Validation of satellite data with WIM/WAM

Contents

1 Introduction

Note: a related document that is more suitable to be used as a guide through practical exercises is available at http://www.wimsoft.com/Course/4/Validation_2015_presentation.pdf.

Note: the *.csv* files used are in the *English (US)* format. Corresponding files with semicolon (*_semicolon.csv*) or tab (*_tab.csv*) as list separator must be used if using a different region setting, e.g. *French (France)*. Also, the arguments including decimal point (e.g. 1.5) must be with comma.

Satellite measurements are typically very indirect measurements of our actual target variable. For example, when we want to measure ocean surface chlorophyll concentration (Chla), we actually measure radiance at the top of the atmosphere. It is well known that radiance at the top of the atmosphere is primarily made up by the photons backscattered from the atmosphere with only a small portion of photons (up to 5-10%) originating from the ocean. Removal of the effects of the atmosphere, particularly of the variable part due to aerosols, is error-prone and therefore estimates of water properties are inherently subject to both systematic and random errors. Small errors in calibration and atmospheric correction can easily propagate into large errors in estimated in-water properties. Moreover, a variable like primary production (PP, mg C m^{-2} d^{-1}), estimated from satellite data, is even more indirect measurement as it is typically derived from remotely

estimated Chla a set of and other variables. Adding to the complexity: PP is (1) a derived variable of derived variables (e.g. of Chla, PAR, etc.); (2) PP is a rate (mg C m^{-2} d^{-1}) and not a concentration; and (3) PP depends on simplified models that parameterize various complex relationships (e.g. of the vertical distribution of Chla, of the assimilation number, etc). In spite of these potentially severe errors we need these indirect satellite measurements as they are our primary source for large-scale estimates of variables that are difficult to estimate otherwise. However, it is essential to recognize that satellite estimates may have large errors and may need corrections. Validation is a primary step in the process of making sure that satellite data are indeed a fair representation of the true values. Validation means comparison with some kind of "truth" or "ground truth" data. Validation can be of different types:

- against *in situ* measurements;
- against model or theoretical estimates;
- between different satellite or aircraft sensors.

In many cases we don't have a good "truth" or "ground truth" as measurements are not available, have been made at different time or space scales or have significant errors of their own. Comparison with *in situ* point measurements is often called a *match-up* analysis. Keep in mind that validation is not the same as calibration (and sensor characterization) which apply to the optical characteristics of the sensor or measurement complex.

1.1 Prerequisites

We will be using the WIM/WAM software package (http://wimsoft.com) but similar concepts and tools can be applied using other software. We assume that you have installed WIM/WAM and are familiar with the basics of Windows command prompt. Basic knowledge of WIM is useful but not essential beyond basic opening of image files. To load images from HDF files just click (double-click) on the file name and select WIM as the default application to open these files. You can check out the WIM and WAM manuals, particularly Exercises with WIM/WAM. We assume that you have copied the ***Sat*** folder with satellite data from the DVD or USB stick to the root of your C drive, e.g. ***C:\Sat***. You can use another drive instead of "C" and then you need to remember to replace the "C" with your drive letter.

1.2 Sample in situ data

Let's assume that we have a dataset from a cruise – we will use actual data collected by the CCE-LTER cruise P0810 in October, 2008 in the California Current. We have selected tabular data in a simple text file *C:\Sat\P0810.csv*. This file is in the "comma separated values" or CSV format (*.csv) and corresponds to the English (US) usage of dates and decimal numbers. In many other regions comma is used in decimal numbers and cannot be used as a separator or values. Also, the standard date format is different depending on the region. We have therefore provided copies of the same data in other versions: an Excel file *C:\Sat\P0810.xlsx*, a tab delimited *P0810_tab.csv*, a semicolon delimited *P0810_semicolon.csv*. The last two have comma instead of the decimal point in numbers (e.g. like used in France). Excel file usually takes care of the conversions between different formats of date/time and numbers. Please open one of these spreadsheets and examine its column structure:

Lon_Dec, Lat_Dec, Date, Time, Cruise, Sta_ID, Depth, Temp, Sal, Density, Chl, PO4, SiO4, NO2, NO3, IntChl, NCDepth, MLD

Note that *Longitude* (in decimal degrees) is first, followed by *Latitude* (in decimal degrees). Date and Time are both in GMT (UTC) and Date is in the US format MM/DD/YYYY. After the essential 4 columns (*Lon, Lat, Date, Time*) you may have unlimited additional columns, such

Cruise , *Station* and the measured variable columns. There is a *Depth* column but data have been selected for the near-surface depth only. Columns for the measured *Chl* (fluorometric Chla), nutrients, nutricline depth (*NCDepth*) and mixed layer depth (*MLD*) are included. Our task is to find matching satellite data to these 66 stations. WAM programs should be able to use your regional settings. If there is a problem then you can temporarily change your formats to *English (United States)*. In Win7 you can select *Control Panel – Region and Language*, change your "*Current location*", "*Format:"* to "*English (United States)*". In Win8 you can select *Control Panel – Region*, and then "*English (United States)*".

While in situ measurements are considered "ground truth", for ocean color sensors they are always at a wrong spatial scale, e.g. a sample at <1 m spatial resolution is compared to a typical pixel at ~1 km resolution (or even larger for Level-3 data). Spatial distributions at ~1 km and larger scales cannot be considered constant and fully characterized by a single or a few in situ samples.

2 Using WIM interactively

2.1 Right-click

A simple but most primitive and time consuming way of a match-up analysis is to load an image and interactively looking up pixel coordinates and values. In WIM we can use the right click of the mouse to look up pixel values. For example, load the image from *C:\Sat\MODISA\L2\CAL\A2008_cruise_chl.hdf* (by double-clicking on the filename in *Explorer*) and try to match the latitude/longitude of a station in the *P0810.csv* spreadsheet to a pixel (point) in the image. As you can see, it is extremely time consuming and error-prone.

2.2 Using Geo-Get Vector Objects

A little more advanced method for finding match-ups is to use the *Geo-Get Vector Objects* menu option in WIM.

