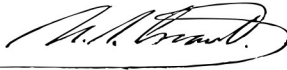





<b>Project:</b>	Ocean Colour Climate Change Initiative (OC_CCI) – Phase Two
<b>Document title:</b>	Product User Guide
<b>Reference:</b>	D3.4 PUG
<b>Issued:</b>	9 April 2015
<b>Issue:</b>	2.0.3
<b>Authored:</b>	 Mike Grant, Thomas Jackson, Andrei Chuprin,, Shubha Sathyendranath, Marco Zühlke, Thomas Storm, Martin Boettcher, Norman Fomferra
<b>Reviewed:</b>	Steve Groom
<b>Approved:</b>	 Shubha Sathyendranath
<b>Copyright:</b>	© Plymouth Marine Laboratory 2015. Licensed under the <a href="http://creativecommons.org/licenses/by/4.0/">Creative Commons Attribution 4.0 International license</a> .

This document may be refined over the lifetime of the v2.0 data release. Please check <http://www.esa-oceancolour-cci.org/> for the latest version.



## Document Change Log

<b>Issue</b>	<b>Date</b>	<b>Comment</b>
2.0.0	27 Feb 2015	Initial creation from v1.0 PUG and updates for v2.0 data release
2.0.1	11 March 2015	Content additions to address internal review comments
2.0.2	20 March 2015	Major layout change from formalised document to a user-friendlier style
2.0.3	9 April 2015	Updated content on uncertainties, other datasets and refresh some images. Many cleanups.

# Table of Contents

1. In brief.....	4
What are these data and what are they intended for?.....	4
What are the key features of the v2.0 dataset compared to v1.0?.....	4
When are the next releases and what's in them?.....	5
What changes are needed to v1.0-compatible programs so they work with v2.0 data?.....	5
Where to get the data / how to get more?.....	5
How do these compare to other data sets?.....	5
Where can I get detailed information?.....	6
Where to get support?.....	6
2. Using the products.....	7
Applicability to different water types.....	7
Interpreting data values.....	7
Composite creation.....	7
Using the uncertainty estimates.....	8
3. Tools and sample programs.....	10
Sample programs in various languages.....	12
4. Known issues.....	14
Major errors.....	14
Data errors.....	14
Non-errors, but care required by users.....	14
Trivial issues.....	14
Informational only.....	14
Noteworthy changes from v1.0 format.....	15
5. The products: scientific overview.....	17
Chlorophyll-a concentration (mg m <sup>-3</sup> ).....	17
Remote Sensing Reflectance (sr <sup>-1</sup> ).....	19
Kd490: the attenuation coefficient for downwelling irradiance (m <sup>-1</sup> ).....	19
Inherent Optical Properties (IOP): total absorption and backscattering coefficients and their components (atot, aph, adg, bbp) (m <sup>-1</sup> ).....	19
Uncertainty characterisation.....	21
Optical water classes.....	21
The data-day approach.....	22
6. The products: technical overview.....	24
General format description.....	24
Filename convention.....	25
Example Filename.....	26
Grid format, map projection and coverage.....	26
File structure.....	28
Flags.....	32
Data sources (number of observations).....	33
High level metadata.....	34
7. How were the products made?.....	35
Stages of processing.....	35
8. Earlier versions.....	38

# 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 by climate change prediction and assessment models. OC-CCI aims to produce the highest quality data and thus may not contain the very latest data, as there are often concerns about calibration.

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

## ***What are the key features of the v2.0 dataset compared to v1.0?***

This data release has targeted specifically the consolidation of multiple improvements across the processing chain, particularly with regard to consistency of how the inputs are processed.

Easily noticeable changes are:

- Significantly reduced speckle effects from mapping SeaWiFS 4 km pixels (cf 1km pixels from MERIS and MODIS) and at swath edges
- Bias maps are now sensitive to seasonal variation rather than being a single (per-pixel) correction
- New water classes that much better represent the waters observed, as well as having a clearer relation to open-ocean/coastal waters
- Data up to the end of 2013 are included

More formally, version 2.0:

- refreshes the input datasets to the latest versions
- extends the time series to the end of 2013 ; 2014 is omitted based on NASA's concerns over MODIS (see section 4)
- greatly improves the in-situ database used for characterisation and quantification of error
- optimises the uncertainty generation for the CCI data. Specifically, the water classes are now based on the v2.0 data rather than on Tim Moore's SeaWiFS-based classes.
- improves consistency in many areas, including unifying the binning/mapping processing (correcting some pixelisation issues noted in v1.0)
- incorporates an improved bias correction, able to respond to temporal variation (primarily seasonal)
- incorporates an improved cloud mask (Idepix 2.0) for MERIS
- benefits from a more automated quality assurance process

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

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

- the initiation of cyclical processing, whereby a reprocessing and release to the user community is undertaken every year (2015, 2016, 2017);
- the 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.

While there will likely be intermediate test releases made, the next official release (v3.0) should be anticipated around February 2016 and is expected to address case 2 waters. The following version, v4.0, is planned for February 2017 and aims to include the first Sentinel-3 OLCI data.

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

A couple of small changes are necessary:

**Take account of a new time dimension on 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..

## **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 download page on the OC-CCI website for links to pages detailing these:

<http://www.esa-oceancolour-cci.org/>

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 to other data sets?***

For a comparison of v2.0 against v1.0, 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/>

### ***Where can I get detailed information?***

All project documentation and related publications can be found at the website:

<http://www.esa-oceancolour-cci.org/> (documents and publications links)

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](http://wiki.esipfed.org/index.php/Attribute_Convention_for_Data_Discovery))
- GlobColour Product User Guide (<http://globcolour.info/>)

### ***Where to get support?***

Feedback and questions regarding the use of the OC-CCI data are welcome – please email:

[help@esa-oceancolour-cci.org](mailto: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 only excluded waters of depth < 10m. Hence, the products are primarily designed for application in case-1 waters, but have some validity in case-2 waters depending on the applicability of the respective algorithms (OC4v6 for chl-a and QAA for IOP) for case-2 waters. 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.

### ***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<sup>3</sup>, and values greater than 100 have been set to 100 mg/m<sup>3</sup>
- Inherent Optical Properties: all values greater than 10 m<sup>-1</sup> have been discarded from the products
- Effect of high pathlength 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.
- R<sub>rs</sub>: all negative R<sub>rs</sub> 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. The uncertainties were first estimated for each of the optical water classes identified. The uncertainties are then assigned to each pixel, also on the basis of optical water classes: using OC-CCI remote-sensing reflectance spectra (R<sub>rs</sub>) for each pixel, the fuzzy membership of each optical

class in that pixel at that time was calculated, and the uncertainties for that pixel for each product were weighted averages of uncertainties, with the the fuzzy membership being 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), with the primary difference between it and v1.0 of OC-CCI being that, here, the optical classification is based on OC-CCI satellite Rrs data as inputs to the classification, instead of in-situ Rrs data. For further details regarding optical classification, please see section 5.

Given  $K$  optical classes, with the root-mean-square difference  $\Delta_k$  and bias  $\delta_k$  for each class, with  $k = 1 \dots K$ , and the membership of each of the  $K$  optical water classes in a particular pixel represented as the weighting function  $w_{k,p}$ , the root-mean-square difference of product  $x$  in that pixel ( $\Delta_p$ ) is given by:

$$\Delta_p = \sqrt{\frac{\sum_{k=1}^K w_{k,p} \Delta_k^2}{\sum_{k=1}^K w_{k,p}}}$$

