| | |
|---|---|
| **Project:** | Ocean Colour Climate Change Initiative (OC_CCI) – Phase Two |
| **Document title:** | Product User Guide |
| **Reference:** | D3.4 PUG |
| **Issued:** | 2 November 2016 |
| **Issue:** | 3.0.1 |
| **Authored:** | Mike Grant, Thomas Jackson, Andrei Chuprin, Shubha Sathyendranath, Marco Zühlke, Thomas Storm, Martin Boettcher, Norman Fomferra |
| **Reviewed:** | Steve Groom |
| **Approved:** | Shubha Sathyendranath |
| **Copyright:** |  |

**This document may be refined over the lifetime of the v3.0 data release. Please check http://www.esa-oceancolour-cci.org/ for the latest version.**

# Document Change Log

| Issue | Date | Comment |
|---|---|---|
| 3.0.0rc1 | 30 June 2016 | Initial creation from v2.0 PUG with updates for v3.0rc1 data release |
| 3.0.0rc2 | 10 Aug 2016 | Updates for v3.0rc2 (following flagging corrections to v3.0rc1) |
| 3.0.0 | 02 Sep 2016 | First public release with v3.0 dataset. |
| 3.0.1 | 02 Nov 2016 | Fix page numbering in table of contents |
| | | |
| | | |

# Table of Contents

# 1. In brief

## *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 thirteen projects, each addressing a particular ECV, complemented by the Climate Modelling User Group. 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 is continuing phase 2 with another 3 years of improvement and annual data releases.

The OC-CCI project is providing ocean colour ECV data, with a focus on case 1 waters, which can be used, for example, in climate change prediction and assessment models. OC-CCI aims to produce the highest quality data, perhaps not containing the very latest data, which may be adjusted in the light of recalibration or assessment.

The dataset is created by band-shifting and bias-correcting MERIS, MODIS and VIIRS data to match SeaWiFS data, merging the datasets and computing per-pixel uncertainty estimates.

## *What are the key features of the v3.0 dataset compared with v2.0?*

This data release has targeted a significant increase in input data (including VIIRS and SeaWiFS LAC), refreshing processing versions of the inputs and improvements to chlorophyll retrievals (including in case 2 waters).

Easily noticeable changes are:

- Inclusion of SeaWiFS LAC (1km) and VIIRS (2012 onwards)
- Improvements to POLYMER for MERIS resulting in higher numbers of retrievals, particularly in case 2 waters
- Change of the primary MODIS algorithm from SeaDAS l2gen to POLYMER, improving retrievals in glint conditions
- Further improvements in the binning algorithm to eliminate speckle effects from areas with low sampling resolution (e.g. swath edges or SeaWiFS 4km GAC pixels); binning now gives an approximation to flux-preserving algorithms.
- Bias maps are now generated with weekly composites rather than dailies as the input, while remaining sensitive to seasonal variation. This gives a smoother, fuller correction.
- Significantly more data (~7% vs v2.0), attributed to the improvements above
- Refreshes the input datasets to the latest versions as of Q1 2016
- Extends the time series to the end of 2015
- Further improvements to the in-situ database used for characterisation and quantification of error, now publicly available at http://www.earth-syst-sci-data.net/8/235/2016/

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

The key objectives for Phase 2 of the OC-CCI project include:

- initiation of cyclical processing, whereby a reprocessing and release to the user community is undertaken every year (2015, 2016, 2017);
- extension of products to Case 2 waters that contain substances such as suspended sediments and dissolved organic matter that modify ocean colour independently of phytoplankton;
- improvements to inter-sensor consistency in atmospheric correction algorithms;
- improvements to the uncertainty characterisation;
- exploring the possibility of using data from other sensors to fill the gap between MERIS and Sentinel-3 OLCI;
- preparation for Sentinel-3 OLCI.

The next official release (v4.0) is planned for February 2017. It will further address Case 2 waters (improvements to A/C), cloud flagging (Idepix improvements) and will take account of the first Sentinel-3 OLCI data, although it is unlikely that significant amounts of OLCI data can be included, as there will be few routine/operational data with sufficient quality control available by the release date.

## What physical variables are in OC-CCI?

| **Data variable** | **Accompanying uncertainty variables** | **Notes and further references** |
|---|---|---|
| Rrs_412 | Rrs_412_rmsd | Remote sensing reflectance at SeaWiFS wavelengths |
| Rrs_443 | Rrs_443_rmsd | |
| Rrs_490 | Rrs_490_rmsd | |
| Rrs_510 | Rrs_510_rmsd | POLYMER ATBD |
| Rrs_555 | Rrs_555_rmsd | NASA SeaDAS/l2gen documentation |
| Rrs_670 | Rrs_670_rmsd | |
| | Rrs_412_bias | |
| | Rrs_443_bias | |
| | Rrs_490_bias | |
| | Rrs_510_bias | |
| | Rrs_555_bias | |
| | Rrs_670_bias | |
| chlor_a | chlor_a_log10_rmsd<br>chlor_a_log10_bias | Chlorophyll-a, estimated using a blended combination of OC3, OCI (OC4+CI) and OC5 algorithms.<br><br>Blending ATBD<br>In-water RR doc |
| atot_412<br>atot_443<br>atot_490<br>atot_510<br>atot_555 | *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) |

| | | |
|---|---|---|
| atot_670 | | QAA paper |
| adg_412 | adg_412_rmsd | QAA absorption due to detrital |
| adg_443 | adg_443_rmsd | and dissolved matter |
| adg_490 | adg_490_rmsd | |
| adg_510 | adg_510_rmsd | QAA paper |
| adg_555 | adg_555_rmsd | |
| adg_670 | adg_670_rmsd | |
| | adg_412_bias | |
| | adg_443_bias | |
| | adg_490_bias | |
| | adg_510_bias | |
| | adg_555_bias | |
| | adg_670_bias | |
| bbp_412 | *Insufficient in-situ data to make* | QAA backscatter due to |
| bbp_443 | *a plausible estimate* | particulate matter |
| bbp_490 | | |
| bbp_510 | | QAA paper |
| bbp_555 | | |
| bbp_670 | | |
| kd_490 | kd_490_rmsd | Attenuation coefficient (Lee |
| | kd_490_bias | algorithm with Zhang backscatter coefficients) |
| water_class1 | *n/a* | Water class memberships |
| water_class2 | | according to Moore et al. (2009) |
| water_class3 | | and class definitions per the CCI |
| water_class4 | | derivations (broadly, classes |
| water_class5 | | range from open ocean to coastal |
| water_class6 | | waters as the class number |
| water_class7 | | increases) |
| water_class8 | | |
| water_class9 | | Water class ATBD |
| water_class10 | | |
| water_class11 | | |
| water_class12 | | |
| water_class13 | | |
| water_class14 | | |

## *What changes are needed to v2.0-compatible programs so they work with the new v3.0 data?*

v3.0 should generally be format-compatible with v2.0 for all the core parameters, attributes and variables.. There will obviously be minor updates to the content of attributes, such as version numbers or algorithm descriptive text, but the structure and content are the same.

The only format change is in the data type of the number-of-observations variables (*total_nobs, MERIS_nobs, MODIS_nobs, SeaWiFS_nobs* and the new *VIIRS_nobs*). Previously these were integers, reflecting a direct count of the number of observations falling into a cell. In v3.0, they are now floats, meaning that there may be "partial" observations from a sensor into 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.

## *What changes are needed to v1.0-compatible programs so they work with v2.0/v3.0 data?*

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 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. In v3.0, they are now floats, meaning that there may be "partial" observations from a sensor into 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.

## *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 Compliant WMS/WCS services and OPeNDAP are available. Please see the data product description page on the OC-CCI website for links to pages detailing these:

http://www.esa-oceancolour-cci.org/?q=products%20description

Many of the advanced data services, including a visual product browser in the style of the NASA oceancolor portal, are available on the general ocean colour portal:

http://www.oceancolour.org/

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

## How do these compare with other data sets?

For a comparison of v3.0 against earlier versions of OC-CCI, please see section 5.

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)

Space precludes a detailed comparison here, but CCI's primary focus is on 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:

http://www.esa-oceancolour-cci.org/?q=node/208

## Where can I get detailed information?

All project documentation and related publications can be found at the website:
http://www.esa-oceancolour-cci.org/ (on the left side, click “resources” to expand it, then there are menu items for documents and publications)

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 1.6) – 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)
- GlobColour Product User Guide (http://globcolour.info/)

## 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 citation in the acknowledgements, alongside any description within the methodology:

Ocean Colour Climate Change Initiative dataset, Version [*Version Number*], European Space Agency, available online at http://www.esa-oceancolour-cci.org/

We would also appreciate being notified so that we can list publications at:

http://www.esa-oceancolour-cci.org/?q=publications

## *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:

http://www.esa-oceancolour-cci.org/?q=contact%20points

# 2. Using the products

The first point to highlight is that these are novel 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.

