EOF (Empirical Orthogonal Function) and PCA (Principal Component) analysis with WAM

1 Introduction

Empirical Orthogonal Function (EOF) analysis and the related Principal Components Analysis (PCA) are a set of powerful methods to extract information from large datasets, e.g. series of satellite images. The terminology of these methods is rather confusing as similar methods are used under a number of different names. In this exercise we will use clearly defined mathematical methods and use terminology that suits us. Please keep in mind that other software is probably using different names.

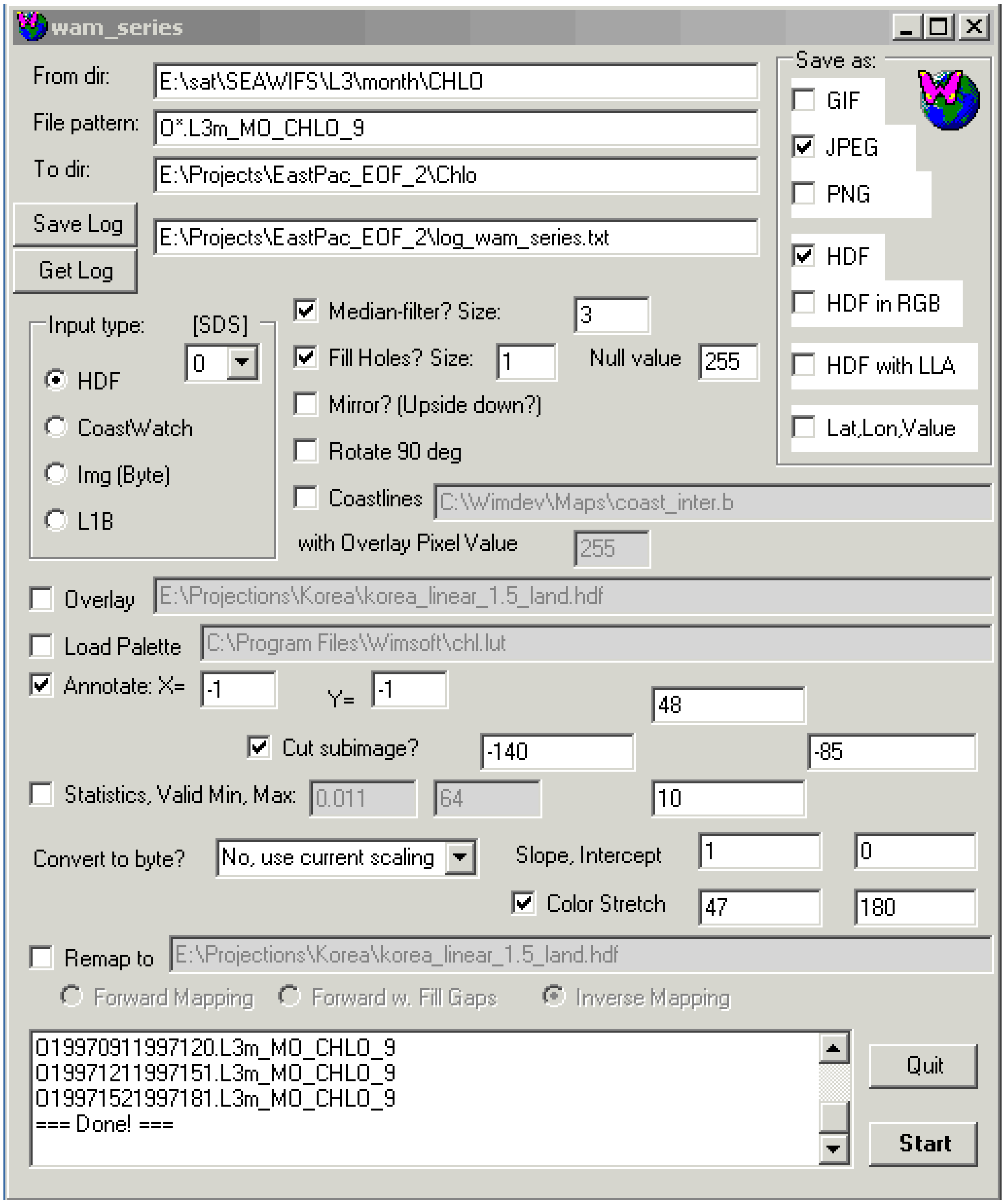
2 Prerequisites

We assume that you are familiar with the basics of the command line, i.e. how to open the command window, change directory, issue a command, etc. We also assume that you are familiar and able to run WAM command line programs. If not, please check out the basic WAM and WAM exercises manuals. We also assume that you have a set of images that you are going to use. In our example we use standard mapped monthly chlorophyll images from OCTS, SeaWiFS and MODIS-Aqua.

3 Selecting the standard map and area of interest

Before you start processing the data you need to pick your area of interest and a suitable target map projection. In this exercise we will just use subsets of the standard 9-km SMI (standard mapped) images of Chl-a from two sensors - OCTS and SeaWiFS. You can also use the recently released merged SeaWiFS/MODIS-Aqua data (starting from July, 2002) but you must consider the different scaling issue (Int16 pixels instead of the Byte pixels and the short time span of the merged data). Here we pick Eastern Pacific as our area of interest. You can choose another area and/or a different map projection. A different map projection is appropriate, e.g., for polar data.

To create the subset dataset we use *wam_series* to cut out our area of interest and perform several operations on the way. If you want to your area of interest in a different projection then, instead of cutting out a subset, you need to remap it to a target projection (typically another HDF file), e.g. with *wam_series*. A screenshot of *wam_series* is shown below.