- Load the Chla image *A2008_cruise_chl.hdf* in *C:\Sat\MODISA\L2\CAL*. Use menu option *Geo-Get Vector Objects-Point* (*Bitmap Only, Geographic Lon, Lat, Float Lon Lat*) to load the station data from *C:\Sat\P0810.csv*. If you don't get any match-ups then there is something wrong. For example, WIM may assume that you have *Longitude* first and *Latitude* second. Use the *Settings* (the  icon on the Toolbar)-*Misc* and uncheck the *Lat first* checkbox.
- You should get a something like Fig.1 with small colored circles showing the stations on the image.

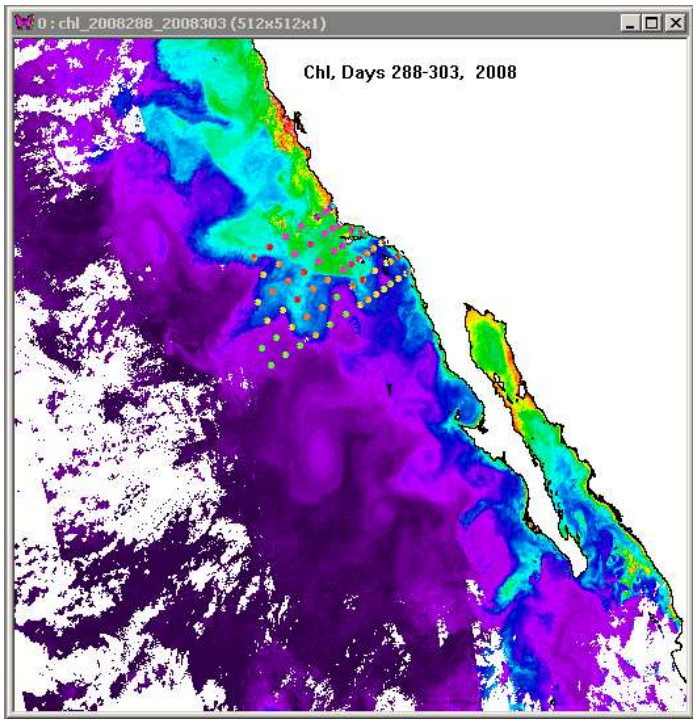


Fig. 1. Overlaying stations on an image with *Geo-Get Vector Objects*.

- Select a point in the *Vector objects* table and see which one is blinking. Select *Statistics* for the selected point. You need to make sure that the valid range is properly specified. For a scaled byte image with *Log-Chl* scaling the proper range is Min = 0.011, Max = 64. The statistics is calculated for a 3 x 3 pixel window centered at the point. Note the values of ***N (in-range*** = number of pixels in the valid range) and ***N (out-range*** = number of pixels **out** of the valid range). As we are using a monthly composite, the image is fairly good and has few missing pixels due to clouds.

  - You can select one or more (or all) the points and save all the statistics into a new file, e.g. *test.csv* with the *Save Lat,Lon,Value* button (**not** the *Save* button that saves it in a HDF format). Load the saved *test.csv* file into a text editor, e.g. *Notepad*, or MS *Excel* and view the file format. The header line has column names:

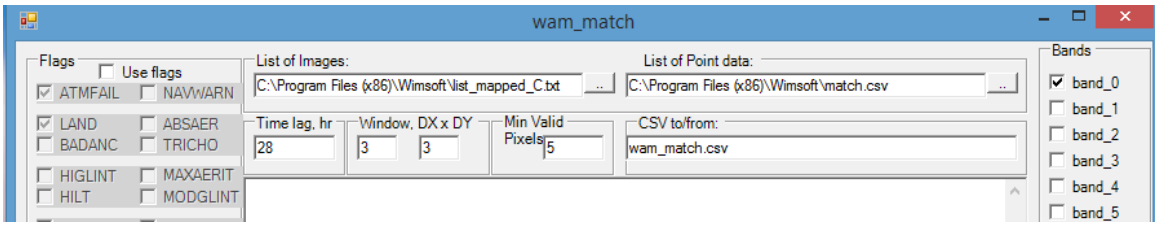
  Lon Lat Value N_Valid N_Invalid Min Max Mean Median

  Reading it into *Excel* is a bit tricky as there is space between *Lon* and *Lat* and tab between the other columns.

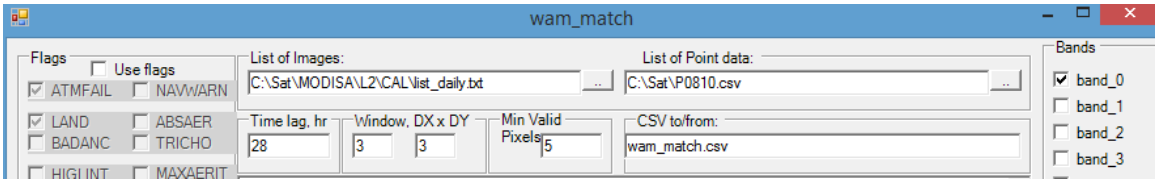
Using *Geo-Get Vector Objects* interactively in WIM works fine if you have a number of *in situ* values that you want to match with a **single** satellite image. In reality, you typically have many satellite images and the closest in time may be cloudy, so that and you need to match your in situ points with many other images. Which image to pick for each of the points becomes a difficult problem to solve as you need to consider (1) is the image area corresponding to the point clear; (2) which match-up image to pick (going forward in time or back in time) if you have more than one clear scene.