## *Applicability to different water types*

The focus of phase 1 of OC-CCI and the first year of phase 2 was 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 Case-2 waters. 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, v3.0 has given extra consideration 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 attempts to weight the outputs of the best-performing algorithms (from OCI [OC4+CI], OC3 and OC5) 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 therefore 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 will still offer better performance, but global scale analyses will find improved results in v3.0.

## *Interpreting data values*

Upper and lower limits to the products have been applied, based on what we know to be realistic in Case-1 waters and also based on the range in the values used for validation and error 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 from the products
- 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.
- A filter was applied to MERIS and MODIS that discriminated against 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 670 nm.

## *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, $\Delta$) and bias ($\delta$), 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 (Rrs) 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). The v3.0 optical classification includes 14 classes, identical to those in v2.0 (whereas v1.0 had 9 classes). The primary difference between v3.0/v2.0 and v1.0 of OC-CCI is that in v2.0 the optical classification is based on OC-CCI satellite Rrs data as input to the classification, instead of in-situ Rrs data. For further details regarding optical classification, please see section 5.

The root-mean-square difference ($\Delta_p$) for each product, for each pixel $\mathrm{p}$, 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 $\mathrm{k} = 1 \ldots \mathrm{K}$ are the optical classes, $\Delta_k$ is the root-mean-square difference for each class, and $\mathrm{w}_{k,p}$ are the weighting factors, (the fuzzy memberships) of each of the $\mathrm{K}$ optical water classes for that pixel.

The bias $\delta_p$ is computed similarly:

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

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

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

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 $R_{rs}$. . Other known problems, such as those at high latitudes, are not accounted for in this error budget.

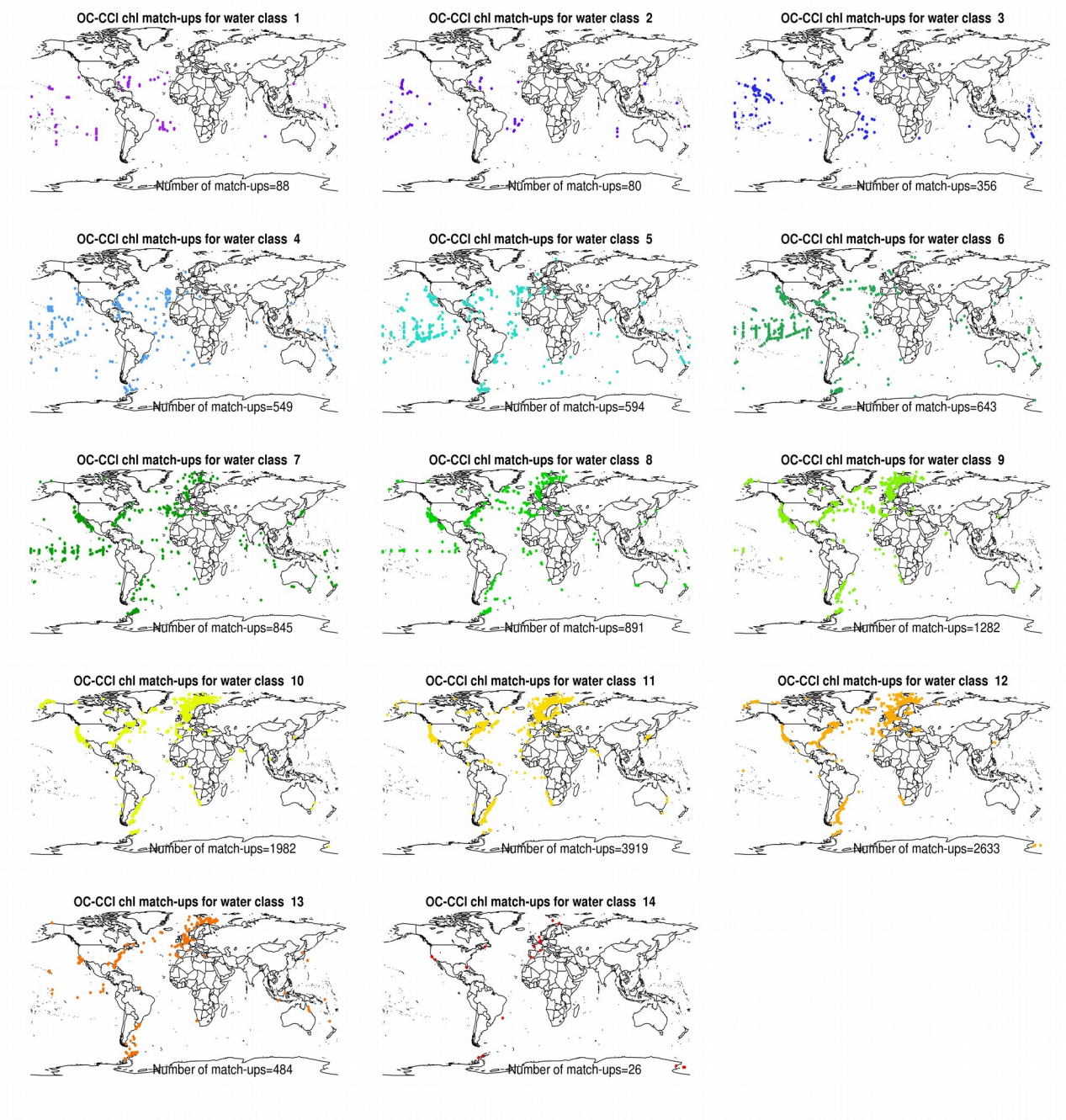


*Figure 1: Geographical coverage of water types for Chl-a matchups*

## Computation of unbiased data

The error statistics provided with the v3.0 products allow the user to compute unbiased values of the products. Unbiased values $_p$ of any variable X can be computed as:

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

In the case of chlorophyll data, $\delta_p$ is provided for $\log_{10}$ chlorophyll. So, for the particular case of chlorophyll, the unbiased chlorophyll at pixel p, say $_p$, 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 $\delta_p$, hence $\delta_p < 0$ implies an overestimation by the satellite.

## *Statistical Properties of the Chlorophyll Fields*

It is important to bear in mind that uncertainties in chlorophyll ($\Delta_p$, $\delta_p$, and hence $\sigma_p$) 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 $\sigma_p$ included in the products, to its corresponding value $\sigma_e$ for natural-log-transformed data:

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

A similar relationship exists between $\mu_p$, the mean of the common-log-transformed distribution and the corresponding mean $\mu_e$ for the natural-log-transformed data:

$$\mu_e = \ln(10)\mu_p \quad \text{(Equation 2.7)}$$

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, $_p$, 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 $\mu_e$ according to the equation below:

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

or,

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

such that $\mu_e$ can be estimated from the quantities provided. This equation clearly shows that the mean of the log-transformed data ($\mu_p$) 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:

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

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

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

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 $\mu_p \pm 2\sigma_p$

$\mu_p \pm 2\sigma_p$ 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 $10^{(\mu_p+2\sigma_p)}$ in the untransformed units. Similarly, the lower confidence limit is $(\mu_p - 2\sigma_p)$ in the transformed units or $10^{(\mu_p-2\sigma_p)}$ in the untransformed units. Confidence limits calculated in this fashion lie about the quantity $10^{\mu_p}$, which is the geometric mean (or the median) of the untransformed chlorophyll values.

## *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 $\mathrm{X}_c$ for a composite c of N pixels can be computed as:

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

the composite root-mean-square deviation as:

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

and the bias as:

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

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

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

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.

