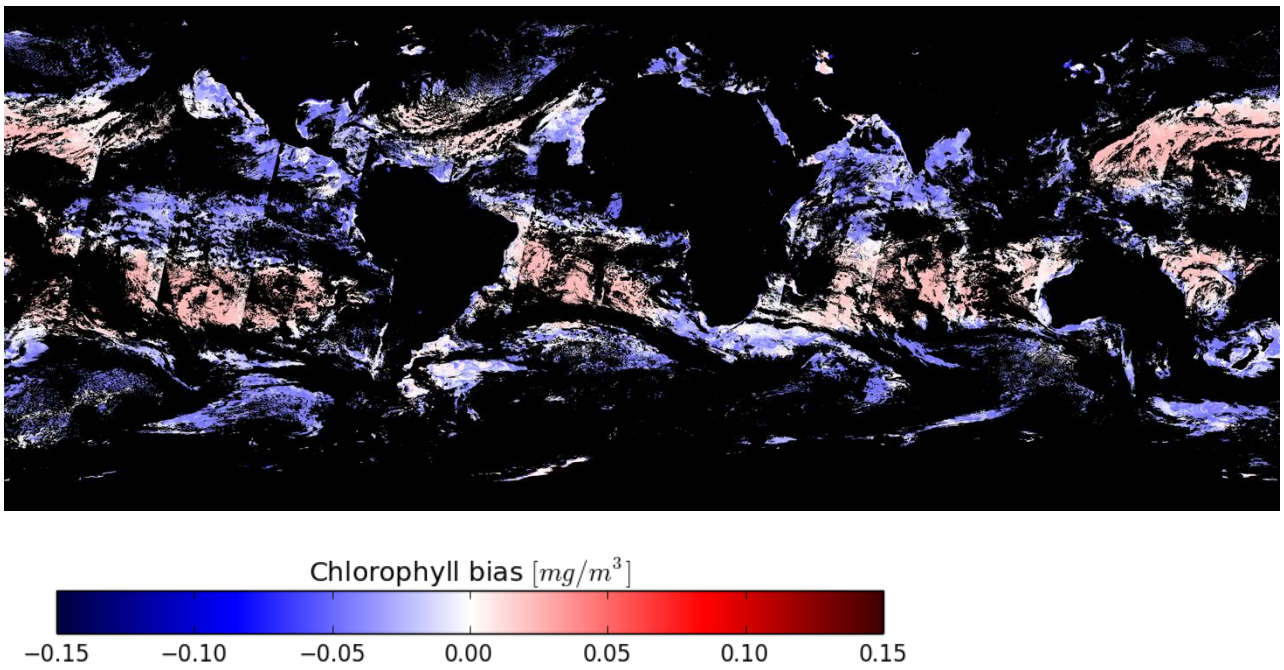


Figure 13: Bias of chlorophyll-a concentration (1st Jan 2003)

6.2.2 Remote Sensing Reflectance (sr^{-1})

The OC-CCI products also include daily composites of remote-sensing reflectance (R_{rs}) at the sea surface, at a resolution of ~4 km/pixel. R_{rs} values are provided for the standard MERIS wavelengths (412, 443, 490, 510,

560, 665nm) with pixel-by-pixel uncertainty estimates for each wavelength. These are merged products based on SeaWiFS, MERIS, Aqua-MODIS, VIIRS and OLCI data. Atmospheric correction was carried out using the POLYMER algorithm for MERIS & MODIS (see the POLYMER Algorithm Theoretical Baseline Document) and SeaDAS v7.3 processor for SeaWiFS and VIIRS. The R_{rs} values from SeaWiFS, MODIS and VIIRS were band-shifted to MERIS wavebands if necessary, and SeaWiFS and MODIS were corrected for inter-sensor bias when compared with MERIS in the 2003-2007 period. VIIRS and OLCI were also corrected to MERIS levels, via a two-stage process comparing against the MODIS-corrected-to-MERIS-levels (2012-2013 for VIIRS and 2016-2019 for OLCI).

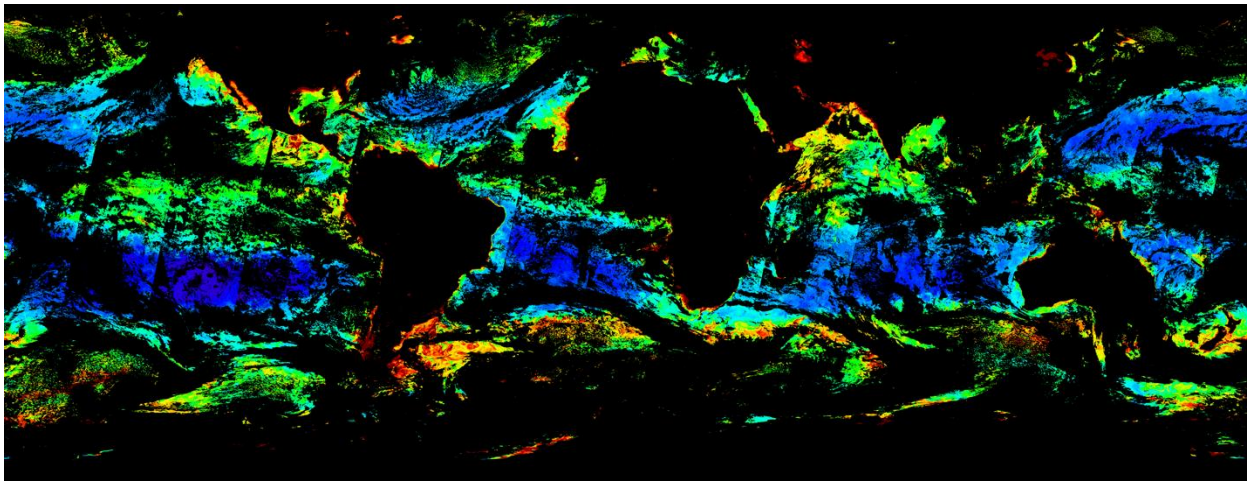
6.2.3 Kd490: the attenuation coefficient for downwelling irradiance (m^{-1})

The attenuation coefficient at 490nm for downwelling irradiance, which is an apparent optical property, is one of the OC-CCI products. It is provided at daily resolution and spatial resolution of ~4 km/pixel. It is computed from the inherent optical properties (see below) at 490 nm and the sun-zenith angle, using the Lee et al. (2005) algorithm.