3 Using wam_match GUI application

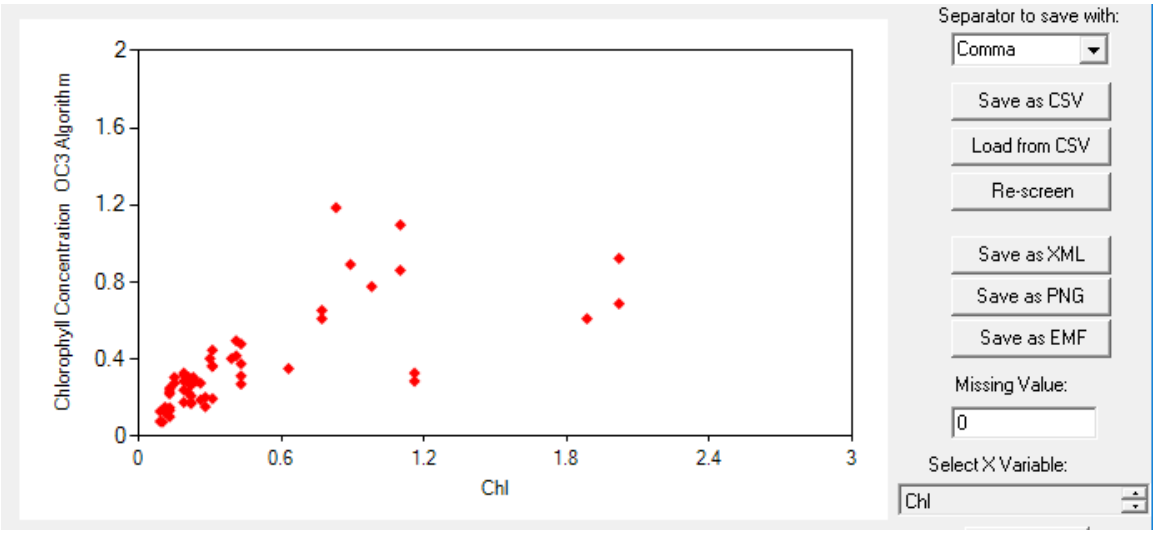
wam_match is a GUI (graphical user interface) that finds match-ups for a *List of Images* and a *List of points* and allows selecting a set of options. You need to select those list files in the text boxes. From the default settings



For the *List of Images* select C:\Sat\MODISA\L2\CAL\list_daily.txt and for the List of Points select *C:\Sat\P0810.csv*.



Click *Start* and you should get a scatter plot of satellite-detected Chl-a against MLD. Now use *Select X Variable* (bottom right) to scroll up to variable (column) *Chl*. You should have the following scatter plot:



Note that *C:\Sat\P0810.csv* works in *English (US)* region but in *French (France)* region you need to use, e.g. *C:\Sat\P0810_tab.csv* or *C:\Sat\P0810_semicolon.csv*.

4 Using WAM command line applications

4.1 Finding match-ups with wam_match_nearest

With the command line program *wam_match_nearest* you can find match-ups between a set of points and a set of image files. For each point, it finds the nearest image in time with at least *MinValid* (e.g. 3) valid pixels within the 3 x 3 pixel window. If an image has less than *MinValid*

valid points within the 3 x 3 pixel window centered at the point then it jumps to the next nearest image in time (going both forward in time and back in time) and this process continues forward and back until the image with enough valid pixels is found or, the process stops if no suitable match-up is found within *MaxDiffDays* (e.g. 5) days. You can control this selection process with many options.

We use *wam_match_nearest* and find match-ups with daily standard mapped image (SMI) files from SeaWiFS, MODIS-Aqua and MODIS-Terra. After that we find match-ups with MODIS-Aqua Level-2 files using a different command (*wam_l2_match*). Global SeaWiFS SMI data are available at 9-km resolution. For compatibility, we use the same for other sensors. We run the following commands:

```
cd C:\Sat
wam_match_nearest P0810.csv C:\Sat\MODISA\L3\Daily\CHL_9\2008\A*.hdf maxDiffDays=5 plotNthColumn=10 xMin=-2 xMax=2 yMin=-2 yMax=2
```

Note: If your Windows uses a different *Region* then you may need to use instead of *P0810.csv* you may need to use *P0810_tab.csv* or *P0810_semicolon.csv*.

```
wam_match_nearest P0810_tab.csv C:\Sat\MODISA\L3\Daily\CHL_9\2008\A*.hdf maxDiffDays=5 plotNthColumn=10 xMin=-2 xMax=2 yMin=-2 yMax=2
```

We must have a result looking like that:

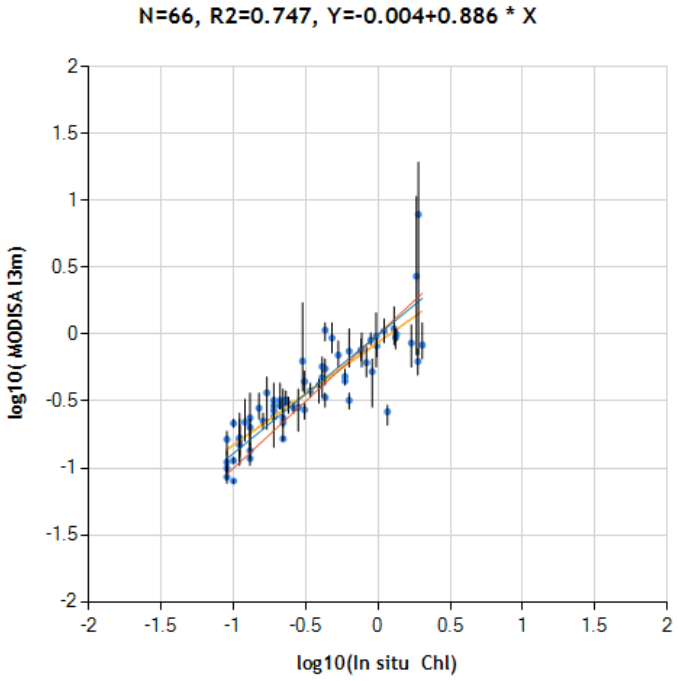


Fig. 3. Match-ups between in situ Chla and MODIS-Aqua daily SMI images at 9-km resolution.

With default options *wam_match_nearest* finds 66 match-ups, i.e. one for each station. However, the default maximum time lag is 5 days. We can refine the match-up file saved in previous step and select a shorter maximum time lag and possibly other conditions:

```
wam_read_match P0810_MODISA_l3m.csv plotNthColumn=10 xMin=-2 xMax=1 yMin=-2 yMax=1
```

wam_read_match P0810_MODISA_l3m.csv plotNthColumn=10 maxDiffDays=0.5 xMin=-2 xMax=1 yMin=-2 yMax=1

We now get less than 66 match-ups. Make sure you understand why. Look at the output files and try to understand the info in the file names.