### *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 $\mathrm{N}_v$ of valid pixels in each bin, each with errors characterised by RMSD $\Delta_p$, bias $\mathrm{m}_p$, standard deviation $\sigma_p$ and water class membership $\mathrm{W}_p$ .Then the uncertainties in the composite product can be computed as:

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

$$\mathrm{m}_c = \frac{\Sigma_{i=1}^{N_v} m_p}{N_v}$$

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

$$\mathrm{W}_c = \frac{\Sigma_{i=1}^{N_v} W_p}{N_v}$$

# 3. 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 (released 2008) 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

Although we expect the new SNAP (http://step.esa.int/main/toolboxes/snap/) toolbox to take over eventually, the current analysis package recommended is the BEAM toolbox, which is specifically developed by ESA for the exploitation of Earth Observation data products. BEAM is open source and freely available from http://earth.esa.int/beam Regarding the OC-CCI products, BEAM 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 BEAM (see screenshot below)
- band arithmetic using a fast expression language

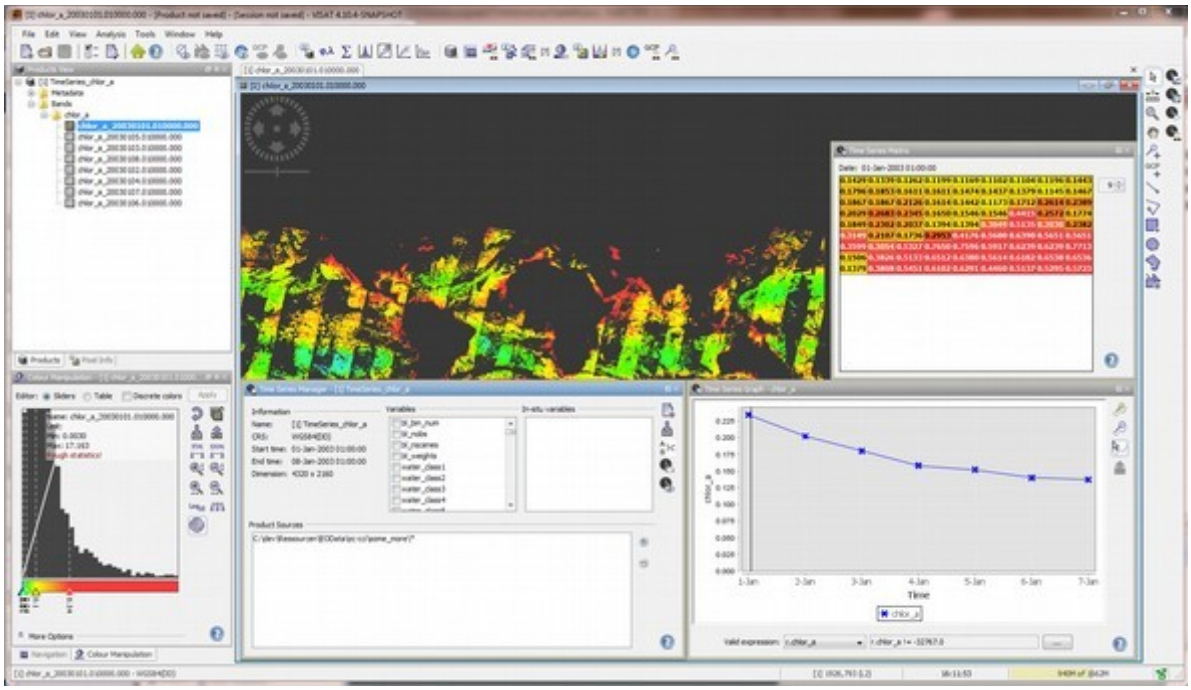

*Figure 2 - Time series of chlor-a variable*

At the time of writing (June 2016) the current released version of BEAM (5.0) in its default configuration requires an update of the modules before it is capable of reading the v2.0/v3.0 sinusoidal product format, but can read the geographic format without difficulties. To perform the update:

1. Make sure you have BEAM version 5.0 installed
2. Open the Module Manager by selecting Help -> Module Manager…

3. Open the Module Updates tab
4. Select the Level-3 Binning Processor and select the Update button
5. Start the update process with the Ok button

An easy 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 framework with extensions that provide the functionality of previous versions of SeaDAS with a tilt towards the NASA products. SeaDAS versions 7.3 and up handle both CCI projections without updates.

Additionally, there is the Panoply data viewer that NASA provides free of the 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.

### *Sample programs in various languages*

Please see the website for more examples.

## Python

There are several netCDF capable libraries, but PML most commonly uses “netCDF4” (available from http://code.google.com/p/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()
```

## 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)
```