The bias  $\delta_p$  can be computed similarly:

$$\delta_p = \frac{\sum_{k=1}^K w_{k,p} \delta_k}{\sum_{k=1}^K w_{k,p}}$$

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

$$\psi = \sqrt{(\Delta_p^2 - \delta_p^2)}$$

It is important to note that uncertainties in chlorophyll are provided for data that have been  $\log_{10}$  transformed. Let  $\sigma$  be the standard deviation of the log-transformed chlorophyll data, computed from the RMSD and bias that are provided. If  $\mu$  is the mean of the log-transformed data, and assuming that the data follow a log-normal distribution, the mean  $m$  of the non-log-transformed data (the expected value, equal to the satellite observation at that pixel) is related to  $\sigma$  and  $\mu$  as follows:

$$\mu = \log_{10}(m) - \frac{1}{2}\sigma^2$$

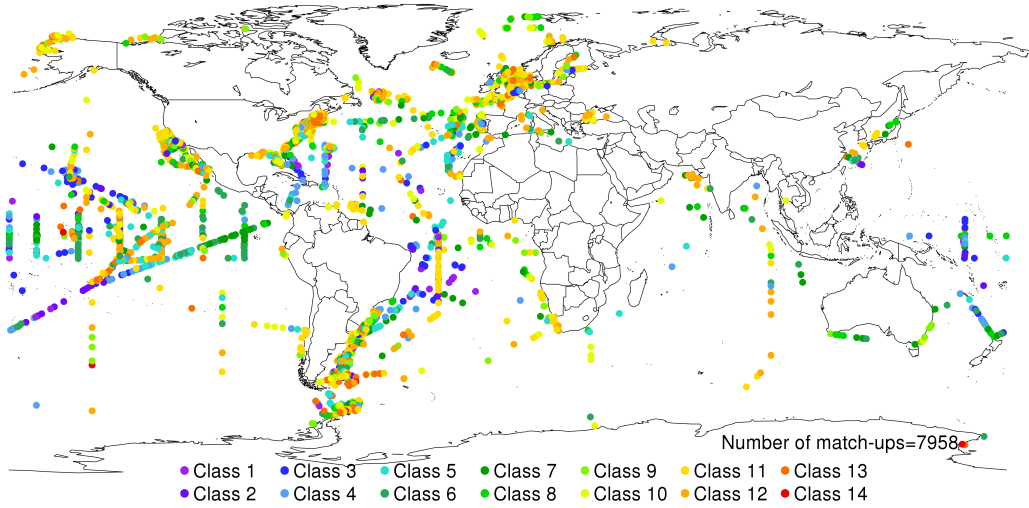
and the standard deviation  $\psi$  of the un-transformed data is related to  $\sigma$  and  $m$  according to:

$$\sigma = \frac{1}{\sqrt{\log_e(10)}} \sqrt{\log_{10}\left(1 + \frac{\psi}{m^2}\right)}$$

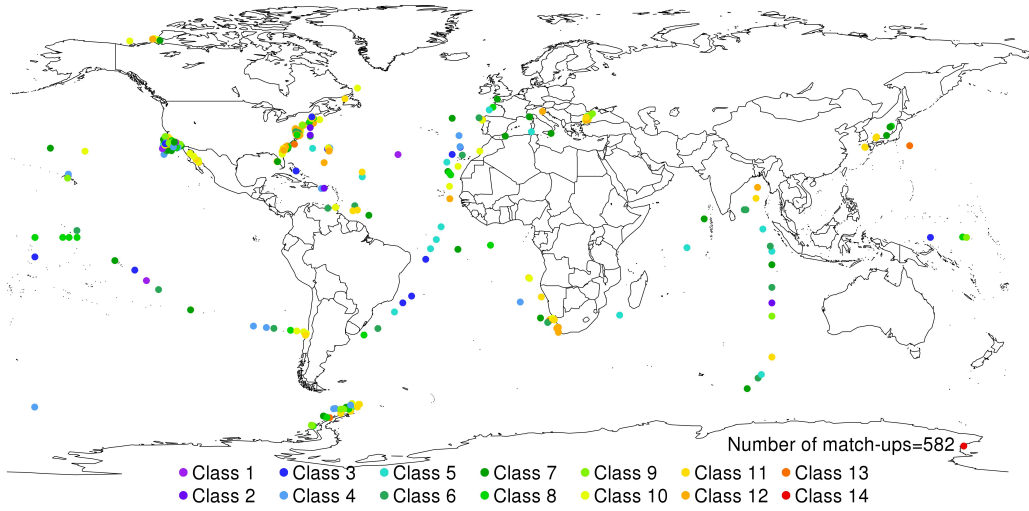
Note that these estimates are only as good as the quality and representativeness of the in-situ match up data sets that were available for uncertainty estimation. Geographical coverage and representation of different water types were best for chlorophyll, followed by  $K_d$  and then by  $R_{rs}$ . See Figure 1. Other known problems, such as those at high latitudes are not accounted for in this error budget.



### Chlorophyll match-up locations and water class types



### Kd(490) match-up locations and water class types



### Rrs match-up locations and water class types

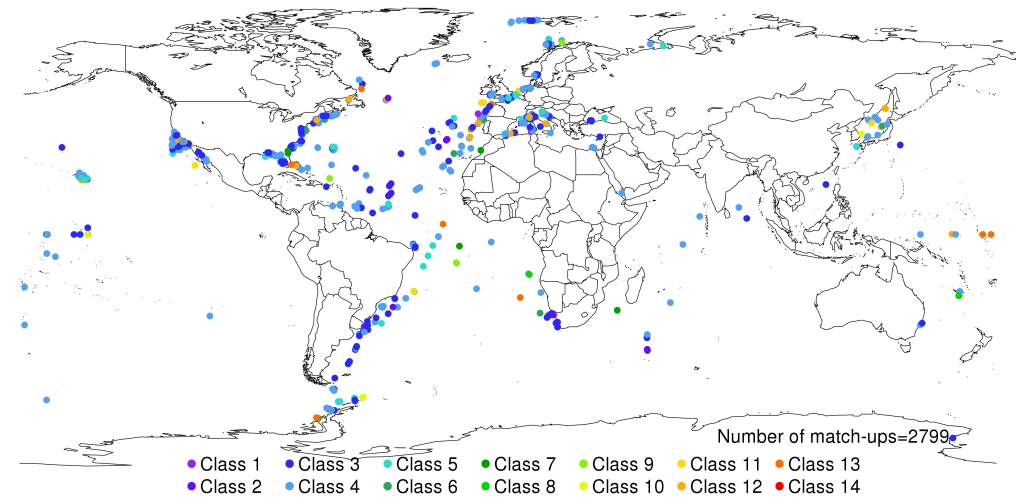


Figure 1: Geographical coverage of water types for a) Chl-a matchups; b) Kd490 ; c) Rrs

## **Creating composites of uncertainty variables**

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

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

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

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

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

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

### 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) in order to support transparent internal compression and read the products. Examples include the NetCDF operators, ncview, the Python netCDF4 library, R's netcdf package, etc.

The recommended package is the BEAM toolbox, which is specifically developed by ESA for the exploitation of Earth Observation data products. BEAM, for example, features the interpretation of flag-coding, provides image interpretation information, handles missing data gracefully and allows band arithmetic using a fast expression language.

BEAM is open source and freely available from <http://earth.esa.int/beam>

Regarding the OC-CCI products, BEAM could for example be used 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)

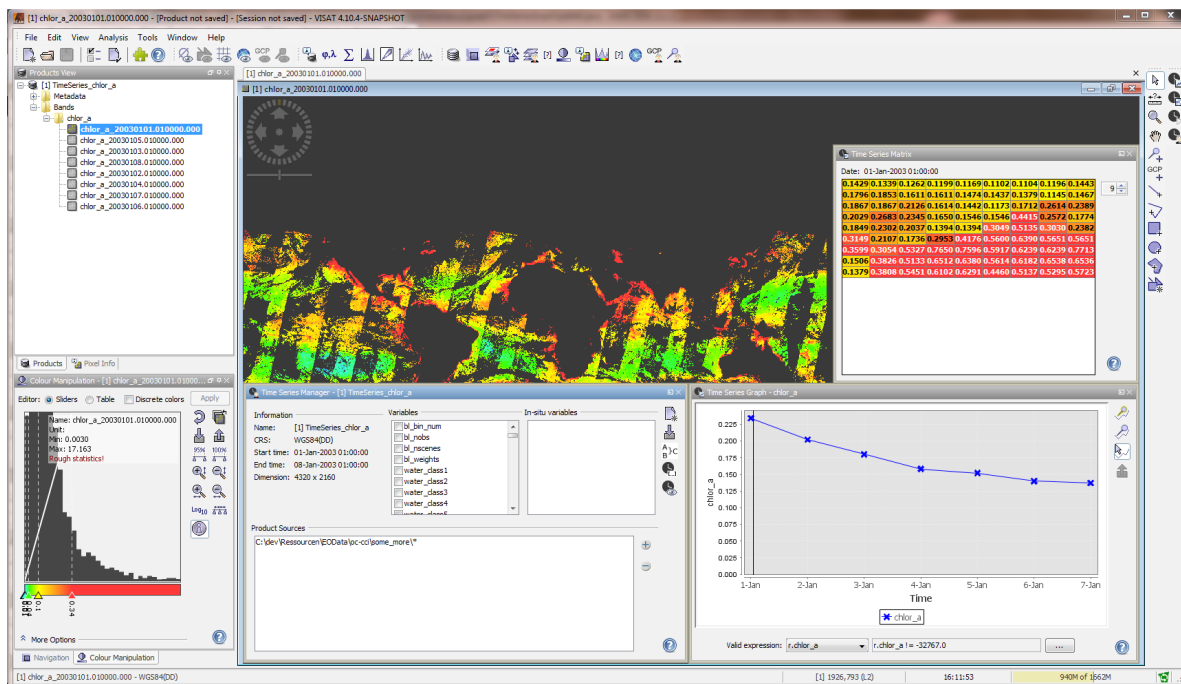


Figure 2 - Time series of chlor-a variable

At the time of writing (April 2015) 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 sinusoidal product format. After an update of the “Level-3 Binning Processor”, BEAM will be able to work with these products. Future version versions of BEAM will support this format by default. This can be done as follows:

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 alternative for working with the OC-CCI products is the SeaDAS Visualization tool. The latest version (SeaDAS 7.0.1) is the result of collaboration with the developers of ESA's BEAM software package. The core visualization package for SeaDAS 7 is based on the BEAM framework, with extensions that provide the functionality provided by previous versions of SeaDAS. To work with the OC-CCI products please update the "Level-3 Binning Processor" module in the same way as described for BEAM.