6.2.4 Inherent Optical Properties (IOP): total absorption and backscattering coefficients and their components (a_{tot}, a_{ph}, a_{dg}, b_{bp}) (m^{-1})

The OC-CCI product includes inherent optical properties (IOP): the total absorption and particle backscattering coefficients, and, additionally, the fraction of detrital & dissolved organic matter absorption (a_{dg}) and phytoplankton absorption (a_{ph}). The *total absorption* (units m^{-1}), the *total backscattering* (m^{1}), the *absorption by detrital and coloured dissolved organic matter* a_{dg} (m^{-1}), the *backscattering by particulate matter* (m^{1}), and the *absorption by phytoplankton,* a_{ph} (m^{1}) share the same resolution of ~4 km. The values of IOP are reported for the standard MERIS wavelengths (412, 443, 490, 510, 560, 665nm). They were computed from daily, merged R_{rs} values using the Lee et al. (2009) algorithm updated to the QAA_v6. Note that total absorption coefficient is the sum of absorption coefficients of pure water (a_w) according to Pope and Fry (1997), a_{ph} and a_{dg} i.e. $a_{tot} = a_w + a_{ph} + a_{dg}$ for each wavelength. The backscattering coefficient reported is particle backscattering (b_{bp}), and does not include the contribution to total backscattering from water. Uncertainty estimates (RMSD and bias) are reported for the components of absorption (a_{ph} and a_{dg}) but not for a_{tot} or b_{bp}.

Figures 11 to 13 show global daily images of total absorption, absorption of detrital and dissolved matter, and absorption by phytoplankton at 443 nm.



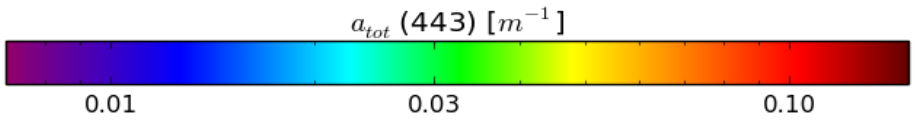


Figure 14: Total absorption at 443 nm (1st Jan 2003)

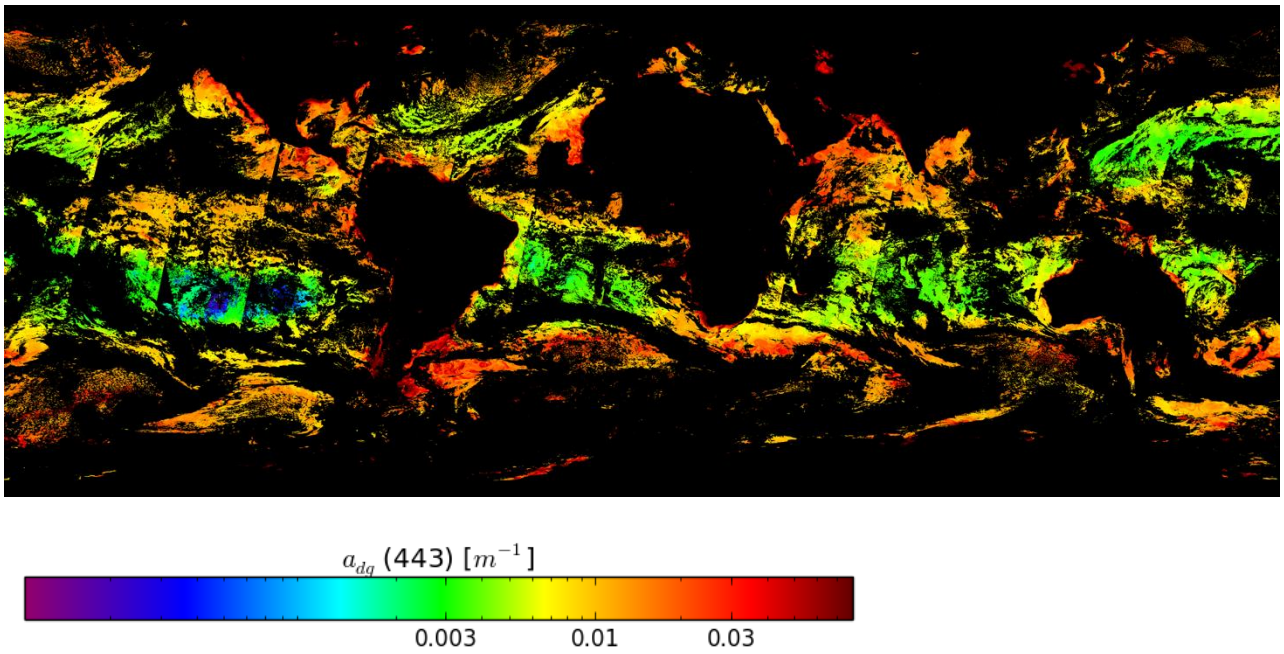
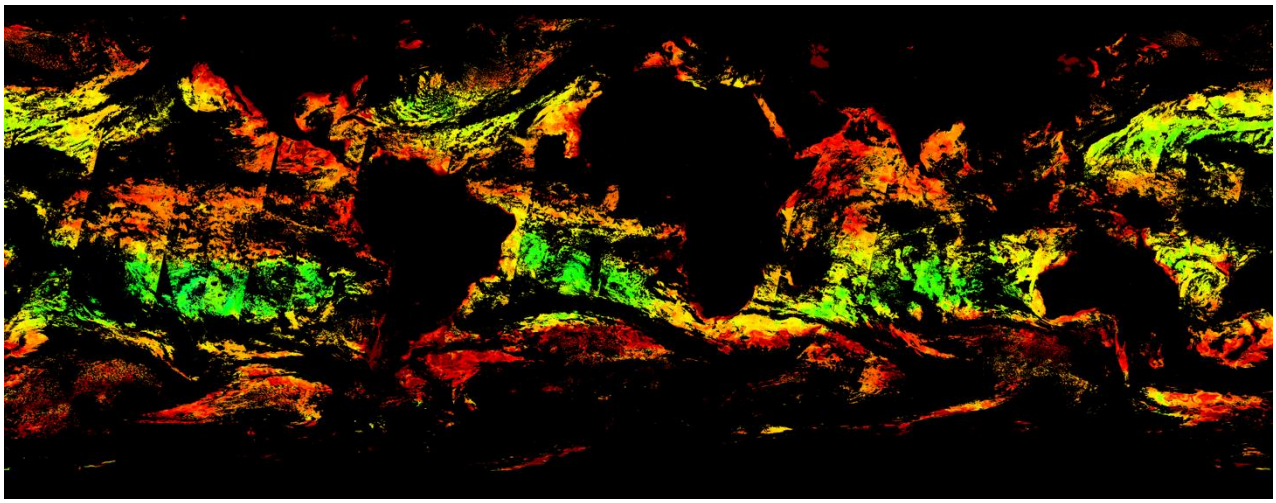


Figure 15: Absorption by detrital and dissolved matter at 443 nm (1st Jan 2003)