## 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)
```

# 4. Known issues

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 with the correction incorporated.

## *Major errors*

None found so far.

## *Data errors*

None found so far

## *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 ones that are relatively uncommon / few in number through the time series, though they may include noteworthy pixels such as coccolithophore blooms.

**Decay of MODIS calibration:** NASA monitor the calibration of MODIS (Aqua) and regularly adjust it. The last processing was r2014.0, working to correct some of the degradation noted. It is to be expected that data after ~2013 will have reduced quality and further improvements are likely to be attempted by NASA. The OC-CCI v3.0 dataset is less vulnerable to this degradation than v2.0, as it incorporates VIIRS (now stated by NASA to be more reliable than MODIS Aqua). However, data in this period should be analysed with caution.

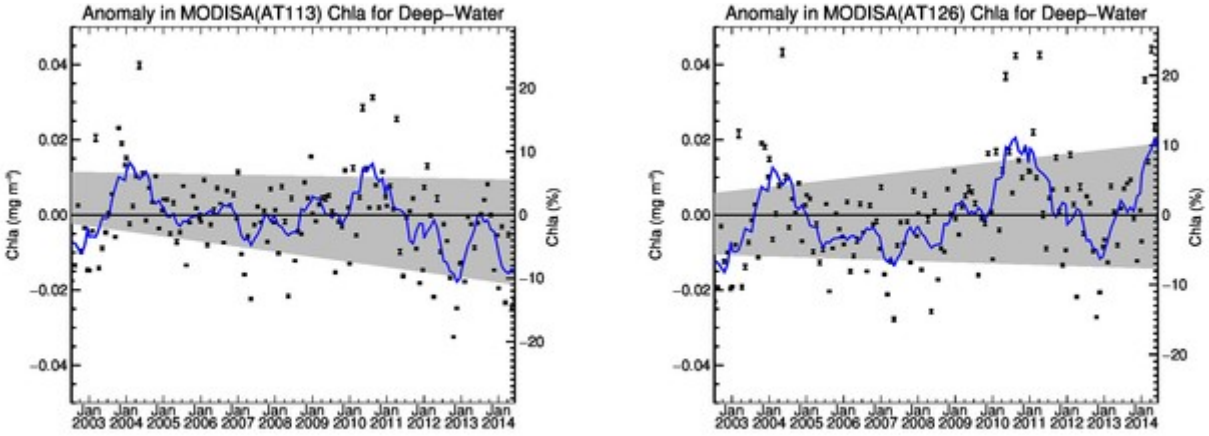


*Figure 3: Improvement in the chlorophyll anomalies between NASA MODIS reprocessing versions. Further information available at* [http://oceancolor.gsfc.nasa.gov/cms/reprocessing/OCReproc20140MA.html](http://oceancolor.gsfc.nasa.gov/cms/reprocessing/OCReproc20140MA.html)

## *Trivial issues*

**IOP standard_name attributes** in the NetCDFs are insufficiently descriptive (no distinction between aph, adg, etc). This is because improved names have not yet been accepted into the CF standard name list, despite apparent consensus having been reached a number of years ago. This can only be resolved only when suitable names are accepted into CF.

## *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, MODIS from July 2002 and on-going, and VIIRS from 2012 and on-going).

**No uncertainty for atot and bbp:** atot 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. 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.

## *Noteworthy changes from v2.0 format*

v3.0 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* 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 v3.0, they are now floats, meaning that there may be "partial" observations from a sensor into 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.

## *Noteworthy changes from v1.0 format (applies to v2.0 & v3.0)*

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

**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 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. In v3.0, they are now floats, meaning that there may be "partial" observations from a sensor into 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.

# 5. 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 in section 6.

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.