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 one-dimensional geolocation.

## ***Sample programs in various languages***

Please see the website for more examples.

### **Python**

There are a number of 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 a number of 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 dataset 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 this may include noteworthy pixels such as coccolithophore blooms.

**Decay of MODIS calibration:** NASA monitor the calibration of MODIS and regularly adjust it. The last processing was r2013.1, completed around the beginning of 2014, and intended to correct problems noted since August 2013. Since then, decay of the sensor has continued, with increasing problems at the blue end. NASA has noted these and have been working on a further reprocessing (nominally r2014.0, though 2014 is already past). This has not yet been completed, and CCI data from 2014, which rely solely on MODIS, are considered to be at high risk of errors and have been omitted from the public release.

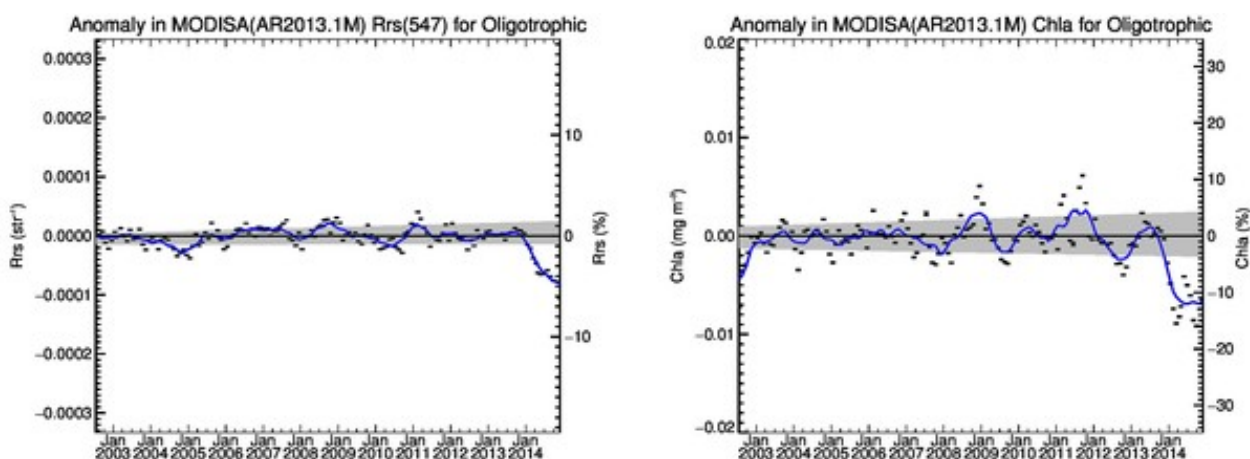


Figure 3: Rrs547 and chlorophyll anomalies from a NASA presentation at IOCCG (Mar '15). Accompanying text: "In 2014, all radiometry shifted (up in blue, down in green), with commensurate decline 10% in clear-water chlorophyll. This is clearly a calibration error!"

## ***Trivial issues***

**IOP standard\_name attributes** in the NetCDFs are insufficiently descriptive (no distinction between  $a_{ph}$ ,  $a_{dg}$ , etc). This is because improved names have not yet been accepted into the CF standard name list, despite apparent consensus having been reached over 1.5 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

**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 and MODIS from July 2002 and on-going at the time of writing).

**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  $bbp$ , 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 were 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.

**“Moire effects” / “pepper noise” / black speckles:** small black speckles organised in a Moire pattern may be seen variously throughout the dataset, particularly in the 1997-2002 period. These are due to the lower resolution of the SeaWiFS GAC data being undersampled by the geographic grid (the effect is also present to a more limited degree in the sinusoidal data, but this is commonly reprojected to geographic when inspecting it). While the SeaWiFS GAC data are nominally 4 km, this is not true further from nadir and, at the edges of a swath, it may be significantly more. Consequently, when these data are sampled onto a 4 km grid using a nearest-neighbour binning algorithm, there may be pixels where no data are available. The pattern of black dots falls naturally into a Moire effect due to the shape of the Earth and the projection method. The same effect can be seen in NASA SeaWiFS outputs at 4 km and is the reason they mainly sample at 9 km. When the higher-resolution MERIS and MODIS join the time series in 2002, the effect largely disappears except where only SeaWiFS data are present.

### **Noteworthy changes from v1.0 format**

**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.



## 5. The products: scientific overview

After a comparison with the v1.0 data, the following sections provide 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.

### OC-CCI v2.0 and OC-CCI v1.0

The OC-CCI data and each of its subsets are comprised of 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.

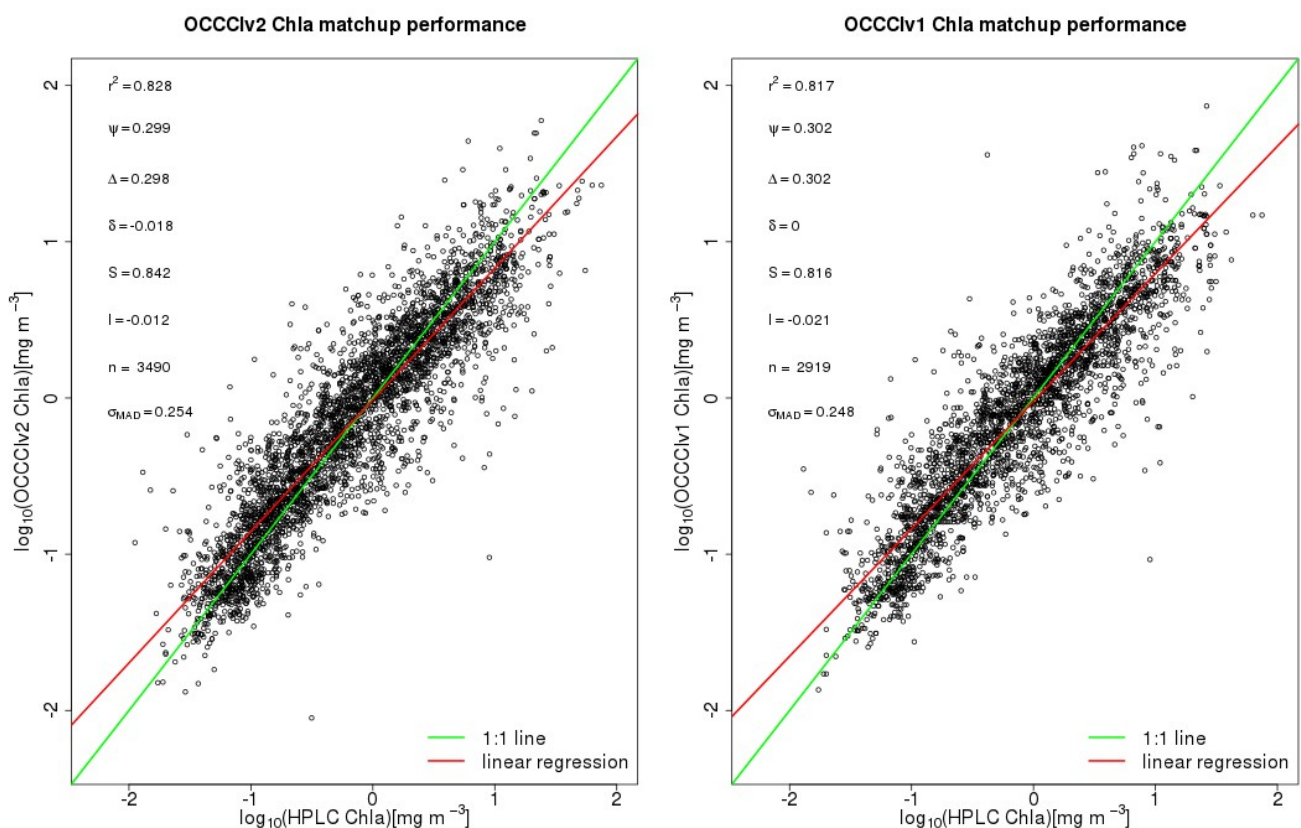


Figure 4: Comparison of v1.0 and v2.0 of the OC-CCI chl product when matched against in-situ HPLC chl-a measurements. Summary statistics shown are correlation coefficient, root-mean-square-difference (RMSD), un-biased RMSD, bias, slope of regression, intercept of regression, number of match-ups and the median-absolute-deviation of residuals.