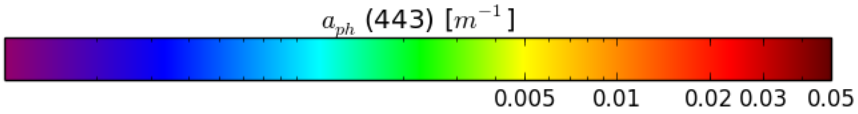


Figure 16: Phytoplankton absorption at 443 nm (1st Jan 2003)

6.2.5 Uncertainty characterisation

Each product has pixel-by-pixel uncertainty characterisation (root-mean square difference and bias), with the exception of b_b where insufficient supporting in-situ data were available to make a viable estimate of uncertainty, and for a_{tot}, which is a convenience product based on the other absorption components, all of which have associated uncertainty. These uncertainties are based on comparison of satellite products with in-situ match-up data. To extrapolate from point observations to global scales, uncertainties are first computed for different optical water types in the ocean. The membership of the various optical water types is determined for each pixel: that is, each pixel can exhibit the characteristics of more than one class. The uncertainties are then calculated for each pixel as the weighted sum of the uncertainties for each water class, according to the pixel water class membership. The approach follows the work of Moore et al. (2009) and Jackson et al. (2017).

Note that the uncertainty for chlorophyll is based on the log10 chlorophyll values, because the underlying natural distribution is logarithmic.

6.2.6 Optical water classes

The uncertainty estimates for each pixel and product are computed based on a classification of the optical water type using fuzzy logic, following Moore et al (2009). In CCI v1.0, Moore's eight water classes based on SeaWiFS were used; in v2.0 (and used unchanged in v3.0, v3.1 and v4.0), 14 specific classes were derived that best match the observations. This process is described in Jackson et al. (2017, DOI 10.1016/j.rse.2017.03.036). For v5.0 a new optical water class set was required, as the data used MERIS as the reference. These classes were also used for v6.0 following a check that they were still suitable. The water class set is based on reflectances following a normalisation of each band to the total spectral integral across the 6 optical bands. This focusses the emphasis on the spectral shape when differentiating classes. Figure 16 shows the spectral shapes of the final classes used in v6.0 processing:

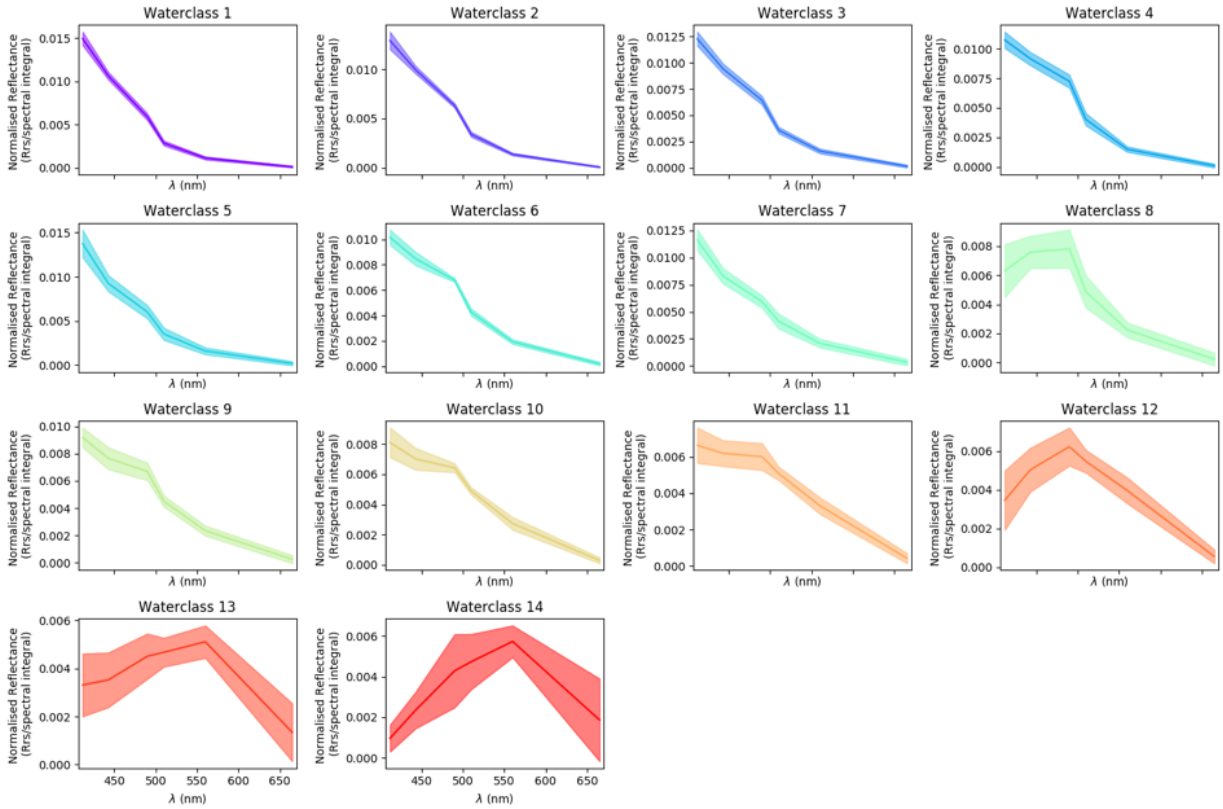


Figure 17: Normalised spectral response of the water types used in OC-CCI v6.0 (solid lines are class means and shaded region shows standard deviation).

6.2.7 The data-day approach

A new spatial and temporal definition of a data-day has been used for the production of these products. This approach has been adapted from the findings of the GlobColour project:

"The aim of the data-day definition is to avoid mixing pixels observed at two different times. As for other classic definitions, we accept to increase the duration of a day in order to include the previous and next day data. Then, at the same spatial area we could select the best input, i.e. the one leading to the lowest temporal discrepancies. A data-day therefore may represent data taken over a 24 to 28 hour period.", GlobColour Product User Guide, http://globcolour.info

As the satellites carrying SeaWiFS, MODIS and MERIS satellites have different orbits, each has its own data-day definition.

To achieve this separation, the following simple algorithm was adopted to distinguish between three different data-days:

```
if ( h < CNT + ( φ +180)* τ ) then
    pixel is attached to data-day (d-1)
else if ( h > CNT + ( φ +180)* τ + 24) then
    pixel is attached to data-day (d+1)
else
```

```
    pixel is attached to data-day (d)
end if
```

Where the variables have the following meaning:

- CNT (in hours): crossing nodal time in ascending track
- τ (hr/°): slope of the data-day definition lines
- d (UTC date): UTC date (day) of the measured pixel
- h (UTC hour): UTC date (four) of the measured pixel
- φ (deg): longitude of the measured pixel

Note: τ has a constant value equal to -24/360.