## *Comparison of OC-CCI v3.0 and OC-CCI v2.0*

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 4.

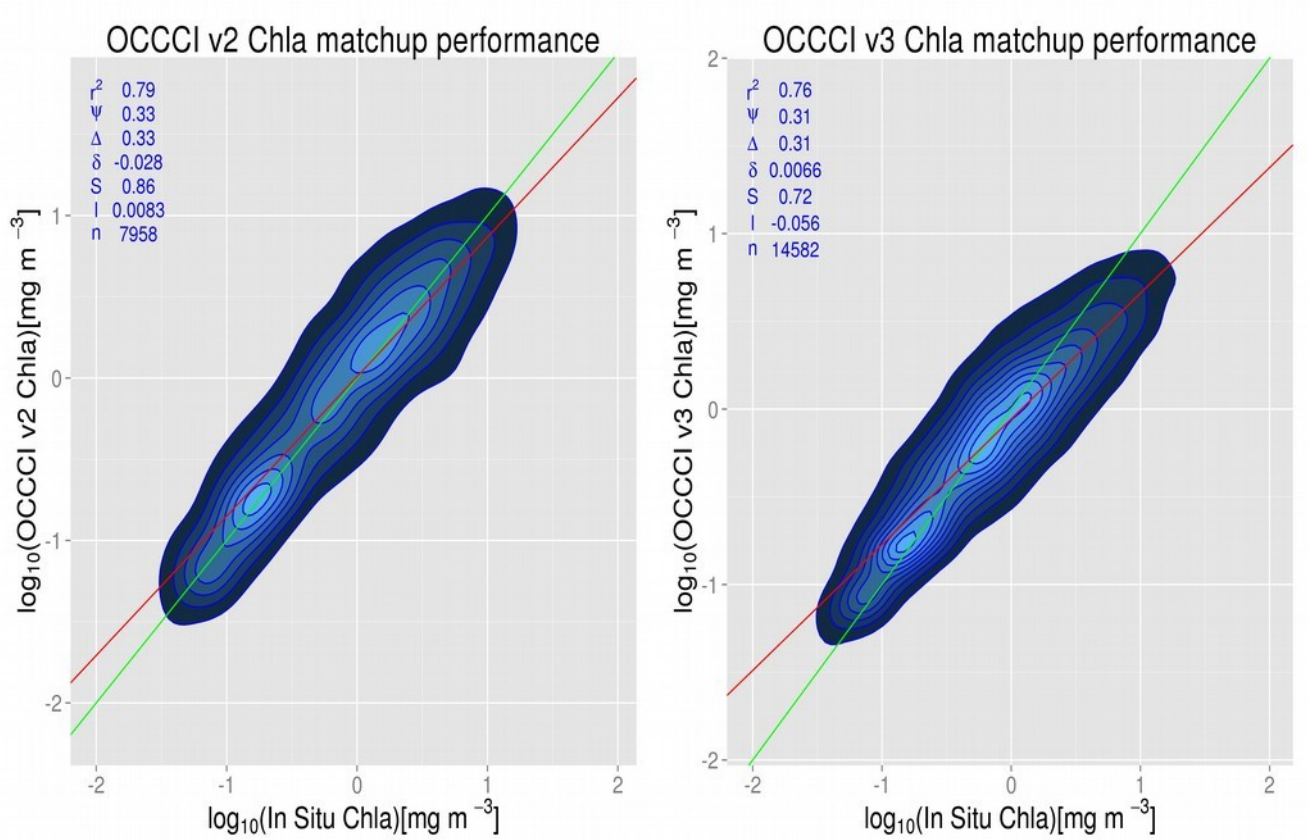


*Figure 4: Comparison of v2.0 and v3.0 of the OC-CCI chl product when matched against in-situ chl-a measurements. Summary statistics shown are correlation coefficient, root-mean-square-difference (RMSD), un-biased RMSD, bias, slope of regression, intercept of regression, number of match-ups and the median-absolute-deviation of residuals.*

It should be noted that the change in the performance statistics for a product, such as chlorophyll, between v2.0 and v3.0 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 v3.0 chlorophyll product has a greater number of match-ups, smaller bias and smaller RMSD than v2.0.

To compare the uncertainty variables between differing OC-CCI datasets one can look at the large scale (global) range and distribution of uncertainty values. Figure 5 show an example of the differences in uncertainty between the two version of OC-CCI data. Overall the uncertainties have reduced between v2.0 and v3.0 but the change is not geographically uniform. The low chlorophyll waters have seen a reduction in the rmsd error and the higher chlorophll waters tend to have a smaller bias.

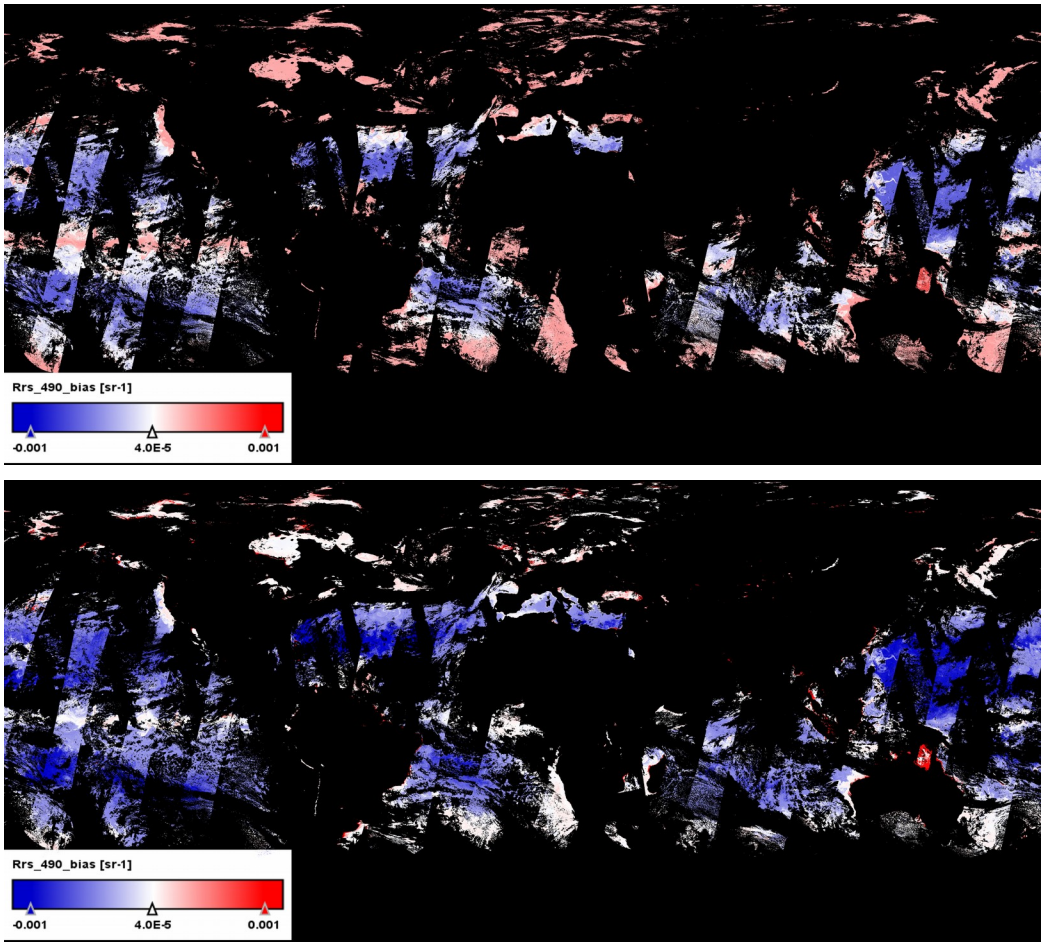


*Figure 5: Comparison of the Rrs 490 bias uncertainty as given in v1.0 (top) and v2.0 (bottom).*

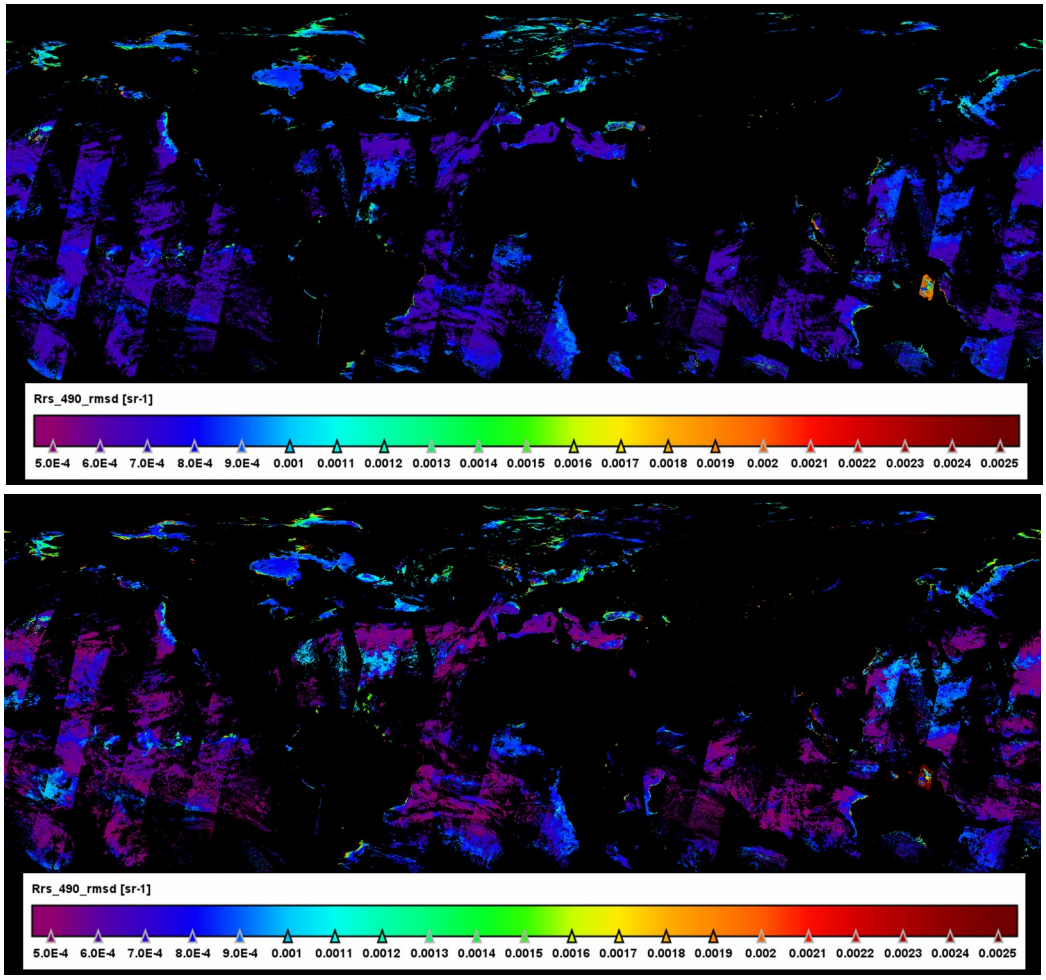


*Figure 6: Comparison of the Rrs 490 rmsd uncertainty as given in v2.0 (top) and v3.0(bottom).*

Overall the OC-CCI v3.0 data has an increased spatial coverage of data, increased performance at retrieving chlorophyll concentrations and a more representative set of uncertainties than the v2.0 dataset.

## *Product overview*

## 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-OCAB document. For v3.0 this involved the blending of the OCI algorithm (as implemented by NASA, itself a combination of CI and OC4), the OC5 algorithm (NASA 2010) and the OC3 algorithm. Each algorithm utilises the same OC-CCI merged $R_{rs}$ products described below.

Figures 7-9 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.

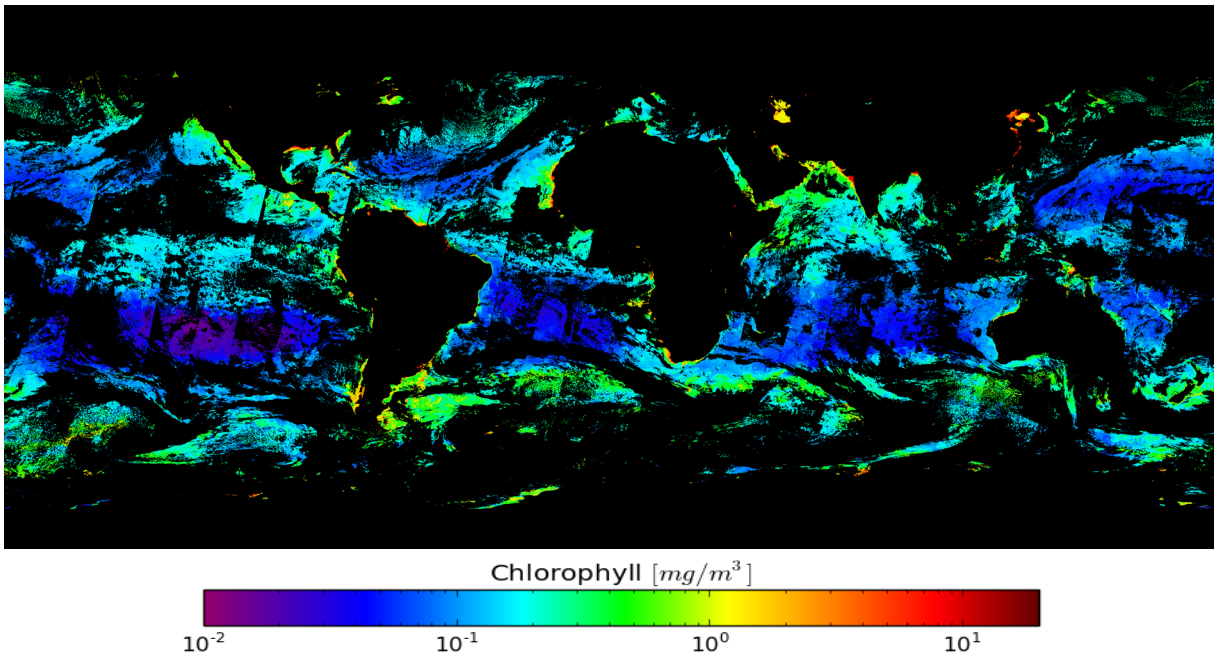


*Figure 7: Chlorophyll-a concentration (1st Jan 2003)*

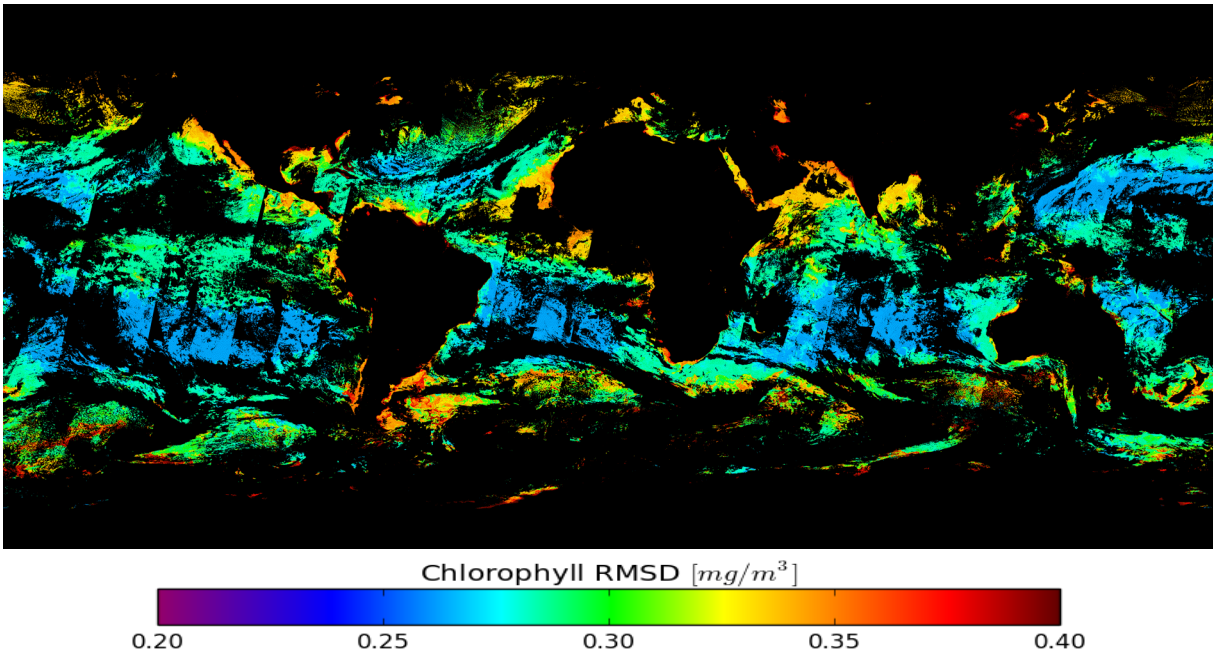


*Figure 8: Root mean square difference of chlorophyll-a concentration (1st Jan 2003)*

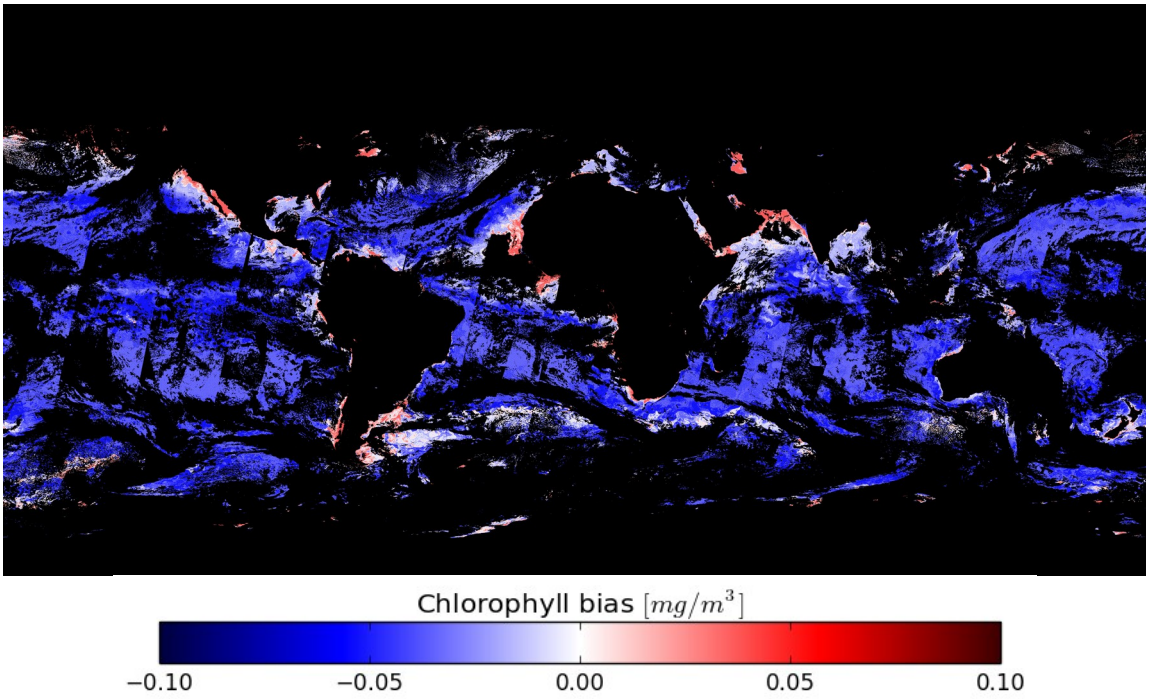


*Figure 9: Bias of chlorophyll-a concentration (1st Jan 2003)*