It should be noted that the change in the performance statistics for a product, such as chlorophyll, between v1.0 and v2.0 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 v2.0 chlorophyll product has a greater number of match-ups, better correlation and smaller RMSD than v1.0. The slope and intercept of the regression have also improved.

To compare the uncertainty variables between differing OC-CCI datasets one can look at the large scale (global) range and distribution of uncertainty values. Figures 5 and 6 show an example of the differences in uncertainty between the two version of OC-CCI data. It should also be noted that although the uncertainties in some regions have become greater in magnitude, the confidence in the uncertainty values given is greater. This is due to the increased performance of the water classification (the gyres, for example, were very poorly classified in OC-CCI v1.0) as shown in Figure 7.

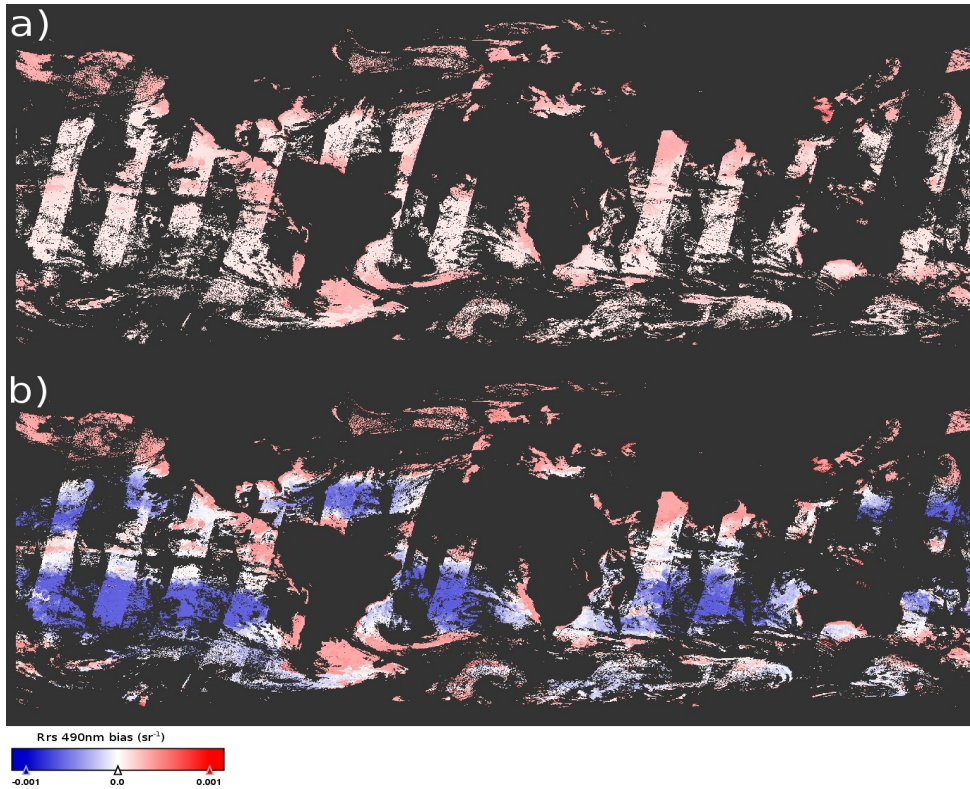


Figure 5: Comparison of the Rrs 490 bias uncertainty as given in v1.0 (a) and v2.0 (b).

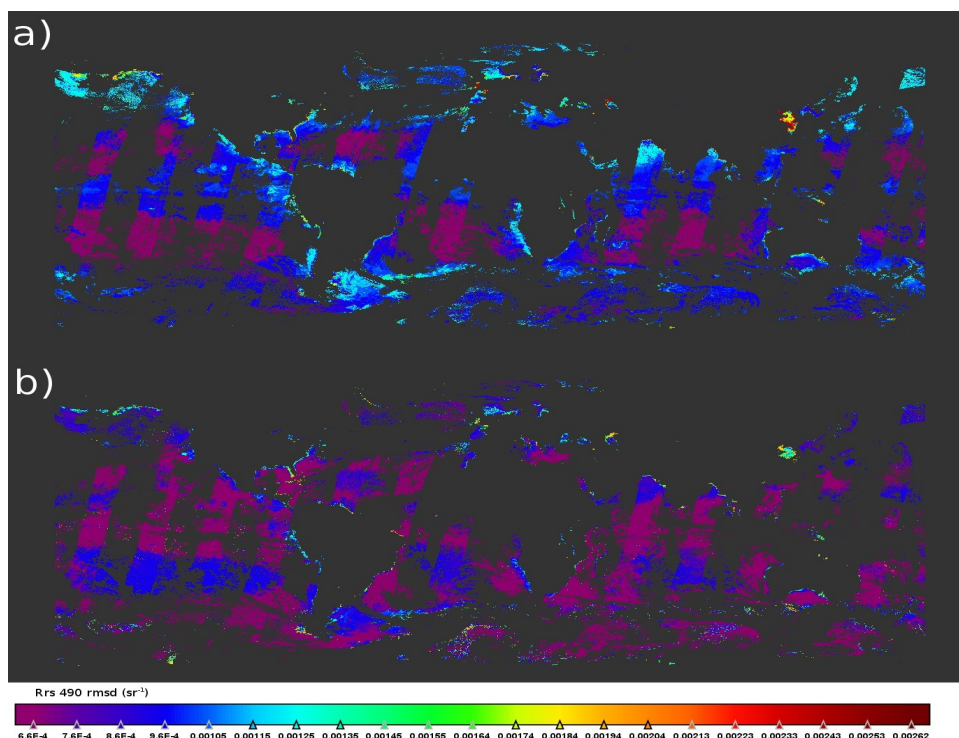
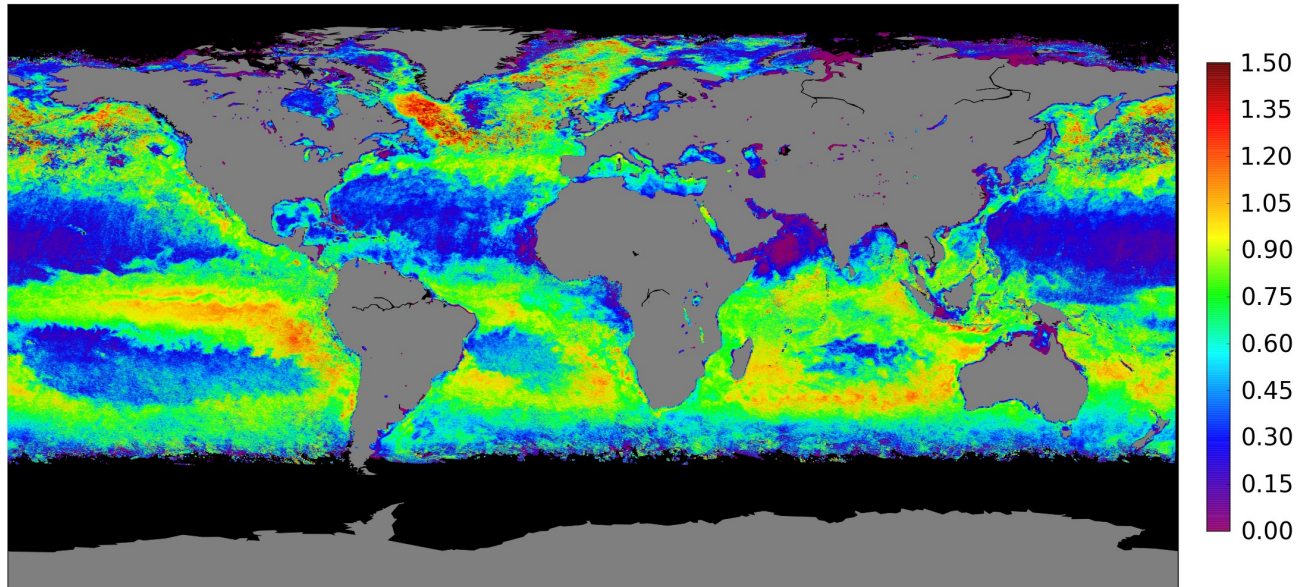
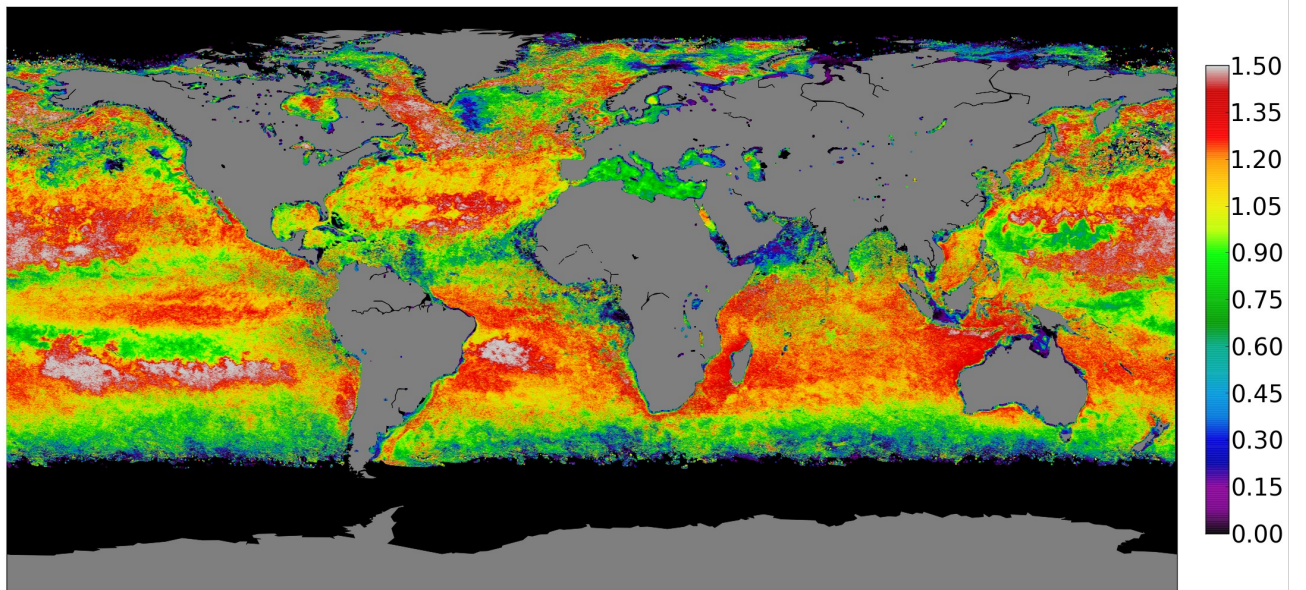