The crossing nodal time (CNT) is a constant depending on the satellite:

- MODIS (Aqua): 13.5
- SeaWiFS (Orbview-2): 12.0
- MERIS (ENVISAT): 10.0
- VIIRS: 13.5
- OLCI(3A and 3B): 10.0

7 The products: technical overview

This section provides an in-depth description of the format of the OC-CCI data products.

7.1 General format description

The outputs of the OC-CCI processing chain are level 3 mapped daily composites, generated from multiple sensors, with a spatial resolution of 4 km/pixel. The data are stored as CF-compliant NetCDF as has been mandated by the ESA CCI Data Standards Working Group. NetCDF version 4 is used because it allows for transparent internal compression of the data, which would otherwise be approximately 15 times larger using NetCDF 3; hence, users need to ensure that their NetCDF libraries are at least version 4.0.0 (released 2008) or higher to be able to read these files.

Familiarity with NetCDF terminology and general usage is assumed for this section.

For the v6.0 data release, a typical netCDF file containing the full set of products for a single day is approximately 1.3GB. Subsetted versions of these files containing only related product groups (e.g. chlorophyll, R_{rs}, IOPs, etc) and advanced data services (e.g. OPeNDAP) are available to mitigate download size problems.

7.2 Filename convention

The name convention for OC-CCI processed products follows the second form required in [AD4]. The filename convention is:

```
ESACCI-OC-<Processing Level>-<Product String>-<Data Type>-<Additional
Segregator>-<Indicative Date>[<Indicative Time>]-fv<File version>.nc
```

With the components above being:

<Processing Level>	see [AD-4]; for the OC-CCI processed products, 'L3S' will apply.
<Product String>	The Product String defines the source of the data set and depends on the processing level. For the OC-CCI processed products, 'MERGED' will apply
<Data Type>	This should contain a short term describing the main data type in the data set.
<Additional Segregator>	This is an optional part of the filename, containing information about spatial and temporal resolution, length of time period, processing centre etc.
<Indicative Date>	The identifying date for this data set. Format is YYYY[MM[DD]].
<Indicative Time>	The identifying time for this data set in UTC. Format is [HH[MM[SS]]].
<File version>	Dataset version for GHRSST compatibility; always the same as the CCI dataset version, e.g. “6.0” for the v6.0 data

7.3 Example filename

An example filename is:

```
ESACCI-OC-L3S-OC_PRODUCTS-MERGED-1D_DAILY_4 km_GEO_PML_OCx_QAA-20031225-fv6.0.nc
```

With components being:

Filename component and alternates	**Description**
ESACCI-OC	Fixed prefix
L3S	Processing Level (fixed)
OC_PRODUCTS	Data Type string indicating all products in one file
CHLOR_A	chlorophyll-related product subset
RRS	R_{rs} and water class product subset
IOP	IOP product subset
K_490	Kd490 product subset
MERGED	Data is from more than one sensor (fixed, though may be used in future releases of individual sensors)

1D	Additional Segregator Element: Composite data (1 day, may be other variants here)
DAILY	Additional Segregator Element: Length of time period covered
4 km	Additional Segregator Element: Spatial Resolution
GEO *SIN*	Additional Segregator Element: Projection type (Geographic or Sinusoidal)
PML	Additional Segregator Element: Processing Centre (fixed)
OCx_QAA	Additional Segregator Element: Algorithm(s) (varies)
20030907	Indicative Date

7.4 Grid format, map projection and coverage

The products are available in two projections: sinusoidal and geographic (also known as equidistant cylindrical, equirectangular, Plate Carrée, etc).

Sinusoidal projection better preserves the area covered by a data cell, especially at the poles.

Geographic projection is simplest to use as a simple rectangular array but misrepresents the area at the poles unless this is specifically accounted for.

All files contain CF-compliant latitude and longitude (and time) dimensions, allowing each data cell to be specifically associated with a location. All latitudes and longitudes are given in WGS/84 datum.

7.5 Geographic grid format

The most commonly used projection, geographic, is a direct conversion of latitude and longitude coordinates to a rectangular grid, typically a fixed multiplier of 360x180. The OC-CCI “GEO” NetCDFs follow the CF convention for this projection with a resolution of 8640x4320.

7.6 Binned grid format

The primary projection used in the OC-CCI processing chain is a global, sinusoidal equal-area grid (see Fig. 11), matching the NASA standard level 3 binned projection [RD 3]. The default number of latitude rows is 4320, which results in a vertical bin cell size of approximately 4 km. The number of longitude columns varies according to the latitude, which permits the equal area property. Unlike the NASA format, where the bin cells that do not contain any data are omitted, the CCI format retains all cells and simply marks empty cells with a NetCDF fill value. The compression built into NetCDF version 4 achieves nearly the same space efficiency as that possible with NASA’s omission of these cells while making the CCI product significantly easier to use.