## 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 SeaWiFS wavelengths (412, 443, 490, 510, 555, 670nm) with pixel-by-pixel uncertainty estimates for each wavelength. These are merged products based on SeaWiFS, MERIS, Aqua-MODIS and VIIRS 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 MERIS, MODIS and VIIRS were band-shifted to SeaWiFS wavebands if necessary, and MERIS and MODIS were corrected for inter-sensor bias when compared with SeaWiFS in the 2003-2007 period. VIIRS was also corrected to SeaWiFS levels, via a two-stage process comparing against the MODIS-corrected-to-SeaWiFS-levels (2012-2013).

## 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.

## 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 SeaWiFS wavelengths (412, 443, 490, 510, 555, 670nm). They were computed from daily, merged $R_{rs}$ values using the Lee et al. (2009) algorithm. 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 10 to 12 show global daily images of total absorption, absorption of detrital and dissolved matter, and absorption by phytoplankton at 443 nm.

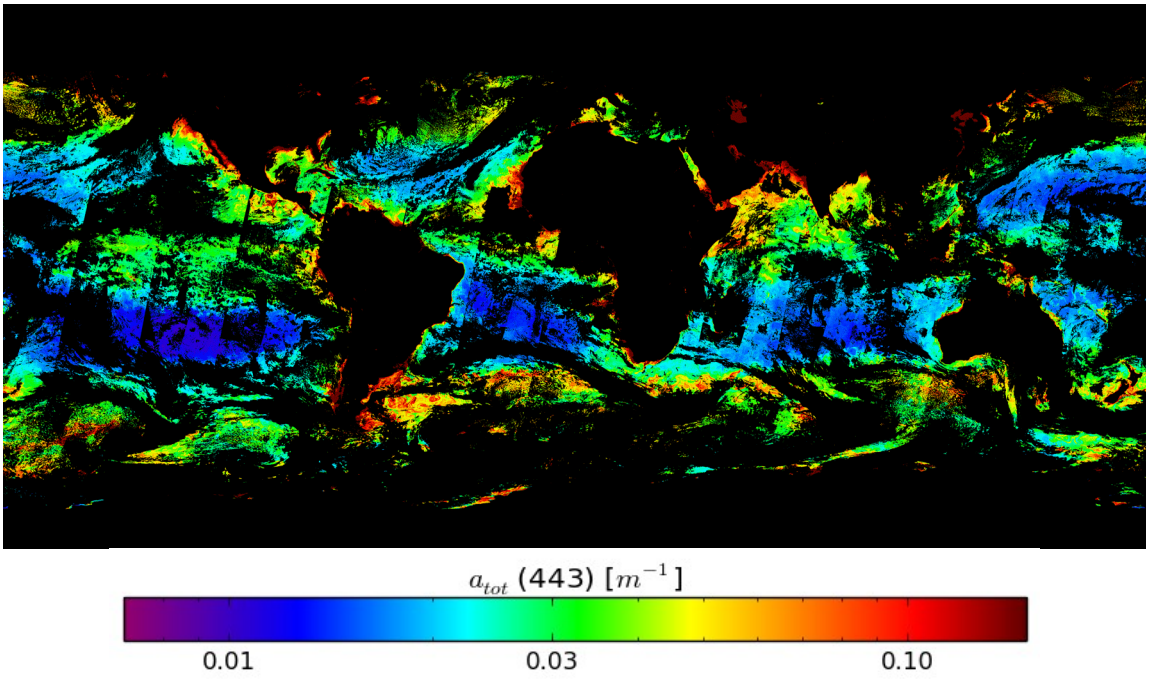


*Figure 10: Total absorption at 443 nm (1st Jan 2003)*

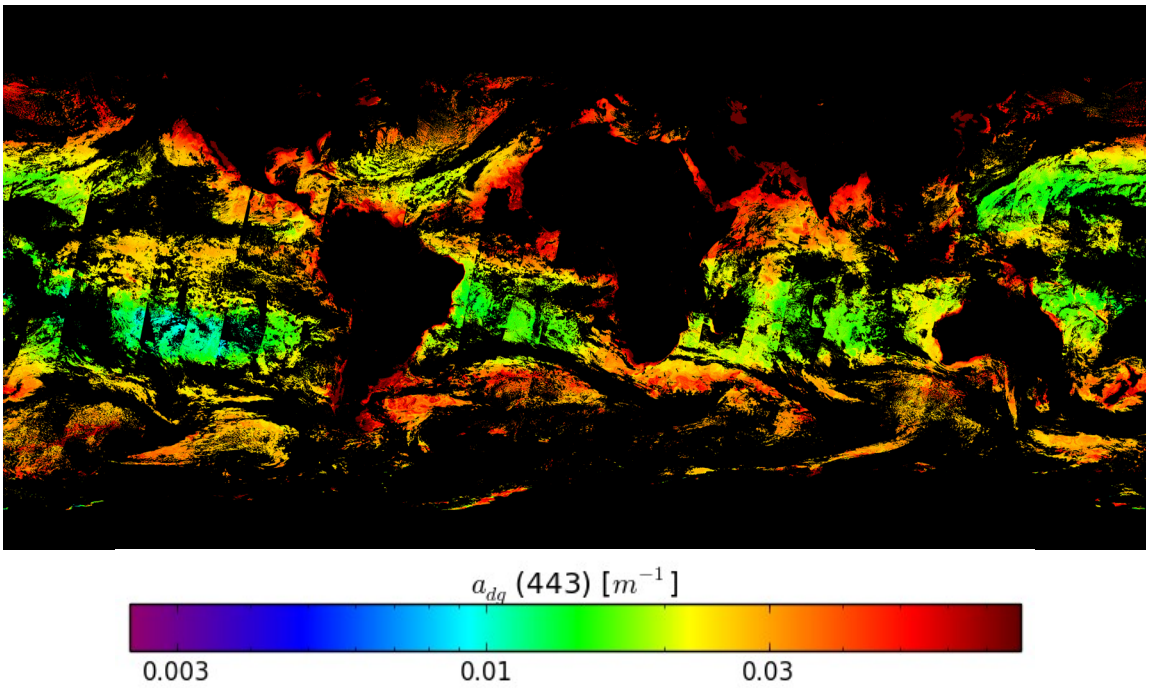


*Figure 11: Absorption by detrital and dissolved matter at 443 nm (1st Jan 2003)*

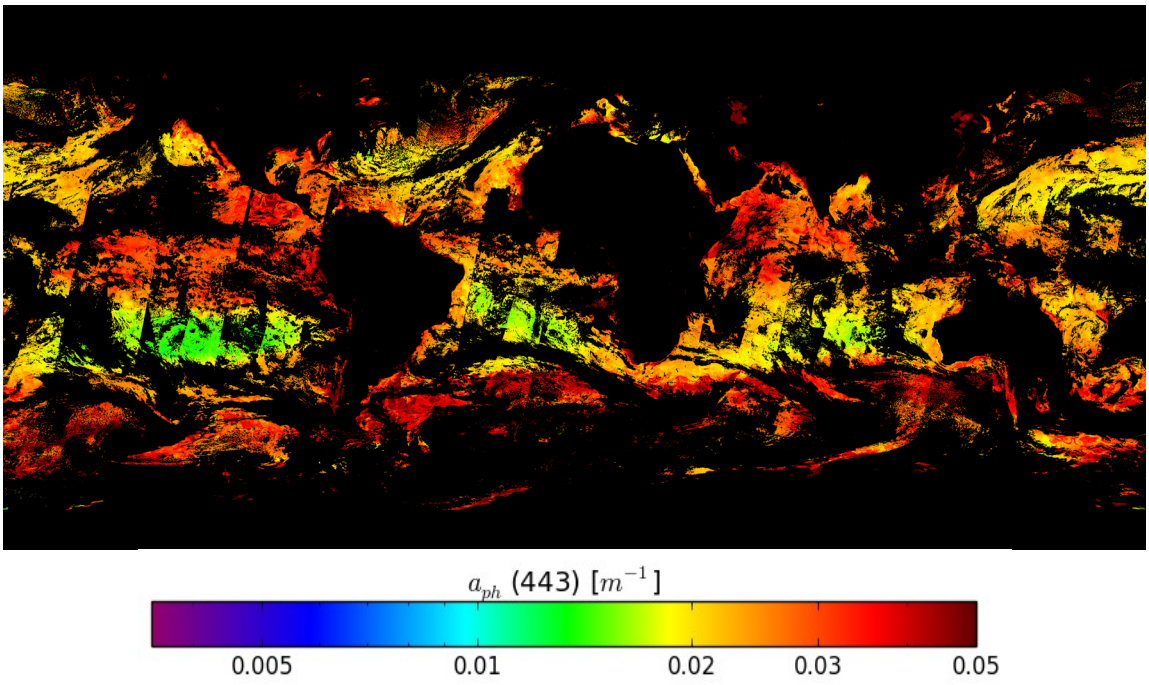


*Figure 12: Phytoplankton absorption at 443 nm (1st Jan 2003)*

## 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).

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

## 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), 14 specific classes have been derived that best match the observations. Each has differing spectral reflectance shapes that allow the separation of waters with similar chlorophyll concentrations but differing composition and optical properties (Moore et al 2009). Figure 13 shows the spectral shapes of the final classes:

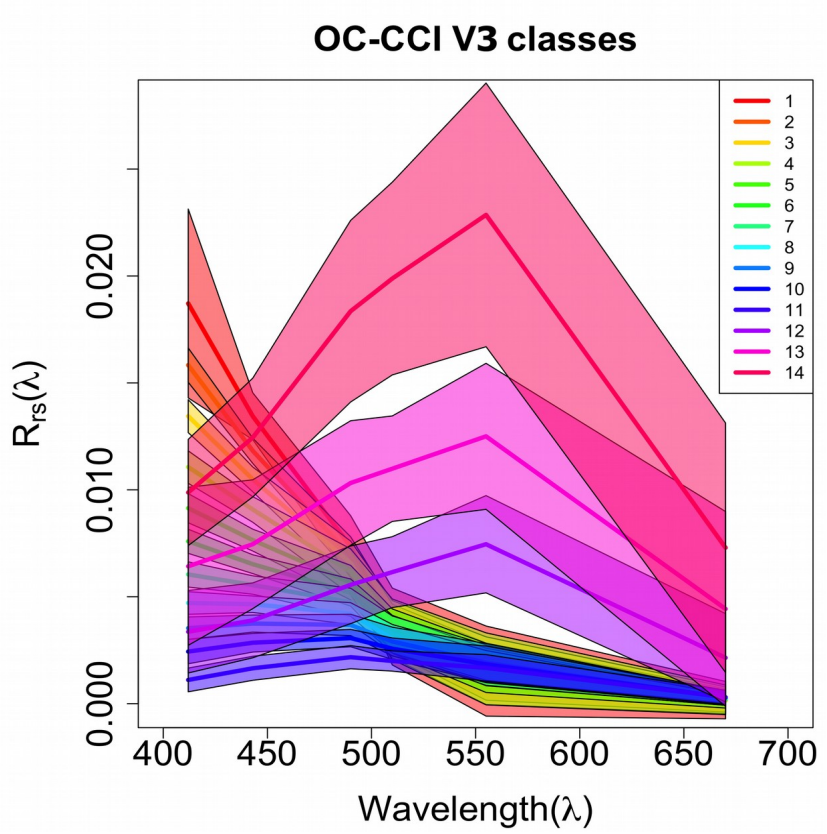


*Figure 13: Spectral response of the water types used in OC-CCI v2.0 & v3.0 products (hard lines are class means and shaded region shows standard deviation).*

## *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 too 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

# 6. The products: technical overview

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

## *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 v3.0 data release, a typical netCDF file containing the full set of products for a single day is approximately 1.7GB. Subsetted versions of these files containing only related product groups (e.g. chlorophyll, Rrs, IOPs, etc) and advanced data services (e.g. OPeNDAP) are available to mitigate download size problems.

## *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 “3.0” for the v3.0 data |

## Example filename

An example filename is:

```
ESACCI-OC-L3S-OC_PRODUCTS-MERGED-1D_DAILY_4 km_GEO_PML_OC4v6_QAA-20031225-fv3.0.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* | Rrs 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*<br>*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 |

## *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.

## 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.

## 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.

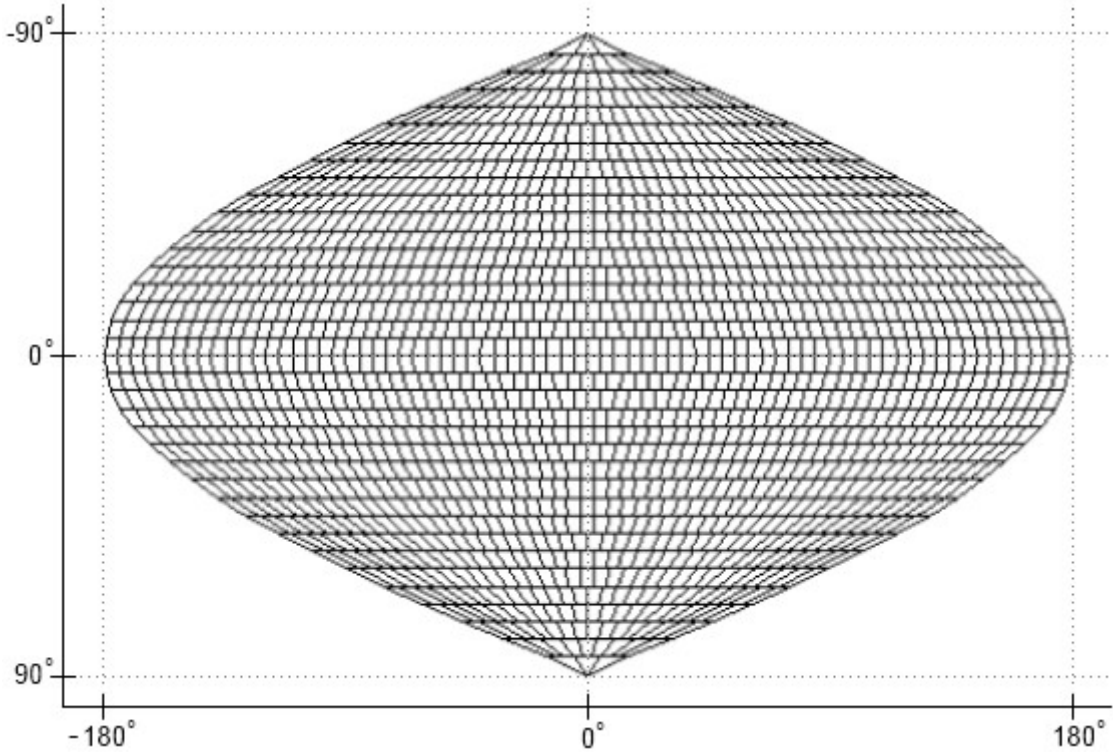


*Figure 14: The sinusoidal grid*

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.

### *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.

## 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.

## 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.

## 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.

## 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).

## 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 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_412_rmsd | Remote sensing reflectance at SeaWiFS wavelengths |
| Rrs_443 | Rrs_443_rmsd | |
| Rrs_490 | Rrs_490_rmsd | |
| Rrs_510 | Rrs_510_rmsd | |
| Rrs_555 | Rrs_555_rmsd | |
| Rrs_670 | Rrs_670_rmsd | |
| | Rrs_412_bias | |
| | Rrs_443_bias | |
| | Rrs_490_bias | |
| | Rrs_510_bias | |
| | Rrs_555_bias | |
| | Rrs_670_bias | |
| chlor_a | chlor_a_log10_rmsd<br>chlor_a_log10_bias | Chlorophyll-a concentration in seawater (not log-transformed), generated using a blended combination of OCI (OC4v6 + Hu's CI), OC3 and OC5, depending on water class memberships |

| | | |
|---|---|---|
| atot_412<br>atot_443<br>atot_490<br>atot_510<br>atot_555<br>atot_670 | *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<br>aph_443<br>aph_490<br>aph_510<br>aph_555<br>aph_670 | aph_412_rmsd<br>aph_443_rmsd<br>aph_490_rmsd<br>aph_510_rmsd<br>aph_555_rmsd<br>aph_670_rmsd<br>aph_412_bias<br>aph_443_bias<br>aph_490_bias<br>aph_510_bias<br>aph_555_bias<br>aph_670_bias | QAA absorption due to phytoplankton |
| adg_412<br>adg_443<br>adg_490<br>adg_510<br>adg_555<br>adg_670 | adg_412_rmsd<br>adg_443_rmsd<br>adg_490_rmsd<br>adg_510_rmsd<br>adg_555_rmsd<br>adg_670_rmsd<br>adg_412_bias<br>adg_443_bias<br>adg_490_bias<br>adg_510_bias<br>adg_555_bias<br>adg_670_bias | QAA absorption due to detrital and dissolved matter |
| bbp_412<br>bbp_443<br>bbp_490<br>bbp_510<br>bbp_555<br>bbp_670 | *Insufficient in-situ data to make a plausible estimate* | QAA backscatter due to particulate matter |
| kd_490 | kd_490_rmsd<br>kd_490_bias | Attenuation coefficient (Lee algorithm with Zhang backscatter coefficients) |
| water_class1<br>water_class2<br>water_class3<br>water_class4<br>water_class5<br>water_class6<br>water_class7<br>water_class8<br>water_class9<br>water_class10<br>water_class11 | *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) |