Figure 6: Comparison of the Rrs 490 rmsd uncertainty as given in v1.0 (a) and v2.0 (b).

*Cumulative optical class membership, July 2003 v1.0*



*Cumulative optical class membership, July 2003 v2.0*



*Figure 7: Improvement in the total optical class membership for pixels in OC-CCI v2.0 compared to OC-CCI v1.0.*

Overall the OC-CCI v2.0 data has an increased spatial coverage of data, increased performance at retrieving chlorophyll concentrations and a more representative set of uncertainties than the v1.0 dataset.

## ***Chlorophyll-a concentration ( $\text{mg m}^{-3}$ )***

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  $\text{mg m}^{-3}$ , and is provided as daily products with a horizontal resolution of  $\sim 4$  km/pixel. Furthermore, the root-mean-square (RMS) uncertainty and the bias in the  $\log_{10}$  chlorophyll-a concentration are provided, based on comparison with match-up in-situ data. The values are calculated applying the NASA OC4.V6 algorithm (NASA 2010) on the OC-CCI merged  $R_{rs}$  products described below.

Figures 8-10 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  $\log_{10}$  values due to the underlying natural distribution.

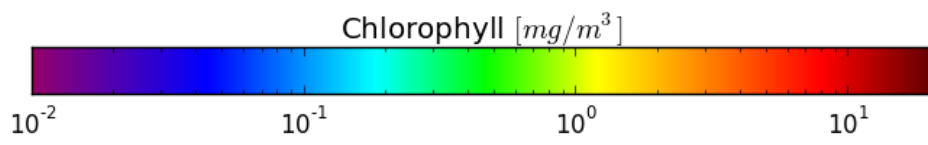
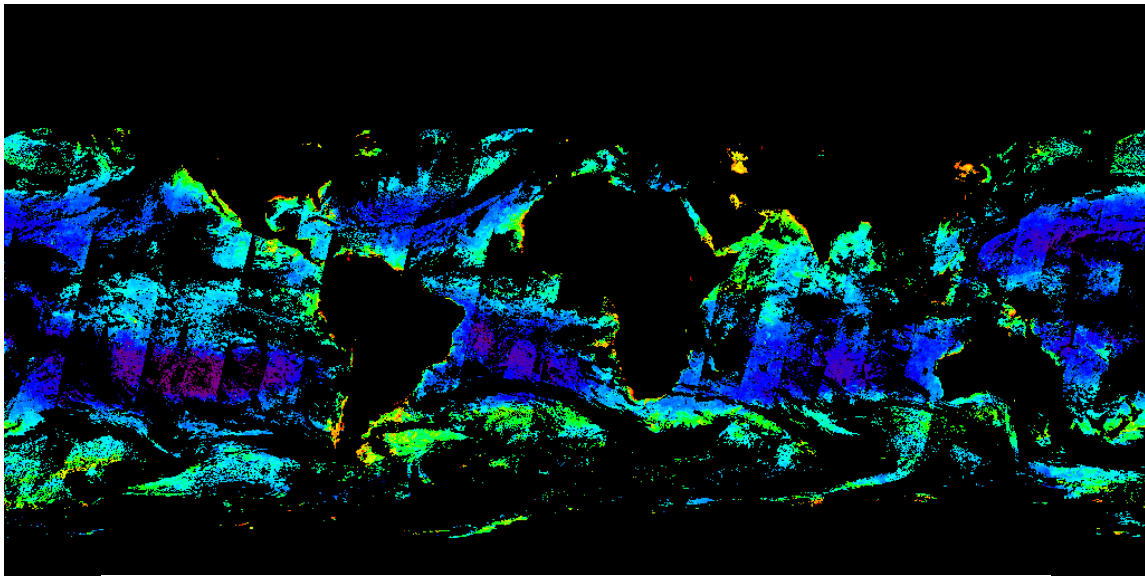


Figure 8: Chlorophyll-a concentration (1<sup>st</sup> Jan 2003)

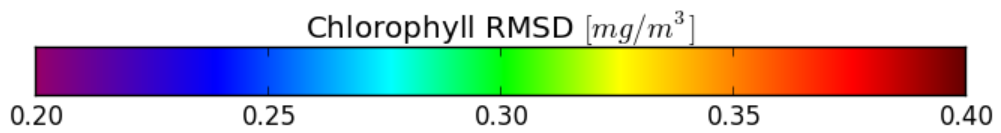
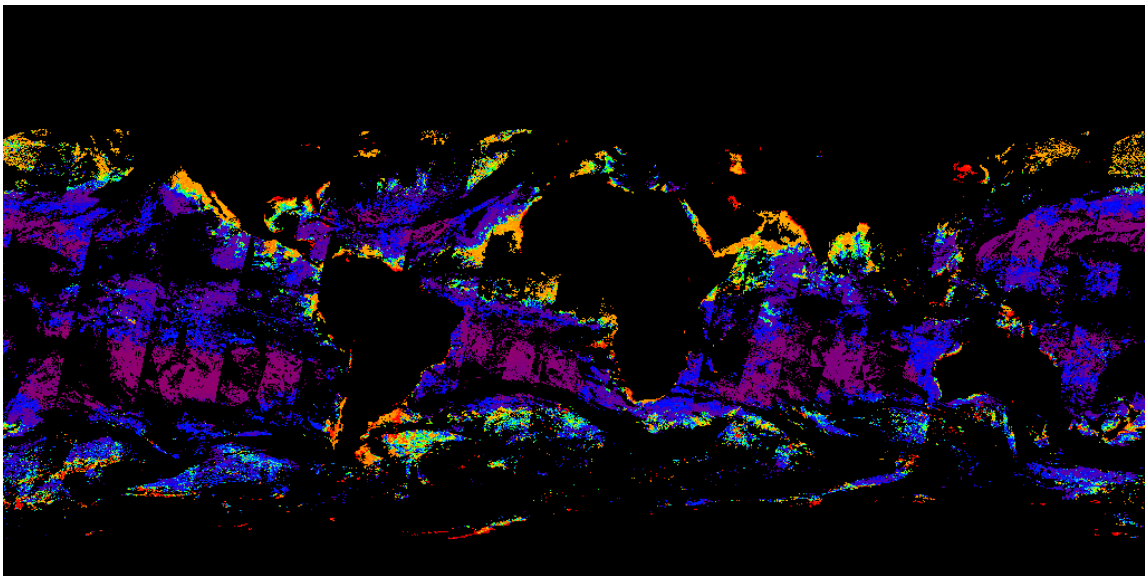