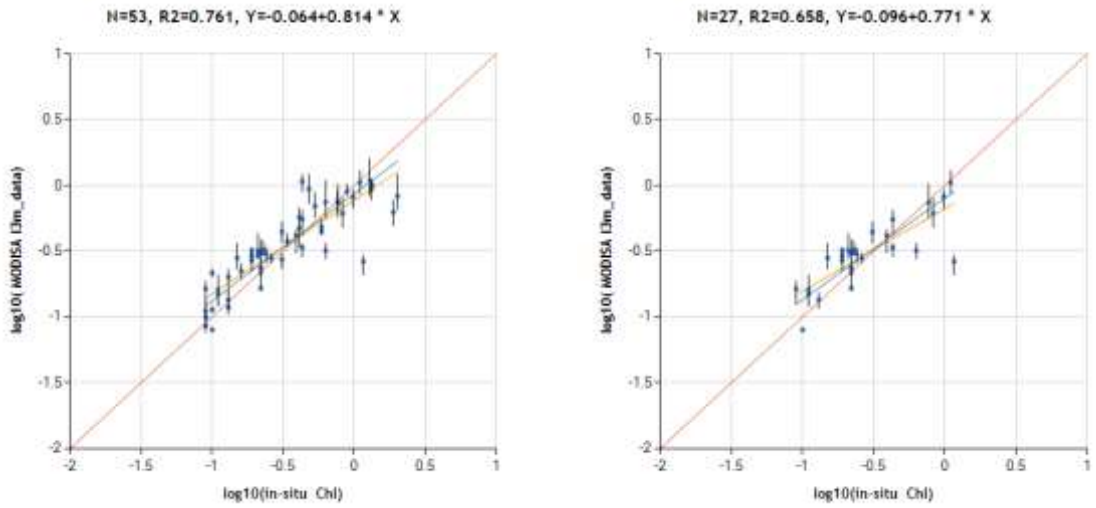


Fig. 4. Output from *wam_match_nearest* (log10 scales) with maximum time lags of 5 day (left) and 0.5 day (right).

Match-ups for MODISA Level-3 data:

wam_match_nearest P0810.csv C:\Sat\MODISA\L3\Daily\CHL_9\2008\A*.hdf maxDiffDays=5 plotNthColumn=10 xMin=-2 xMax=2 yMin=-2 yMax=2

wam_read_match P0810_MODISA_l3m.csv plotNthColumn=10 **maxDiffDays=0.5** xMin=-2 xMax=1 yMin=-2 yMax=1

And now for MODIST Level-3 data:

wam_match_nearest P0810.csv C:\Sat\MODIST\L3\Daily\CHL_9\2008\T*.hdf maxDiffDays=5 plotNthColumn=10 xMin=-2 xMax=2 yMin=-2 yMax=2

wam_read_match P0810_MODIST_l3m.csv plotNthColumn=10 **maxDiffDays=0.5** xMin=-2 xMax=1 yMin=-2 yMax=1

Compare the different sensors. Do you see any difference in quality of the satellite estimates?

4.2 wam_match_l2

In order to find match-ups at the highest spatial resolution and shorter time lags we must use Level-2 data that have multiple unmapped variables at full spatial resolution in the same file. We run *wam_match_l2* with the maximum allowed time difference of 5 days:

wam_match_l2 P0810.csv C:\Sat\MODISA\L2\CAL\2008\A*.hdf maxDiffDays=5

We use 5-days time lag initially and can refine the match-ups later:

wam_read_match P0810_MODISA.csv plotNthColumn=10 xMin=-2 xMax=1 yMin=-2 yMax=1

wam_read_match P0810_MODISA.csv plotNthColumn=10 maxDiffDays=0.125 xMin=-2 xMax=1 yMin=-2 yMax=1

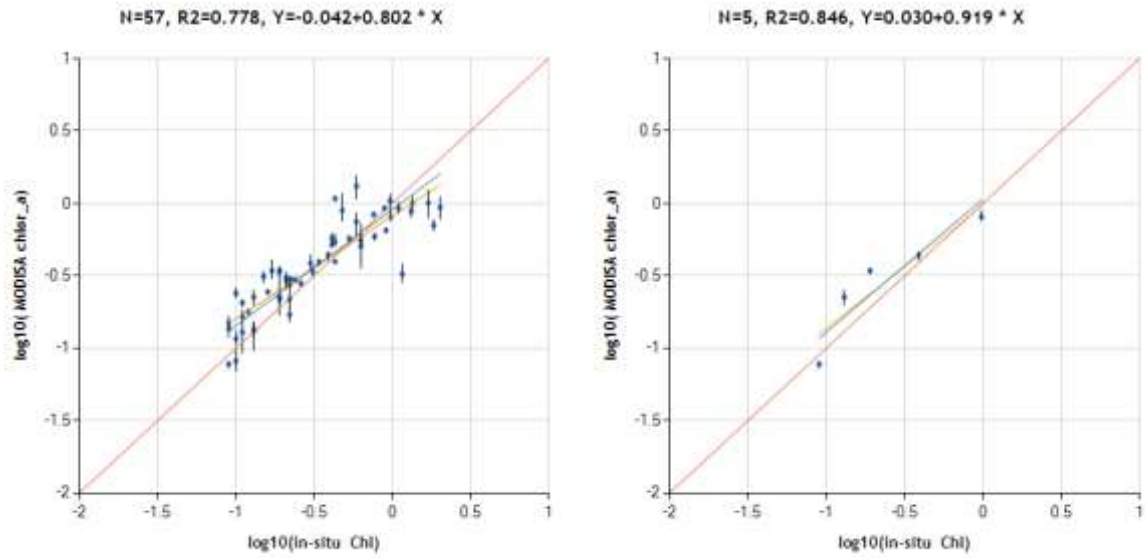


Fig. 5. Output from *wam_match_l2* using maximum time lag of 5 days (left) and 3hr (right).