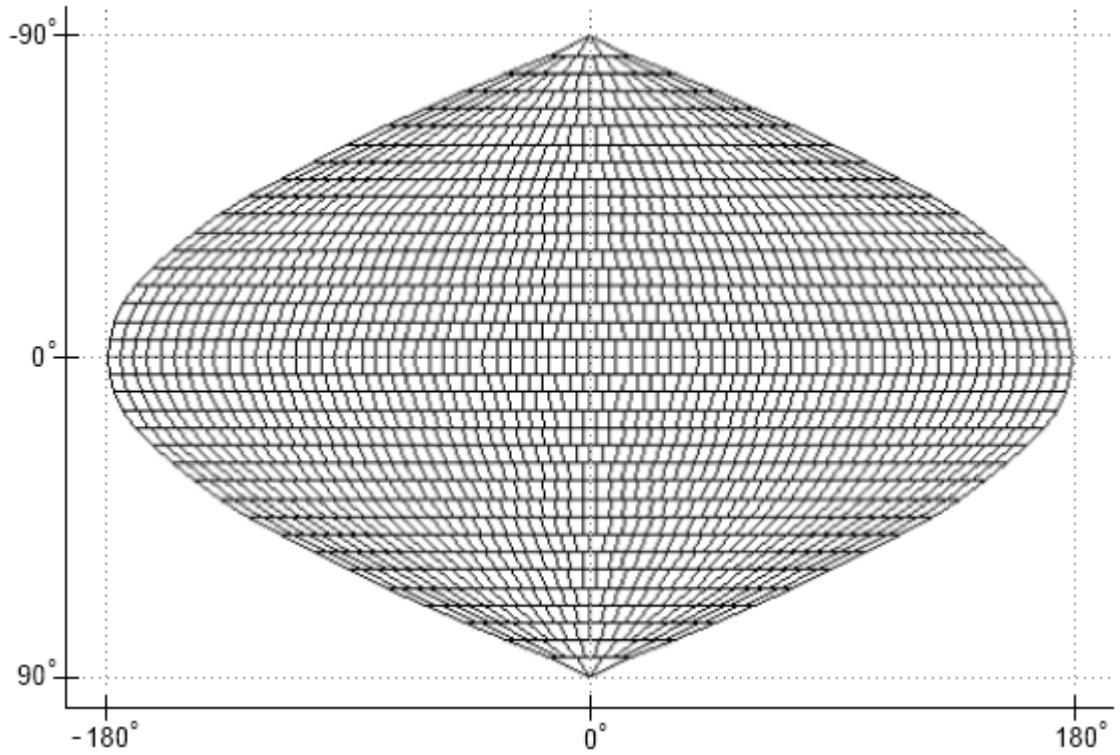


Figure 18: The sinusoidal grid

When written into a NetCDF file, this grid is flattened i.e. the data are stored in one-dimensional variables, where the one dimension of all variables is the total number of bin cells (approximately 23 million). Each NetCDF file contains auxiliary information describing the grid. In the NASA format, the geo-coordinates of every cell must be manually computed. The CCI product instead includes per-pixel latitude and longitude variables for greater ease of use and to meet CF-compliance requirements.

7.7 File structure

This section provides an overview of all the dimensions and variables contained in the OC-CCI processed products. Since the data are provided on two different grids, there are two subsections describing the specific parts of these, while the majority of the variables are covered in one section below.

7.8 Specific elements of the sinusoidal products

```
dimensions:
           time = 1 ;
           bin_index = 23761676 ;
variables:
  int crs ;
            crs:grid_mapping_name = "1D binned sinusoidal" ;
            crs:number_of_latitude_rows = 4320 ;
            crs:total_number_of_bins = 23761676 ;
  float Rrs_412(time, bin_index) ;
            Rrs_412:grid_mapping = "crs" ;
           float lon(bin_index) ;
            lon:standard_name = "longitude" ;
            lon:units = "degrees_east" ;
```

```
            lon:axis = "X" ;
           float lat(bin_index) ;
            lat:standard_name = "latitude" ;
            lat:units = "degrees_north" ;
            lat:axis = "Y" ;
```

The sinusoidal projection has a primary dimension of bin_index, which is used by the data variables. Standard latitude and longitude variables exist and are indexed with the same dimension to provide world coordinates, via the standard “coordinates” attribute linking the data variables to the coordinate variables, per the CF convention. Time is included as a dimension, though is of length 1 for all products.

The ‘crs’ variable is a CF style grid mapping variable that describes and parameterises the sinusoidal projection and can be used as a definitive way to identify a sinusoidally projected variable. The contents of this variable are not yet accepted into the CF convention, but follow the guidelines laid out for new projections.

7.9 Specific elements of the geographic products

```
dimensions:
        time = 1 ;
        lat = 4320 ;
        lon = 8640 ;
variables:
  int crs ;
    crs:grid_mapping_name = "latitude_longitude" ;
  float chlor_a(time, lat, lon) ;
    chlor_a:grid_mapping = "crs" ;
```

The geographic project files are completely CF standard in terms of their projection descriptors. The ‘crs’ variable contains the standard element for a lat/long projection and all variables are dimensioned directly with time, latitude and longitude.

7.10 Product dimensions

The final products' dimensions referenced in the following are:

- **lat**, which determines the latitudinal position. This is indirectly referenced via the “bin_index” dimension in the sinusoidal projection.
- **lon**, which determines the longitudinal position. This is indirectly referenced via the “bin_index” dimension in the sinusoidal projection.
- **time**, which determines the point in time. For all released products, this is a dimension with a length of 1. It is included both for standardisation purposes and to simplify “stacking” of multiple files into a single data cube.

7.11 Flags

As the products are a composite both over time (one day) and of multiple sensors, it is not possible to preserve flags from the source datasets. This is in common with most level 3 compositing approaches. Instead, appropriate filtering was done prior to the level 3 step to exclude pixels flagged as “bad” (details in the SPS).

7.12 Geophysical variables

NetCDF is a self-documenting format, meaning that the majority of the information needed to correctly use and interpret the data are incorporated into the file metadata. Accordingly, this section does not summarise all of the attributes of every variable, but shows one common example from the sinusoidal projection (geographic projection is the same apart from having latitude and longitude dimensions instead of a bin_index that is used to look these up):

```
float chlor_a(time, lat, lon) ;
        chlor_a:long_name = "Chlorophyll-a concentration in seawater (not log-
transformed), generated by SeaDAS using a blended combination of OCI (OC4v6 +
Hu\'s CI), OC3 and OC5, depending on water class memberships" ;
        chlor_a:units = "milligram m-3" ;
        chlor_a:_FillValue = 9.96921e+36f ;
        chlor_a:ancillary_variables = "chlor_a_log10_rmsd chlor_a_log10_bias" ;
        chlor_a:grid_mapping = "crs" ;
        chlor_a:parameter_vocab_uri =
"http://vocab.ndg.nerc.ac.uk/term/P011/current/CHLTVOLU" ;
        chlor_a:standard_name =
"mass_concentration_of_chlorophyll_a_in_sea_water" ;
        chlor_a:units_nonstandard = "mg m^-3" ;
```

The listing above shows the chlor_a data variable, which, in common with all the others, is of the float32 datatype with some data values missing (represented by the NetCDF standard float32 fill value). The “standard_name” attribute gives the accepted name for the parameter described (see the CF convention standard name table) and is used to allow automatic interpretation of physical values. The parameter_vocab_uri serves the same purpose but using the British Oceanographic Data Centre (BODC) vocabulary services namespace. The long_name provides a human-readable descriptive complement to these. Units are described in udunits compatible format and a “nonstandard” variant interpretable by some other programming libraries. The ancillary_variables attribute indicates this variable is linked to the two other named ones (in this case, they represent the uncertainty parameters for this variable). Finally, the grid_mapping and coordinates attributes indicate which other variables within the netCDF contain information on the projection and which are the axis coordinates respectively.