Figure 9: Root mean square difference of chlorophyll-a concentration (1<sup>st</sup> Jan 2003)

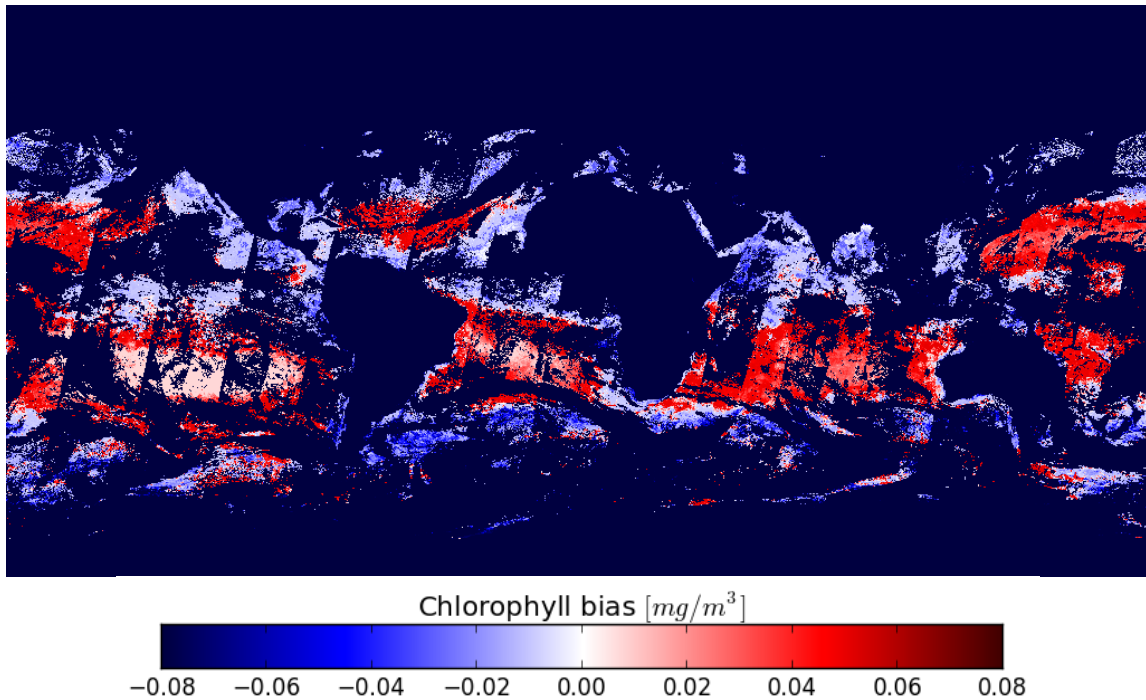


Figure 10: Bias of chlorophyll-a concentration (1<sup>st</sup> Jan 2003)

### **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  $\sim 4$  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 and Aqua-MODIS data. Atmospheric correction was carried out using the POLYMER algorithm for MERIS and SeaDAS v7 processor for MODIS and SeaWiFS (see the Polymer Algorithm Theoretical Baseline Document). The  $R_{rs}$  values from MERIS and MODIS were band-shifted to SeaWiFS wavebands if necessary, and corrected for inter-sensor bias when compared with SeaWiFS.

### **$Kd_{490}$ : the attenuation coefficient for downwelling irradiance ( $m^{-1}$ )**

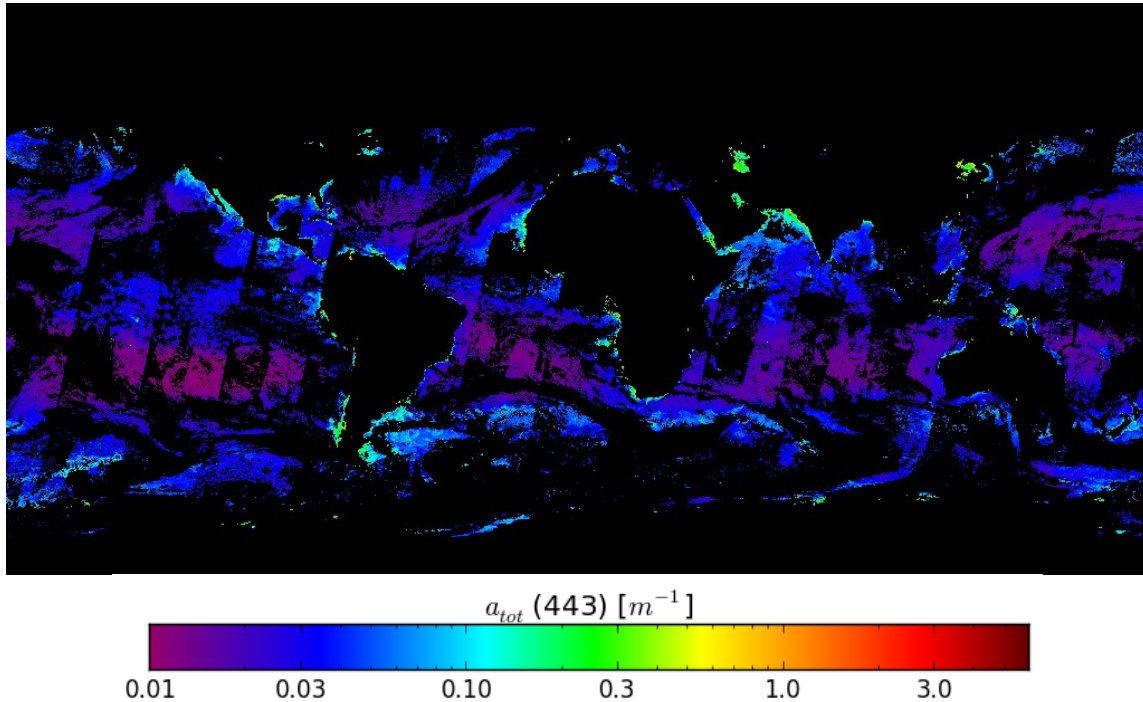
The attenuation coefficient at 490nm for downwelling irradiance, which is an apparent optical property, is part of the OC-CCI products. It is provided at daily resolution and spatial resolution of  $\sim 4$  km/pixel. It is computed from the inherent optical properties (see below) at 490 nm and the sun-zenith 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  $\sim 4$  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 11 to 13 show global daily images of total absorption, absorption of detrital and dissolved matter, and absorption by phytoplankton at 443 nm.



*Figure 11: Total absorption at 443 nm (1<sup>st</sup> Jan 2003)*

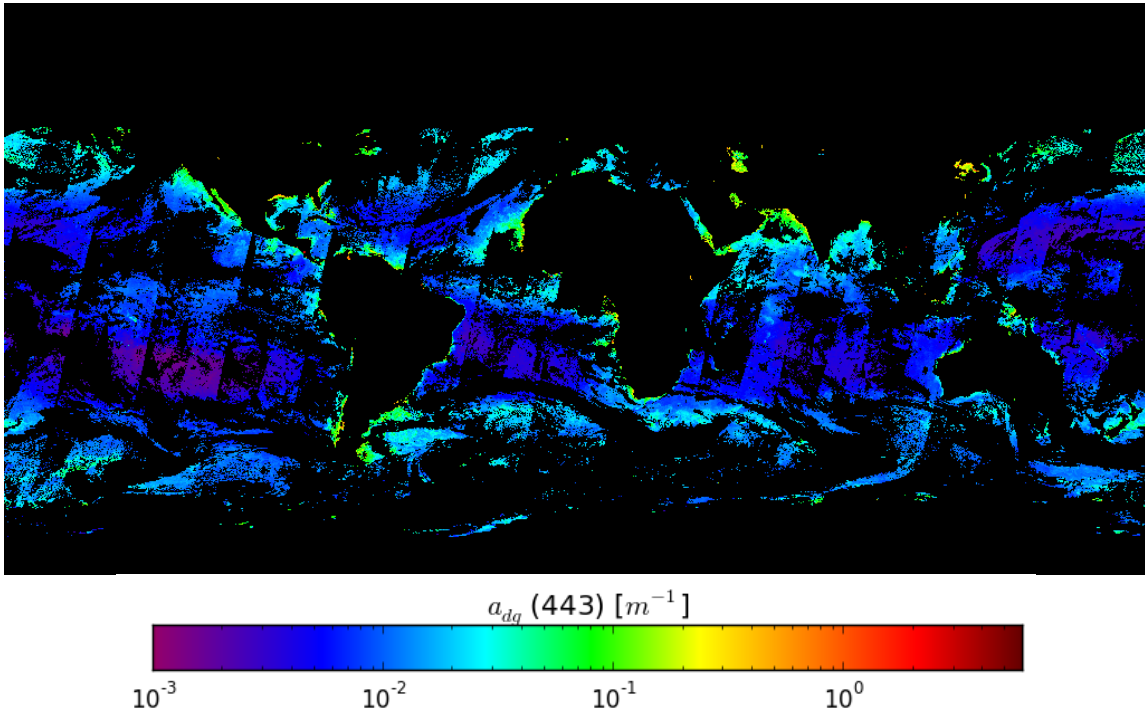


Figure 12: Absorption by detrital and dissolved matter at 443 nm (1<sup>st</sup> Jan 2003)