In this case we used individual Level-2 files at the nominal 1-km resolution (at nadir, lower at swath edges) and we found 57 match-up points with less than 5 day time differenc 5 match-ups with 3 hour lag.

A major difference between L2 and L3 match-ups is that L2 files have a set of variables called L2-flags that report all kinds of potential problems. *wam_match_l2* uses these level-2 flags to screen valid values from invalid values whereas in Level-3 match-ups we can only check if the value itself is valid. Some new data products are also adding uncertainty or estimated error that can be used in Level-3 match-ups for Open the Level-2 match-up file *P0810_MODISA.csv* and try to understand the meaning of the L2-flags reported for each of the 9 match-up pixels.

5 Sensor against sensor validation

5.1 Validating Rrs from different sensors

In previous exercises we compared satellite estimates with in situ measurements. Here we compare one sensor to another sensor by using Level-3 daily standard mapped images. Global images have millions of pixels (~16 million for 9 km images) and therefore we need to limit the area of interest to a smaller area in order to reduce the number of match-ups. As area of interest (mask) we have chosen a rectangular area 100 km wide and 600 km long extending from the coast and covering contrasting environments from high-Chla upwelling area to low-Chla oligotrophic area. The sample mask file is *Mask_strip2_4320x2160.hdf* and it is in the *Sat/Masks* folder. Open the mask file in WIM and try to locate the non-zero area, i.e. California coast. It is easier to find it when you create coastlines with the icon on the Toolbar or with *Geo – Get Map Overlay* and pick the *coast_inter.b* file with background value of 0 and foreground value 255. Then click on the black image and then overlay the coastlines with *Overlay* icon on the Toolbar. If you Stretch colors then you can visualize it like below but the essential part is the pixels with non-zero values (e.g. 1 in this case) represent the area of interest, i.e. the mask. Only those pixels corresponding to the mask will be used in the match-ups between satellite images

and in the plots. You can make your own mask of any other area by basically drawing on the map and filling areas but please don't make the areas too big as otherwise you will get too many matching points.

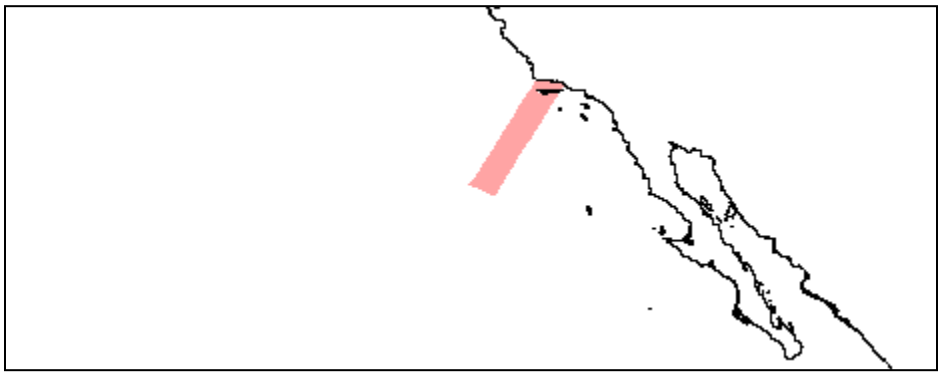


Fig. 6. Mask (area of interest) used in the sensor inter-comparison exercise.

We will now make match-ups between Level-3 global daily mapped *chlor_a* images of MODISA versus SeaWiFS, MODIST and MERIS. Open the command windows, *cd* to the *Sat* folder and run the command *wam_pixelwise_match*, e.g.

```
cd C:\Sat
wam_pixelwise_match  C:\Sat\MODISA\L3\Daily\CHL_9\2008\A*.hdf
    C:\Sat\SEAWIFS\L3\Daily\CHL_9\2008\S*.hdf Masks\Mask_strip2_4320x2160.hdf xMin=-2
    xMax=1 yMin=-2 yMax=1
wam_pixelwise_match  C:\Sat\MODISA\L3\Daily\CHL_9\2008\A*.hdf
    C:\Sat\MODIST\L3\Daily\CHL_9\2008\T*.hdf Masks\Mask_strip2_4320x2160.hdf xMin=-2
    xMax=1 yMin=-2 yMax=1
```

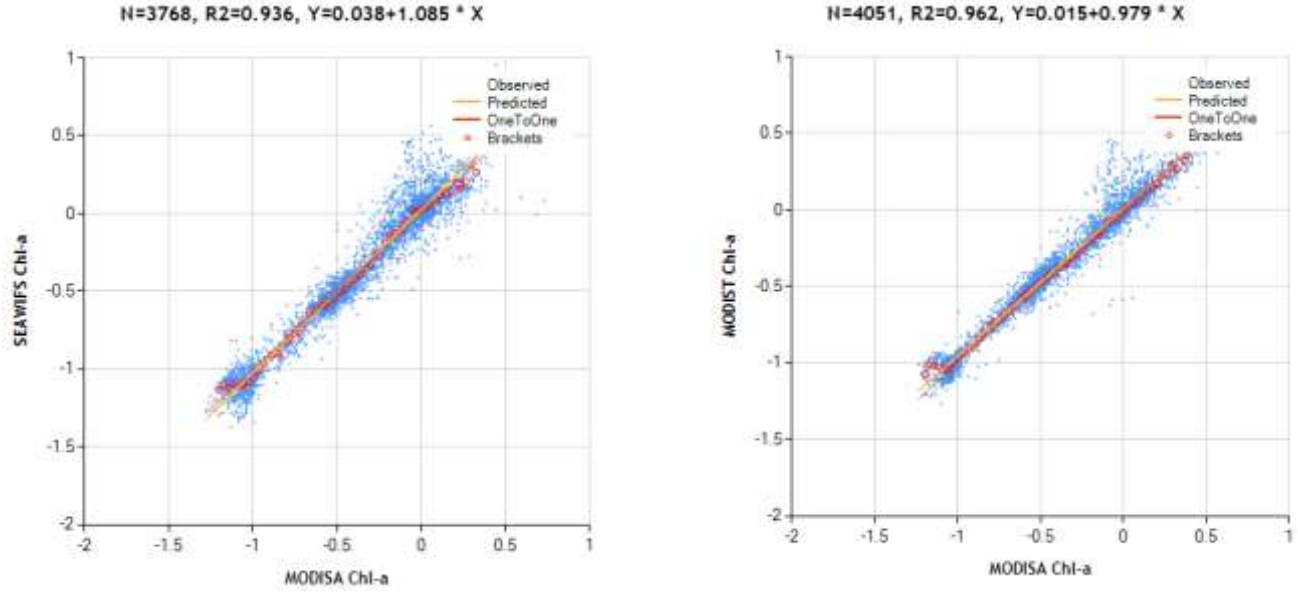


Fig. 7. Pixel-wise comparison of SeaWiFS (left) and MODIST (right) *chlor_a* against MODISA in the California Current region in October, 2008.