The data-bearing variables are:

Data variable	**Accompanying uncertainty variables**	**Notes**
Rrs_412 Rrs_443 Rrs_490 Rrs_510 Rrs_560 Rrs_665	Rrs_412_rmsd Rrs_443_rmsd Rrs_490_rmsd Rrs_510_rmsd Rrs_560_rmsd Rrs_665_rmsd Rrs_412_bias Rrs_443_bias Rrs_490_bias	Remote sensing reflectance at MERIS wavelengths

	Rrs_510_bias Rrs_560_bias Rrs_665_bias	
chlor_a	chlor_a_log10_rmsd chlor_a_log10_bias	Chlorophyll-a concentration in seawater (not log-transformed), generated using a blended combination of OCI, OCI2, OC2, and OCx algorithms, depending on water class memberships
atot_412 atot_443 atot_490 atot_510 atot_560 atot_665	*Not computed separately, as this is a convenience variable*	QAA total absorption ($a_{ph}+a_{dg}+a_w$, though QAA's decomposition method sometimes does not preserve this property)
aph_412 aph_443 aph_490 aph_510 aph_560 aph_665	aph_412_rmsd aph_443_rmsd aph_490_rmsd aph_510_rmsd aph_560_rmsd aph_665_rmsd aph_412_bias aph_443_bias aph_490_bias aph_510_bias aph_560_bias aph_665_bias	QAA absorption due to phytoplankton
adg_412 adg_443 adg_490 adg_510 adg_560 adg_665	adg_412_rmsd adg_443_rmsd adg_490_rmsd adg_510_rmsd adg_560_rmsd adg_665_rmsd adg_412_bias adg_443_bias adg_490_bias adg_510_bias adg_560_bias adg_665_bias	QAA absorption due to detrital and dissolved matter
bbp_412 bbp_443 bbp_490 bbp_510	*Insufficient in-situ data to make a plausible estimate*	QAA backscatter due to particulate matter

bbp_560 bbp_665		
kd_490	kd_490_rmsd kd_490_bias	Attenuation coefficient (Lee algorithm with Zhang backscatter coefficients)
water_class1 water_class2 water_class3 water_class4 water_class5 water_class6 water_class7 water_class8 water_class9 water_class10 water_class11 water_class12 water_class13 water_class14	*n/a*	Water class memberships according to Moore et al. (2009) and class definitions per the CCI derivations (broadly, classes range from open ocean to coastal waters as the class number increases)

7.13 Data sources (number of observations)

The NetCDFs contain variables indicating how many observations were made within a specific data cell. There are total and per-sensor counts, allowing some flexibility in estimating relative importance of the sensors. It should be noted that the SeaWiFS data used was a mixture of LAC (1km) and GAC (4 km) resolution while the MERIS, MODIS, VIIRS and OLCI data were originally 1km prior to binning. Consequently, the sensor contributions can differ ~16 fold per 4 km pixel and the nobs counts will reflect this. The number of observations are float variables (i.e. decimal) because the binning process allows for a partial coverage of a cell (currently in 1/9ths, due using a super-sampling factor of 9).

The number of observations variables are:

```
float total_nobs(time, bin_index) ;
            total_nobs:long_name = "Count of the total number of observations
contributing to this bin cell" ;
float MODISA_nobs(time, bin_index) ;
            MODISA_nobs:long_name = "Count of the number of observations from the
MODIS sensor contributing to this bin cell" ;
float MERIS_nobs(time, bin_index) ;
            MERIS_nobs:long_name = "Count of the number of observations from the
MERIS sensor contributing to this bin cell" ;
float VIIRS_nobs(time, bin_index) ;
            VIIRS_nobs:long_name = "Count of the number of observations from the
VIIRS sensor contributing to this bin cell" ;
float SeaWiFS_nobs(time, bin_index) ;
            SeaWiFS_nobs:long_name = "Count of the number of observations from
the SeaWiFS sensor contributing to this bin cell" ;
```

7.14 High level metadata

The global attributes listed in Table 4 are common to all OC-CCI processed datasets. The global attributes are based on the CF-convention, the Unidata discovery metadata convention and the CCI guidelines to data producers document. Not all global attributes are listed, but the remainder are either unimportant (included to meet compliance requirements) or obvious (e.g., product_version = "6.0").

Element name	Description
Metadata_Conventions	The conventions to which these global attributes are compliant
standard_name_vocabulary	The source of the standard name table
Title	A short description of the dataset.
license	Licensing policy (open)
tracking_id	A UUID allowing this file to be uniquely referenced back against other information in a database, providing complete provenance on request
keywords	A comma separated list of key words and phrases.
id	The file name
history	An audit trail for modifications to the original data.
naming authority	Identifies a namespace provider
creation_date	Time of file creation
date_created	
creator_name	The data creator's name, URL, and email. The "institution" attribute will be used if the "creator_name" attribute does not exist.
creator_url	
creator_email	
institution	
project	The scientific project that produced the data.
platform	Satellites used for these data
sensor	Sensors used for these data
grid_mapping	Link to a document describing the grid.
time_coverage_start	Describe the temporal coverage of the data as a time range.
time_coverage_end	
time_coverage_duration	
time_coverage_resolution	
processing_level	A textual description of the processing level of the data.
geospatial_lat_min	Describe a simple latitude, longitude, and vertical bounding box.
geospatial_lat_max	
geospatial_lat_resolution	
geospatial_lon_min	
geospatial_lon_max	
geospatial_lon_resolution	

Table 4: The global attributes

8 How were the products made?

A thorough description of the OC-CCI processing chain is given in the Ocean Colour System Prototype Specification document (OC-CCI, 2012). This section briefly recapitulates an overview of the processing chain. Please refer to Figure 19 below.

8.1 Input datasets

The input EO datasets were MERIS Reduced-Resolution (1km) L1b 4th reprocessing (including OCL fixes), MODIS level 1 R2018.0 VIIRS level 1R2018.0 , SeaWiFS level 2 LAC (1km / MLAC) and GAC (4 km) R2018.0 a from NASA and Sentinel 3 A and B OLCI 1km data (baseline 2.xx where xx covers the range of 'current' processing steps up to 2020 as shown at https://www.eumetsat.int/olci-processing-baselines).

8.2 Level 2 processing and binning

MERIS, MODIS, VIIRS and OLCI data were processed with the POLYMER algorithm (v4.14) to level 2. SeaWiFS L2 were downloaded from NASA (implicitly meaning they were processed with l2gen from SeaDAS 7.5).

All individual sensors were binned to level-3 4 km (sinusoidal grid) with the BEAM/SNAP binner. The masking scheme included multiple sources for all sensors. MERIS and OLCI were masked using the POLYMER bitmask and IDEPIX v6.0/7.0 cloud and land flags (6.0 for MERIS and 7.0 for OLCI). VIIRS and MODIS were masked using a combination of IDEPIX v6.0, POLYMER bitmasks and NASA L2 Flags. SeaWiFS was masked using a combination of the standard NASA L2 flags and IDEPIX v6.0.