water_class12
water_class13
water_class14

## Data sources (number of observations)

The NetCDFs contain variables indicating how many observations were made of a specific data cell. There is a total and also 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 and VIIRS data were originally 1km prior to binning. Consequently the latter two sensors can contribute ~16 times as many observations 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 1/9$^{th}$s, 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" ;
```

## *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.

| 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

# 7. How were the products made?

A thorough description of the OC-CCI processing chain is given in the Ocean Colour System Prototype Specification document. This section briefly recapitulates an overview of the processing chain. Please refer to Figure 15 below.

## Input datasets

The input EO datasets were MERIS Reduced-Resolution (1km) L1b 3rd reprocessing (including OCL fixes), MODIS level 1A data (implicitly no version identifier), R2014.0.1 level 2 VIIRS from NASA and SeaWiFS level 2 LAC (1km / MLAC) and GAC (4 km) R2014.0.

## Level 2 processing and binning

MERIS and MODIS were processed with the POLYMER algorithm (v3.5) to level 2. MODIS used l2gen to get to level 1c (geometric, sensor corrections and other non-A/C work) before POLYMER created the final Rrs. SeaWiFS and VIIRS L2 were downloaded from NASA (implicitly meaning they were processed with l2gen from SeaDAS 7.3).

All individual sensors were binned to level 3 4 km (sinusoidal grid) with the BEAM binner. MERIS was masked using the IDEPIX v2.0 cloud and land flags. SeaWiFS was masked using a combination of the standard NASA flags and IDEPIX v2.0. MODIS and VIIRS used the NASA flagging as finalised IDEPIX is not yet available for these datasets.

All available data were used, up the end of 2015. It should be noted that NASA considers MODIS' calibration from 2013 onwards to be degrading, although they have tried to compensate for this in r2014. Data after 2013 should therefore be analysed with more caution.

## Band shifting

MODIS, MERIS and VIIRS were band shifted to the six main SeaWiFS bands (412, 443, 490, 510, 555, 670nm) by computing QAA IOPs and back computing the Rrs bands using a high-resolution spectral model. The output Rrs for 412-555nm were cleaned of any negative values, with the data items removed. Negative Rrs values In the 670nm band frequently occur due to low signal levels, and these were clamped to zero.
Nothing was done to the SeaWiFS data.

## Bias correction

The band shifted MERIS and MODIS Rrs were corrected to remove gross differences (biases) against SeaWiFS Rrs. 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 (vs 1 day ones in v2.0), such that the corrections take account of seasonal and regional variations. The biases were computed over the 2003-2007 period of all sensors overlapping and functioning well. Bias adjustments were computed at every location where all sensors had gathered data, with a temporal window of +/- 30 days (weighted by the time difference from the centre point) and spatially-limited interpolation (11 pixels) to fill smaller gaps.
VIIRS is then also corrected to SeaWiFS levels by a similar process, but comparing against MODIS-corrected-to-SeaWiFS-levels rather than directly to SeaWiFS. This indirect comparison is unavoidable due to the lack of temporal overlap.

## Merging

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

## Water class membership

Water classes were computed following Moore et al (2009), but with (14) specific water classes derived originally from the v2.0 CCI Rrs values, but verified against v3.0. The classes were derived by identifying the most representative spectra in the CCI observations and picking the top N classes such that the majority of spectra were covered (with higher N producing more specific classes, but at a higher cost in storage, reduced generality and a reduced number of matchups for each class).

## Product generation

A range of products were computed from the merged Rrs, 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 OC3, OCI (OC4+CI) and OC5, weighted by the relative levels of membership in specific water classes.
- IOP: QAA (with Zhang bb coefficients)
- KD: Lee variant (with Zhang bb coefficients)
- Rrs: Mixed – SeaDAS for MODIS and SeaWiFS; Polymer for MERIS. Bandshifted to SeaWiFS bands and cleaned up.

## Uncertainty estimation

A table of uncertainties for each class were computed from matchups between the CCI in-situ database and the version 3.0 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.

## 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.

## Additional/derived products.

For both projections, product subsets are created so that users wanting only a specific subset (e.g. just chlorophyll, IOP or Rrs 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 (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.

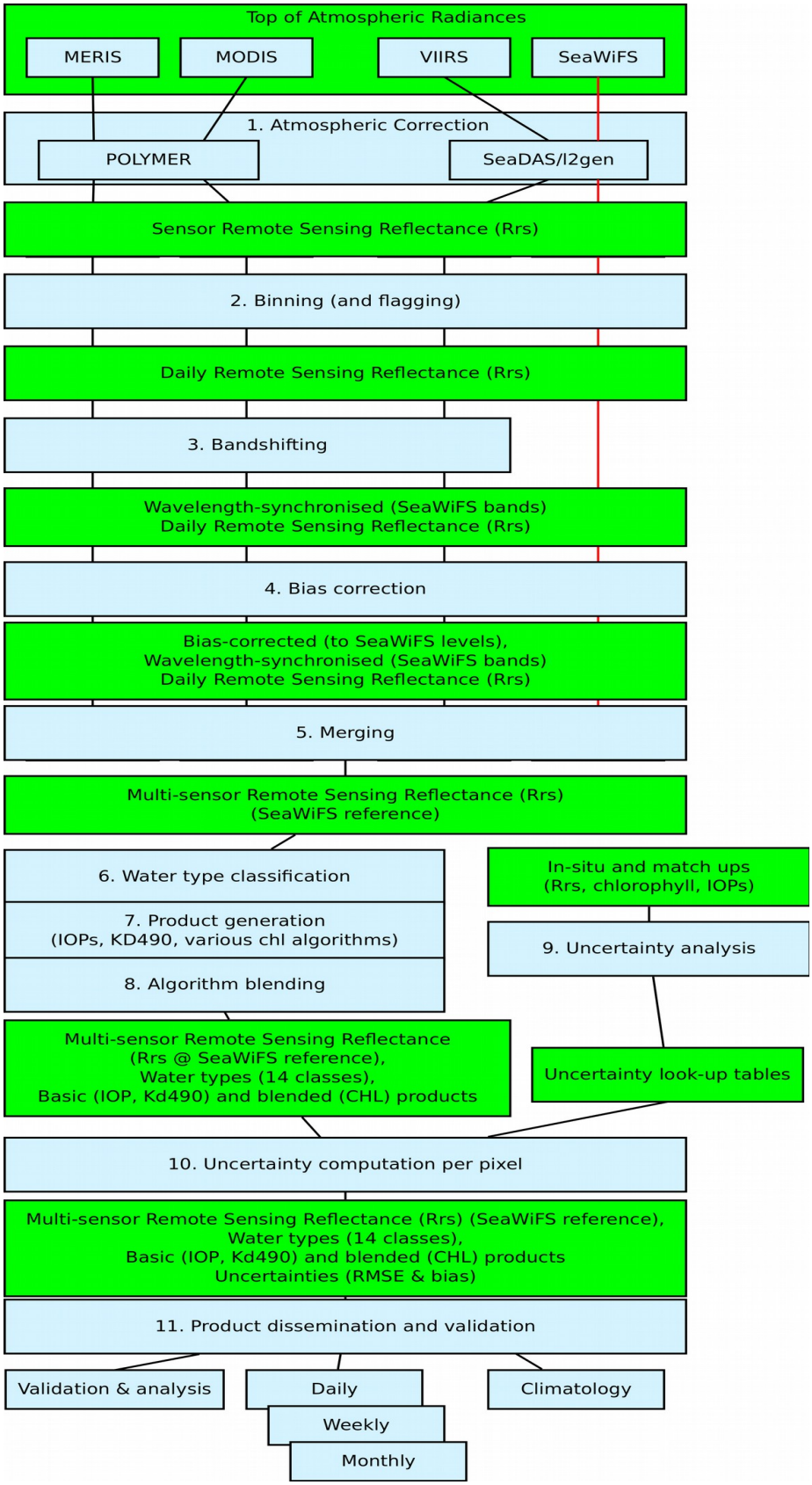


*Figure 15: Data flow in the Ocean Colour ECV production*

# 8. Earlier versions

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.

## A.1 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.

## A.2 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.

## A.3 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 Rrs (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,

## A.4 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.

## A.5 v2.0 (April 2015)

v2.0 extended the time series to the end of 2013, improved the in-situ database used for characteristion 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 pixelisation 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.

## A.6 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 supersampling (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 insitu 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 scrunity, the data were formally released in at the end of August 2016.