wam_pixelwise_match C:\Sat\MODISA\L3\Daily\Rrs_9\2008\A*412_9km
C:\Sat\SEAWIFS\L3\Daily\Rrs_9\2008\S*412_9km Masks\Mask_strip2_4320x2160.hdf
xMin=-3.5 xMax=-1.5 yMin=-3.5 yMax=-1.5
Not included! wam_pixelwise_match C:\Sat\MODISA\L3\Daily\Rrs_9\2008\A*412_9km
C:\Sat\MERIS\L3\Daily\Rrs_9\2008\L3b*RRS412*.hdf Masks\Mask_strip2_4320x2160.hdf
xMin=-3.5 xMax=-1.5 yMin=-3.5 yMax=-1.5

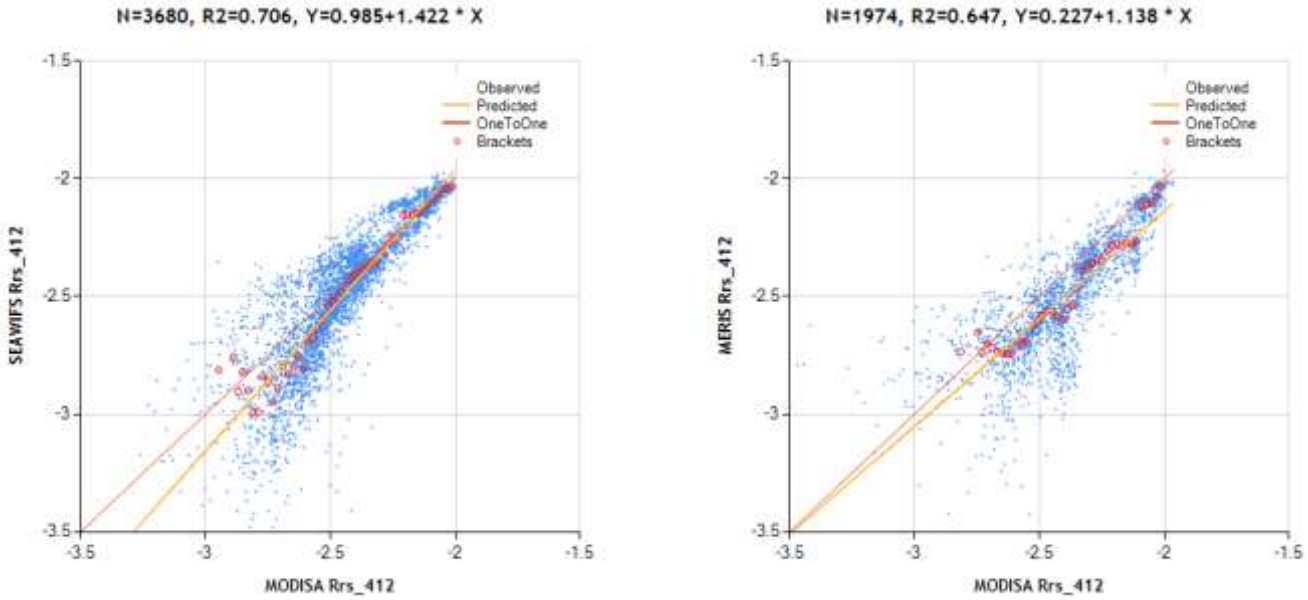
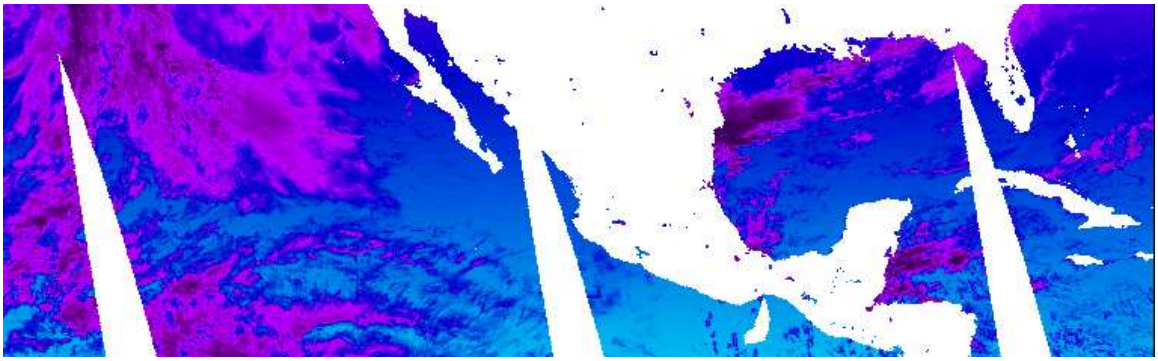


Fig. 8. Pixel-wise comparison of SeaWiFS (left) and MERIS (right) *Rrs_412* against MODISA in the California Current region in October, 2008.

As you can see, there are both random and systematic variations differences between SeaWiFS, MERIS and MODISA *Rrs* values. The same pixel on the same day can have very different values when measured by another sensor during the same day. These differences will certainly be transferred to differences between all derived variables like Chla, PP, etc.