8.3 Band shifting

The non-reference sensors of SeaWiFS, MODIS, VIIRS and OLCI (though OLCI already contains the MERIS bands so this was not required) were band shifted to the six main MERIS bands (412, 443, 490, 510, 560, 665nm) by computing QAA IOPs and back computing the R_{rs} bands using a high-resolution spectral model. The output R_{rs} for 412-560nm were cleaned of any negative values, with the data removed. Negative R_{rs} values in the 665nm band frequently occurred due to low signal levels, and these were clamped to zero.

Nothing was done to the MERIS data.

8.4 Bias correction

The band shifted SeaWiFS and MODIS R_{rs} were corrected to remove gross differences (biases) against MERIS R_{rs}. The correction was done on a per-pixel basis using a temporally-weighted climatology windowed around the date being corrected, and using 7 day composites as the input in v3.0 onwards (vs 1 day ones in v2.0), such that the corrections take account of seasonal and regional variations. The biases for SeaWIFS and MODIS were computed over the 2003-2007 period as all sensors required were overlapping and functioning well. Bias adjustments were computed at every location where all sensors had gathered data, with a temporal window of +/- 45 days (weighted by the time difference from the centre point) and spatially-limited interpolation (11 pixels) to fill smaller gaps.

VIIRS and OLCI-A are then also corrected to MERIS levels by a similar process, but comparing against MODIS-corrected-to-MERIS-levels rather than directly to MERIS. This indirect comparison is unavoidable due to the lack of temporal overlap and made use of data spanning 2012-2014 for VIIRS and 2016-2019. Similarly, OLCI-B lacks overlaps with MERIS and is corrected to the 'MERIS-corrected' OLCI-A using data from 2018-2021.

8.5 Merging

Following de-biasing, the individual sensor data were merged with a simple average.

8.6 Water class membership

Water classes were computed following Moore et al (2009), but with (14) specific water classes derived specifically for the v5.0 data and used in v6.0. The classes were derived using an iterative approach that identified and added classes until the v5.0 could be classified to a satisfactory level.

8.7 Product generation

A range of products were computed from the merged R_{rs}, directly using the validated algorithms in SeaDAS (with the exception of Kd490, which was independent due to implementation issues in the SeaDAS variant). Algorithms were selected from the best performers in the round-robin evaluation:

- Chlorophyll: blended merge of OCI, OCI2, OC2 and OCx, weighted by the relative levels of membership in specific water classes.
- IOP: QAA (with Zhang bb coefficients)
- KD: Lee variant (with Zhang bb coefficients)
- R_{rs}: Mixed – SeaDAS for SeaWiFS; POLYMER for MERIS, MODIS, VIIRS and OLCI. Bandshifted to MERIS bands and cleaned up.

8.8 Uncertainty estimation

A table of uncertainties for each class were computed from matchups between the CCI in-situ database and the v6.0 data. Every individual pixel in a scene has a computed water class membership percentage for each of classes described above, and a pixel-specific total uncertainty is computed using these memberships to weight the uncertainties per-class from the tables.

8.9 Reprojection

All data are re-projected onto a geographic grid in addition to the basic sinusoidal grid. The reprojection engine is that from BEAM. Both projections have CCI-style metadata added.

8.10 Additional/derived products

For both projections, product subsets are created so that users wanting only a specific subset (e.g., just chlorophyll, IOP or R_{rs} related products) can acquire these with a smaller download.

Composites are created using a mean average of all inputs. At present, monthly and 8-day composites are provided as official products, but 5-day and other cycles may also be available depending on user requests – if they are computed for one user, they will be made available to all.

Lower resolution variants have been created for the Observations for Model Intercomparisons Project (Obs4MIPs) at quarter and half degree resolutions. Even coarser products (e.g., 1 degree) may also be created and distributed on a similar basis.

PNG quicklooks are created for all products. The scaling factors are generally the same as NASA and are the same for the complete timeseries (i.e., they do not vary on a daily or monthly basis). Where NASA has no

equivalent product, a scaling range was chosen that gives good contrast, with the constraints of expressing the full range of values available in the timeseries.