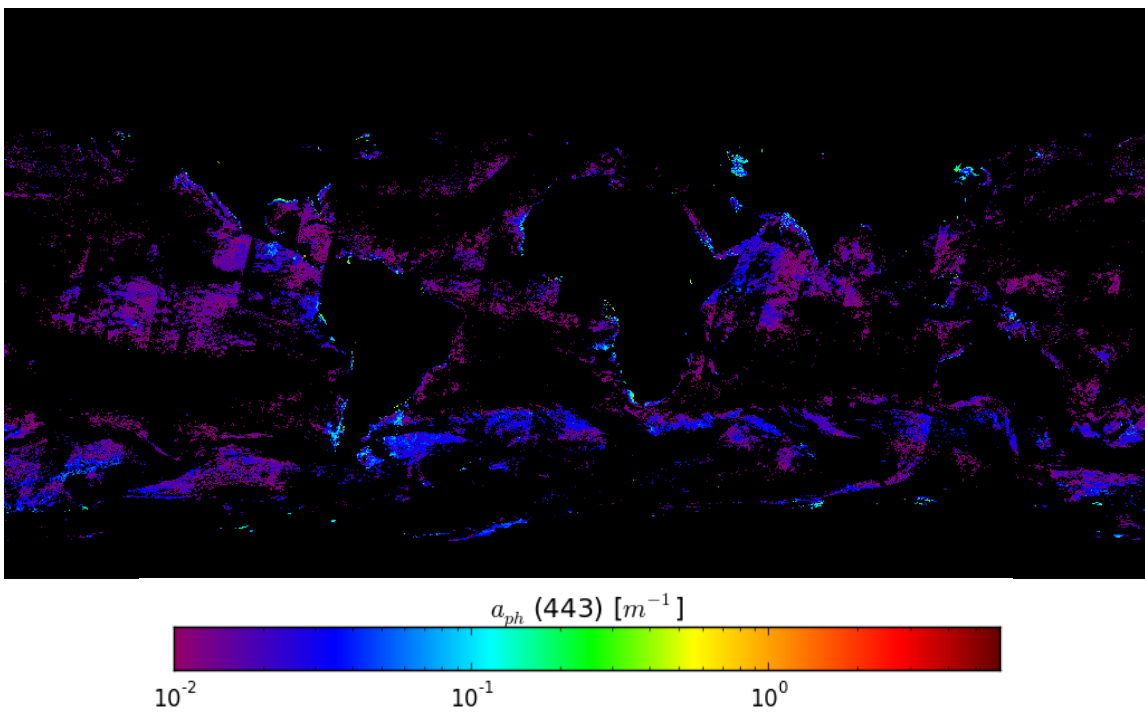


Figure 13: Phytoplankton absorption at 443 nm (1<sup>st</sup> Jan 2003)



## 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, due to the underlying natural distribution being 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, 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 14 shows the spectral shapes of the final classes:

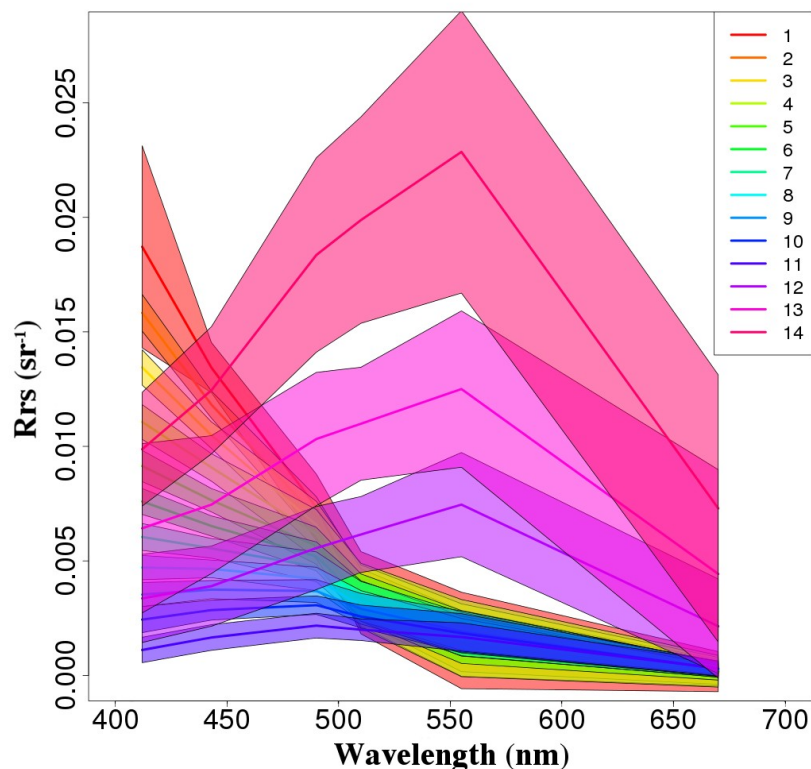


Figure 14: Spectral response of the water types used in OC-CCI V2 products (hard lines are class means and shaded region shows standard deviation).

## 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
- $\tau$  (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
- $\phi$  (deg): longitude of the measured pixel

Note:  $\tau$  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

## 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 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 v2.0 data release, a typical netCDF file containing the full set of products for a single day is approximately 1.6GB. 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 GHRSSST compatibility; always "2.0" for the v2.0 data

## Example filename

An example filename is:

```
ESACCI-OC-L3S-OC_PRODUCTS-MERGED-1D_DAILY_4_km_GEO_PML_OC4v6_QAA-20031225-fv2.0.nc
```

With components being:

<b>Filename component and alternates</b>	<b>Description</b>
<i>ESACCI-OC</i>	Fixed prefix
<i>L3S</i>	Processing Level (fixed)
<i>OC_PRODUCTS</i>	Data Type string indicating all products in one file
<i>CHLOR_A</i>	chlorophyll-related product subset
<i>RRS</i>	Rrs and water class product subset
<i>IOP</i>	IOP product subset
<i>K_490</i>	Kd490 product subset
<i>MERGED</i>	Data is from more than one sensor (fixed, though may be used in future releases of individual sensors)
<i>ID</i>	Additional Segregator Element: Composite data (1 day, may be other variants here)
<i>DAILY</i>	Additional Segregator Element: Length of time period covered
<i>4 km</i>	Additional Segregator Element: Spatial Resolution
<i>GEO</i>	Additional Segregator Element: Projection type (Geographic or Sinusoidal)
<i>SIN</i>	
<i>PML</i>	Additional Segregator Element: Processing Centre (fixed)
<i>OC4v6_QAA</i>	Additional Segregator Element: Algorithm(s) (varies)
<i>20030907</i>	Indicative Date

## **Grid format, map projection and coverage**

The products are available in two projections: sinusoidal and geographic (also known as equidistant cylindrical, equiarectangular, 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.