5.2 Validating PAR from MODIST against MODISA

An important variable in estimates of primary production is PAR (photosynthetically active radiation). Daily PAR images have inter-orbit gaps (see example below) due to limited width of the swath. Confirm that by opening any of the daily PAR images in *\Sat\MODISA\L3\Daily\PAR_9\2008* or *\Sat\MODIST\L3\Daily\PAR_9\2008*.



Missing PAR on daily images limits the estimation of primary productivity, therefore it is desirable to merge PAR data from multiple sensors and 1) eliminate gaps, and 2) reduce errors by averaging over multiple sensors. However, before merging data from multiple sensors we need to confirm that PAR values from different sensors are compatible. Therefore we will examine PAR from MODISA versus PAR from MODIST

First we compare pixel-wise estimates of PAR over the California Current region in correspnding daily 9-km images:

```
cd C:\Sat
wam_pixelwise_match MODISA\L3\Daily\PAR_9\2008\A*.hdf
MODIST\L3\Daily\PAR_9\2008\T*.hdf Masks\Mask_strip2_4320x2160.hdf xMin=0.5 xMax=2.5
yMin=0.5 yMax=2.5
```

Note: In *French (France)* of Windows the arguments including decimal point (e.g. 1.5) must be with comma.

You should get a plot like in Fig. 8 (left). Note that we are using only 20 daily images in January 1-20 of 2012. Now we make a similar comparison but using 12 monthly images of 2004. Note that the new files will **overwrite** the current files! If you want to keep the daily files, either rename them or move to a different folder!

```
wam_pixelwise_match MODISA\L3\Monthly\PAR_9\A*.hdf MODIST\L3\Monthly\PAR_9\T*.hdf
Masks\Mask_strip2_4320x2160.hdf xMin=1 xMax=2 yMin=1 yMax=2
```

You should get a plot like in Fig. 8 (right).

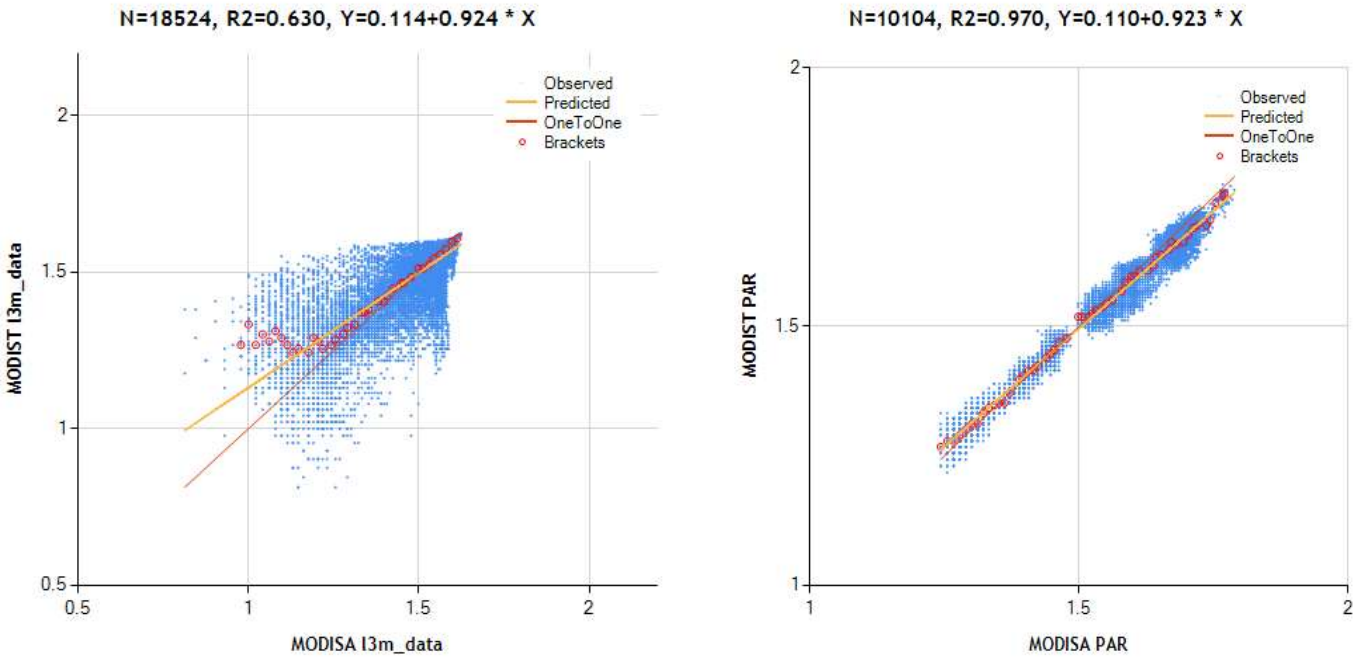


Fig. 9. Pixel-wise comparison of MODIST PAR against MODISA PAR in daily mapped datasets (2008/10/12 to 2-Nov-2012, left) and monthly composited datasets (2008, right).

Note the differences in the ranges (Oct-Nov versus all year!) and that there is a much less scatter in monthly datasets compared to daily PAR datasets. Daily PAR values can be much lower due to overcast skies whereas monthly values tend to be higher as they get at least some clear skies. We conclude that PAR estimates from MODISA and MODIST are compatible.

6 Validating primary production estimates

Saba, V. S., et al. (2011), An evaluation of ocean color model estimates of marine primary productivity in coastal and pelagic regions across the globe, *Biogeosciences*, *8*, 489–503) compared various net primary production (NPP) models and assembled a dataset of in situ NPP measurements. Here we use a subset of their dataset using only data from the Mediterranean (Med) and West Antarctic Peninsula (WAP). The Vertically Generalized Production Model (VGPM) of Behrenfeld and Falkowski (1997) is the most popular model of NPP due to its robustness and simplicity. Global NPP datasets based on satellite data and calculated according to VGPM are available at Oregon SU http://www.science.oregonstate.edu/ocean.productivity/. Here we compare the MedWAP dataset with the OSU estimates and can do our own estimates using satellite data.