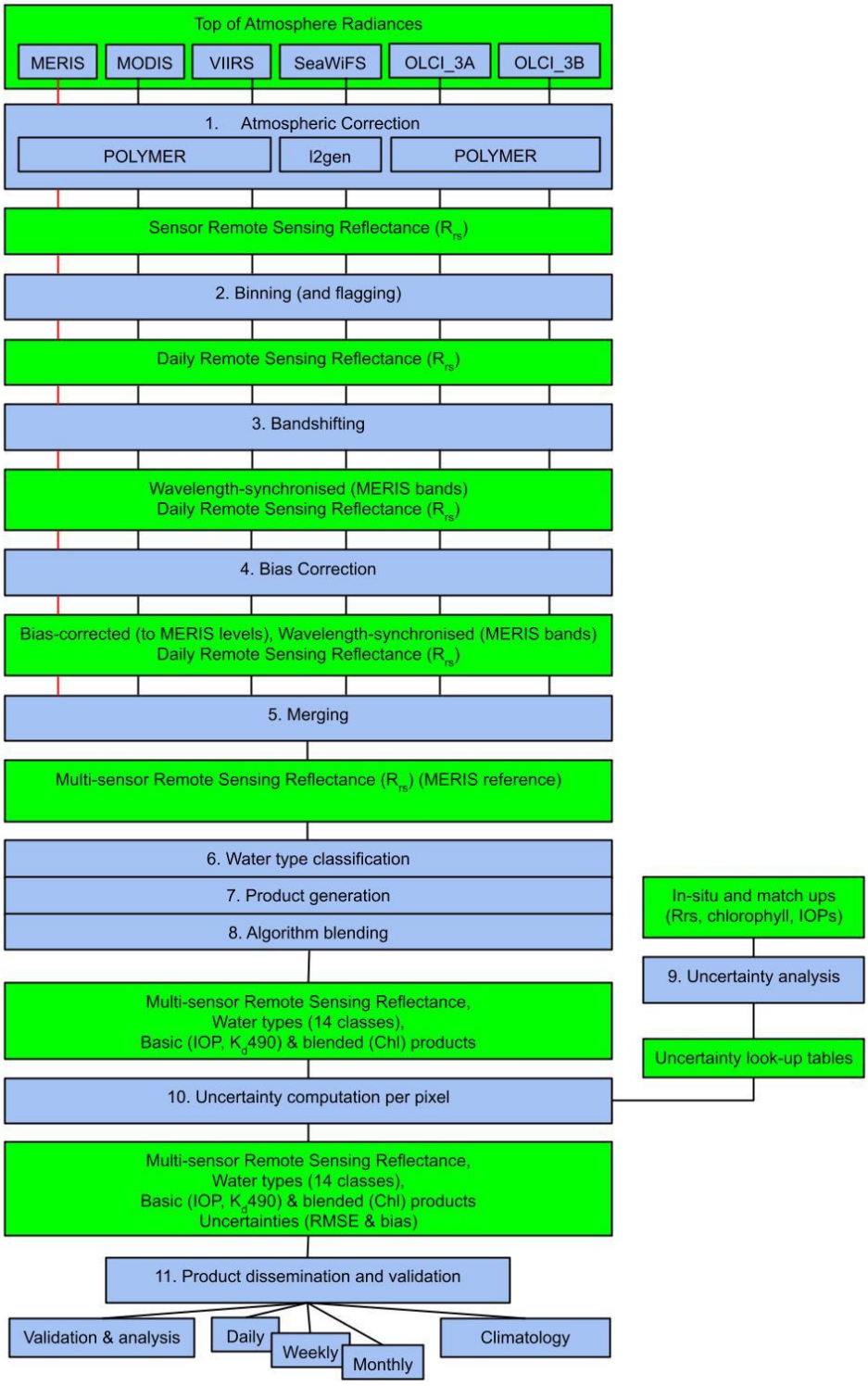


Figure 19: Data flow in the Ocean Colour ECV production

9 References

Jackson T, Sathyendranath S, Mélin F (2017): An improved optical classification scheme for the Ocean Colour Essential Climate Variable and its applications, *Remote Sensing of Environment*: https://doi.org/10.1016/j.rse.2017.03.036

Lee, Z. P., et al. (2005), Diffuse attenuation coefficient of downwelling irradiance: An evaluation of remote sensing methods, *J. Geophys. Res.*, 110(C02017), doi:10.1029/2004JC002573.

Lee, Z., Lubac, B., Werdell, J. and Arnone, R., (2009). *An update of the quasi-analytical algorithm (QAA_v5)*. International Ocean Color Group Software Report, pp.1-9.

Moore, T. S., Campbell, J. W., & Dowell, M. D. (2009). A class-based approach to characterizing and mapping the uncertainty of the MODIS ocean chlorophyll product. *Remote Sensing Environment*, 113, 2424-2430 doi:10.1016/j.rse.2009.07.016 .

OC-CCI (2012) *OC-CCI System Specification document*. Available at: https://docs.pml.space/share/s/6wGtN6XUR_-37HiOXXlZvg , last accessed 22/10/2022

Pope, R. M., & Fry, E. S. (1997, Nov). Absorption spectrum (380--700 nm) of pure water. II. Integrating cavity measurements. *Appl. Opt.*, 36(33), 8710-8723. Retrieved from http://ao.osa.org/abstract.cfm?URI=ao-36-33-8710

10 Earlier versions of OC-CCI dataset

This annex briefly summarises some of the previous OC-CCI data releases to put in context the high level changes. We strongly recommend that the newest data version is used.

V0 (September 2012)

This was the initial test release, consisting of the basic products for 2003 and initial uncertainty estimates. It notably had some excessively high values in higher latitudes.

V0.9 (May 2013) and v0.95 (July 2013)

A first all-years release with many improvements, intended for internal QC and some within-CCI initial evaluations. The majority of the metadata was not present and there were some consistency issues due to the incremental processing used to create the dataset. Some of the high latitude issues present in v0 were corrected by a POLYMER reprocessing and a solar zenith cut off of 70 degrees. A small number of anomalously high and low values made simple evaluations misleading.

V1.0rc1 (November 2013)

The first candidate for public release. The file structure was polished and consistent and a number of significant improvements made, including clamping or filtering anomalous data, removal of over 400 MERIS orbits with bad geolocation, exclusion of negative R_{rs} (which previously silently corrupted some v0.95 merged pixels), increasing the maximum zenith cutoff to 80 degrees to allow more good quality data to be included, switch of fill values from the programmatically difficult NaN to the standard float values,

V1.0rc2 / V1.0 (December 2013)

Following further QC, the zenith cutoff of 80 was changed to an air mass cutoff of 5, which better separated good and bad pixels. Mixed coastal pixels were filtered out. Three significant bugs were corrected: one in bias correction causing errors at high latitudes, one affecting merging with fill values and one resulting in bad uncertainty estimates for products with multiple wavelengths.

This release became the official v1.0 release on 14 Dec 2013; there are no changes between data from v1.0rc2 and v1.0 since then.

V2.0 (April 2015)

v2.0 extended the time series to the end of 2013, improved the in-situ database used for characterisation and quantification of error, developed specific water classes based on the v2.0 data rather than on Tim Moore's SeaWiFS-based classes, switched the NASA sensors to being consistently mapped by BEAM as with MERIS (correcting some pixelization issues noted in v1.0 in the process), incorporates an improved bias correction able to respond to temporal variation (primarily seasonal) and uses an improved cloud mask (Idepix 2.0) for MERIS.

This release was created and evaluated in January – March 2015 and formally released to the public in April 2015.

V3.0 (August 2016)

V3.0 extended the time series to the end of 2015, incorporated VIIRS (2012-) and SeaWiFS LAC (1km, 1997-2010) data, altered the binning algorithm to use super-sampling (better representing contributions of observations to data cells), switched MODIS level 2 processing to POLYMER (based on the AC round robin result), improved POLYMER retrievals especially in case 2, further improved the *in situ* database, changed the chlorophyll algorithm to blend results from multiple algorithms according to the water type memberships and amended the bias correction to have a smoother response to temporal variation.

The initial release candidate was prepared at the end of May 2016, but delayed due to concerns over flagging due to the vastly greater number of retrievals in Case-2 situations. Following flagging improvements and a high level of QC scrutiny, the data were formally released in at the end of August 2016.

V3.1 (May 2017)

V3.1 is essentially an extended version of v3.0, covering 2016 in the initial release. The project is waiting on a sufficient quality OLCI data release to produce the "OLCI-ready" v4.0. Ongoing v3.1 processing using the same processing chain is done on a daily basis, with outputs made available, but these cannot offer the "climate-grade" quality control, not least as recent years are more likely to be subject to sensor recalibrations, but also because daily updates cannot be subjected to the same level of quality control.

V4.2 (May 2019)

V4.2 was an updated release of 4.0 following the identification of an issue in the Kd product. The project was waiting on a sufficient quality OLCI data release to produce the "OLCI-ready" v4 but this then became incorporated into the v5.0 processing.

V5.0 (Oct 2020)

V5.0 was the first version to shift to MERIS as the reference sensor and to include OLCI data (from 3A sensor). In this record VIIRS and MODIS were included in their entirety (they were not cut off at the end of 2019 as in v6.0).