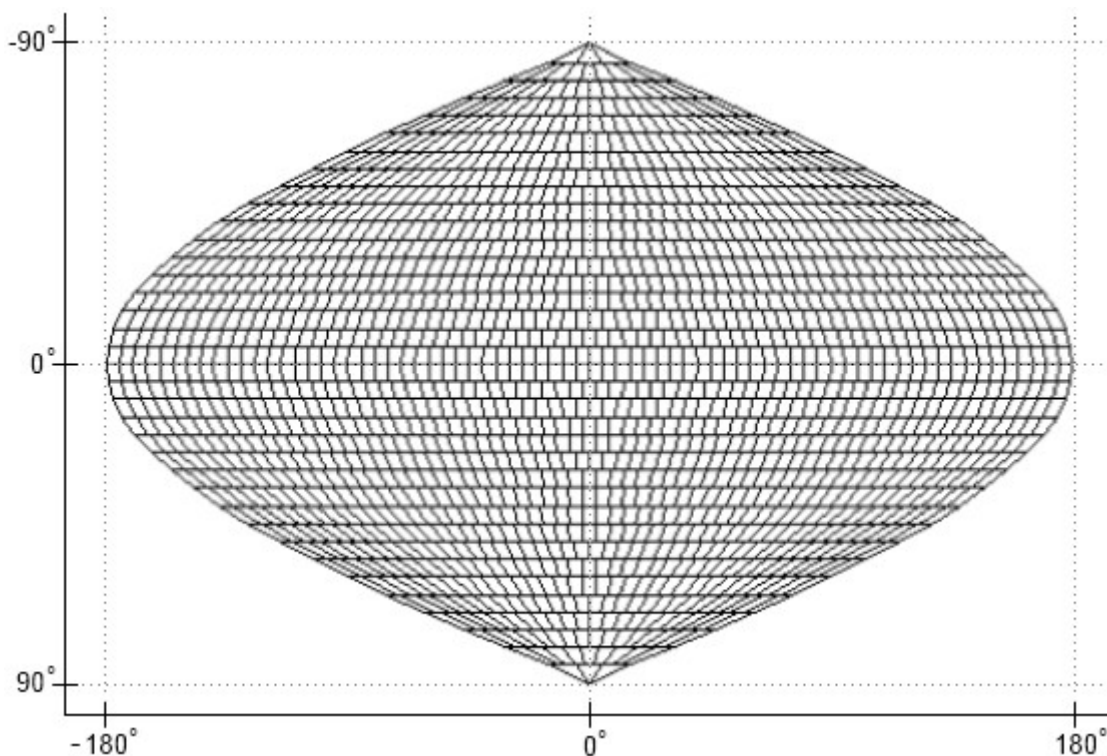
All files contain CF-compliant latitude and longitude (and time) dimensions, allowing each data cell to be specifically associated with a location. All latitudes and longitudes are given in WGS/84 datum.

## 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 15: 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, bin_index) ;  
    chlor_a:_FillValue = 9.96921e+36f ;  
    chlor_a:standard_name =  
"mass_concentration_of_chlorophyll_a_in_sea_water" ;
```

```

    chlor_a:parameter_vocab_uri =
"http://vocab.ndg.nerc.ac.uk/term/P011/current/CHLTVOLU" ;

    chlor_a:long_name = "Chlorophyll-a concentration in seawater,
generated by SeaDAS using OC4v6 for SeaWiFS" ;

    chlor_a:ancillary_variables =
"chlor_a_log10_rmsd,chlor_a_log10_bias" ;

    chlor_a:units = "milligram m-3" ;

    chlor_a:units_nonstandard = "mg m^-3" ;

    chlor_a:grid_mapping = "crs" ;

    chlor_a:coordinates = "lat lon" ;

```

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.

Other variables are:

<b>Data variable</b>	<b>Accompanying uncertainty variables</b>	<b>Notes</b>
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	
atot_412	<i>Not computed separately, as this is a convenience variable</i>	QAA total absorption ( $a_{ph}+a_{dg}+a_w$ , though QAA’s decomposition method sometimes does not preserve this property)
atot_443		
atot_490		
atot_510		
atot_555		
atot_670		
chlor_a	chlor_a_log10_rmsd	Chlorophyll-a, estimated using the OC4v6 algorithm
	chlor_a_log10_bias	
aph_412	aph_412_rmsd	QAA absorption due to phytoplankton
aph_443	aph_443_rmsd	
aph_490	aph_490_rmsd	



aph_510	aph_510_rmsd	
aph_555	aph_555_rmsd	
aph_670	aph_670_rmsd	
	aph_412_bias	
	aph_443_bias	
	aph_490_bias	
	aph_510_bias	
	aph_555_bias	
	aph_670_bias	
adg_412	adg_412_rmsd	QAA absorption due to detrital and dissolved matter
adg_443	adg_443_rmsd	
adg_490	adg_490_rmsd	
adg_510	adg_510_rmsd	
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	<i>Insufficient in-situ data to make a plausible estimate</i>	QAA backscatter due to particulate matter
bbp_443		
bbp_490		
bbp_510		
bbp_555		
bbp_670		
kd_490	kd_490_rmsd kd_490_bias	Attenuation coefficient (Lee algorithm with Zhang backscatter coefficients)
water_class1	<i>n/a</i>	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_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		

## 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 GAC and thus at 4 km resolution while the MERIS and MODIS 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 variables are:

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

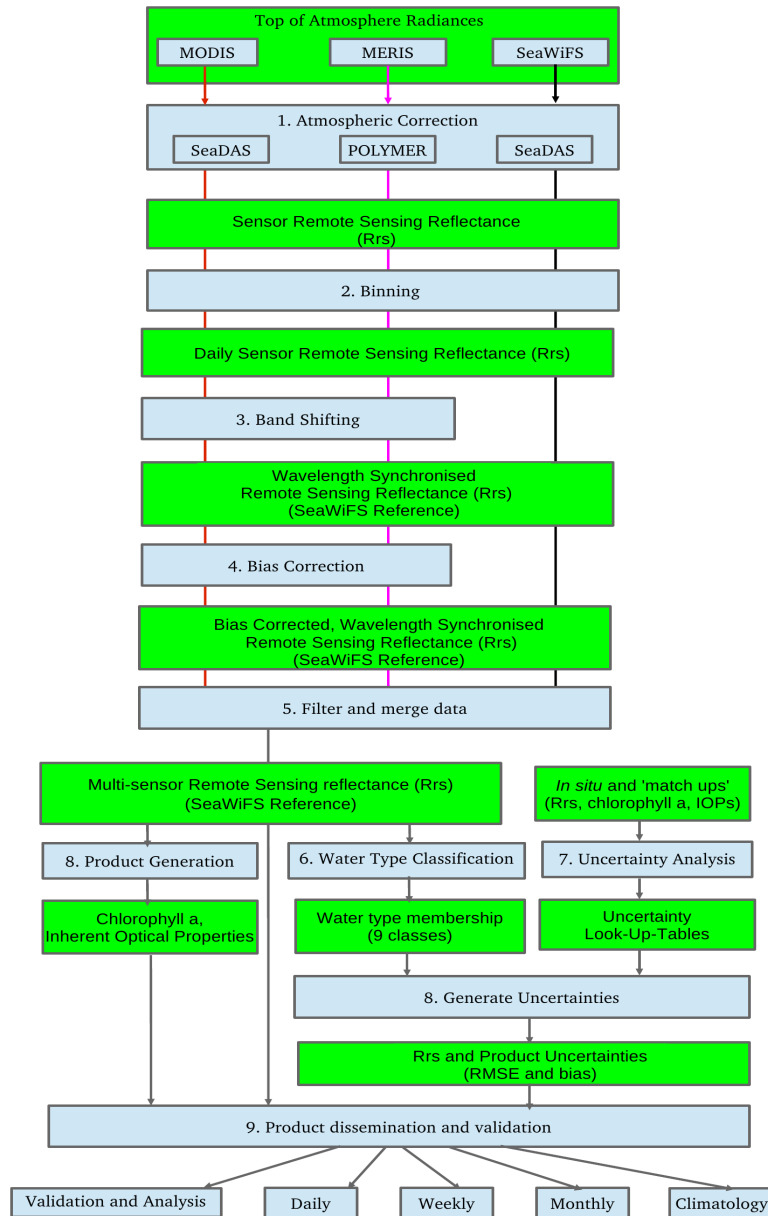


Figure 16: Data flow in the Ocean Colour ECV production

### Stages of processing

A very brief overview of the major CCI processing stages is given here; for more detail, please refer to the SPS.

## **Input datasets**

The input EO datasets were MERIS Reduced-Resolution (1km) L1b 3<sup>rd</sup> reprocessing (including OCL fixes), MODIS R2013.1 level 3 binned (4 km) on a sinusoidal grid, SeaWiFS level 1a GAC (4 km) R2010.0.

## **Level 2 processing and binning**

MERIS was processed with the POLYMER algorithm (v3.0) to level 2. SeaWiFS and MODIS L2 were downloaded from NASA..

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, while MODIS and SeaWiFS were masked using the standard NASA flags.

All available data were used, up the end of 2014. It should be noted that NASA considers MODIS' calibration in 2014 to be degrading, therefore data after 2013 has been omitted from the public release.

## **Band shifting**

MODIS and MERIS 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, 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.

## **Merging**

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

## **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: OC4v6
- 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**

Per-pixel uncertainty estimates were computed following Moore et al (2009), but with (14) specific water classes derived from the v2.0 CCI Rrs vales. 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). A table of uncertainties for each class were computed from matchups between the CCI in-situ database and the version 2.0 data. Every individual pixel in a scene to compute its membership percentage for each of these derived classes, 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.

## **8. Earlier versions**

This annex briefly summarises some of the previous non-public OC-CCI data releases to put in context the high level changes for users who received preliminary data sets. 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 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 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.