6.1 Finding NPP match-ups with wam_match_nearest

Here we use the same command (*wam_match_nearest*) that we used before. Now we find match-up to the MedWAP data using 8-day NPP datasets produced by OSU.

```
C:
cd C:\Sat
wam_match_nearest MedWap_NPP.csv MODISA\L4\8Day\NPP_VGPM_OSU_byte\*.hdf
    plotNthColumn=6 xMin=1 xMax=4 yMin=1 yMax=4
```

Note: the .csv file is in the *English (US)* format. A corresponding tab separated file *MedWap_NPP_tab.csv* must be used if, e.g. in *French (France)*.

After it finishes, it should have found 235 match-ups and created a plot like the following:

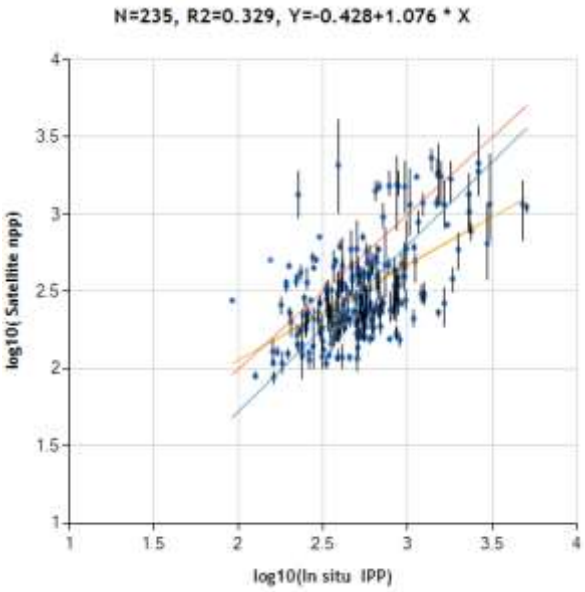


Fig. 10. Output from *wam_match_nearest* showing *in situ* vertically integrated NPP as X variable and NPP from global 8-day VGPM data using MODISA as Y-variable.

Note that the OSU files have been generated using MODISA data but do not have any attribute specifying MODISA in the file name or file attributes. Therefore *wam_match_nearest* was not able to determine the sensor name and reports it as *Satellite*. Also note that we have taken the OSU NPP data and converted the datasets (in float32 format) to byte format with logarithmic scaling. This was done to reduce the size of the files by about 7 times. This conversion did not significantly change the output. As we can see, there is a considerable scatter, both over- and under-estimation by the satellite VGPM algorithm when compared to in situ NPP.

```
wam_match_nearest Med_NPP.csv MODISA\L4\8Day\NPP_VGPM_OSU_byte\*.hdf
    plotNthColumn=6 xMin=1 xMax=4 yMin=1 yMax=4

wam_match_nearest Wap_NPP.csv MODISA\L4\8Day\NPP_VGPM_OSU_byte\*.hdf
    plotNthColumn=6 xMin=1 xMax=4 yMin=1 yMax=4
```

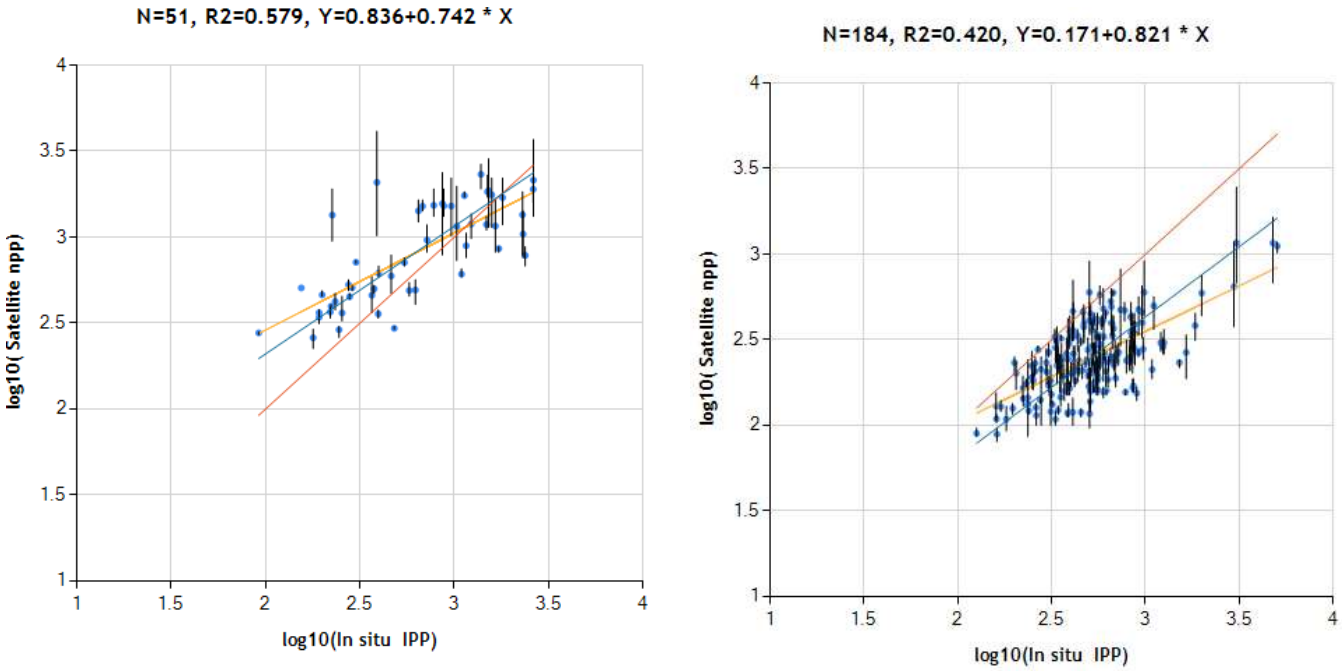


Fig. 11. As Fig. 10 but separated for the Mediterranean Sea (left panel) and Western Antarctic Peninsula (right panel).

Note that the NPP estimates mostly overestimate NPP in the Mediterranean but significantly underestimate NPP in the WAP region. That demonstrates again the need for regional validation and regional model differentiation.