wam_series

From dir: E:\sat\SEAWIFS\L3\month\CHLO

File pattern: O*.L3m_MO_CHLO_9

To dir: E:\Projects\EastPac_EOF_2\Chlo

Save Log / Get Log: E:\Projects\EastPac_EOF_2\log_wam_series.txt

Input type: [SDS] 0
● HDF
○ CoastWatch
○ Img (Byte)
○ L1B

☑ Median-filter? Size: 3

☑ Fill Holes? Size: 1 Null value 255

□ Mirror? (Upside down?)

□ Rotate 90 deg

□ Coastlines C:\Wimdev\Maps\coast_inter.b

with Overlay Pixel Value 255

Save as:
□ GIF
☑ JPEG
□ PNG
☑ HDF
□ HDF in RGB
□ HDF with LLA
□ Lat,Lon,Value

□ Overlay E:\Projections\Korea\korea_linear_1.5_land.hdf

□ Load Palette C:\Program Files\Wimsoft\chl.lut

☑ Annotate: X= -1 Y= -1

48

☑ Cut subimage? -140 -85

□ Statistics, Valid Min, Max: 0.011 64

10

Convert to byte? No, use current scaling

Slope, Intercept 1 0

☑ Color Stretch 47 180

□ Remap to E:\Projections\Korea\korea_linear_1.5_land.hdf

○ Forward Mapping ○ Forward w. Fill Gaps ● Inverse Mapping

```
O19970911997120.L3m_MO_CHLO_9
O19971211997151.L3m_MO_CHLO_9
O19971521997181.L3m_MO_CHLO_9
=== Done! ===
```

Fig. 1. Screenshot of *wam_series* showing the settings for producing a series of subset images for Eastern Pacific (here for OCTS images).

You will probably have to change the folder names when running *wam_series* on your computer. As you can see, we cut out an area between 10N and 48N and 140W and 85W. Please note that EOF analysis needs to create a huge matrix based on the many images and unless you have HUGE amounts of RAM in your computers, you should not use much bigger images. If you want to cover larger area then you need to reduce or remap the images to a smaller target. We perform median smoothing and fill small "holes". We save both HDF and JPG format files. The example above processes OCTS files; simple modifications of the *File pattern* are needed for SeaWiFS. A sample output file is shown below.

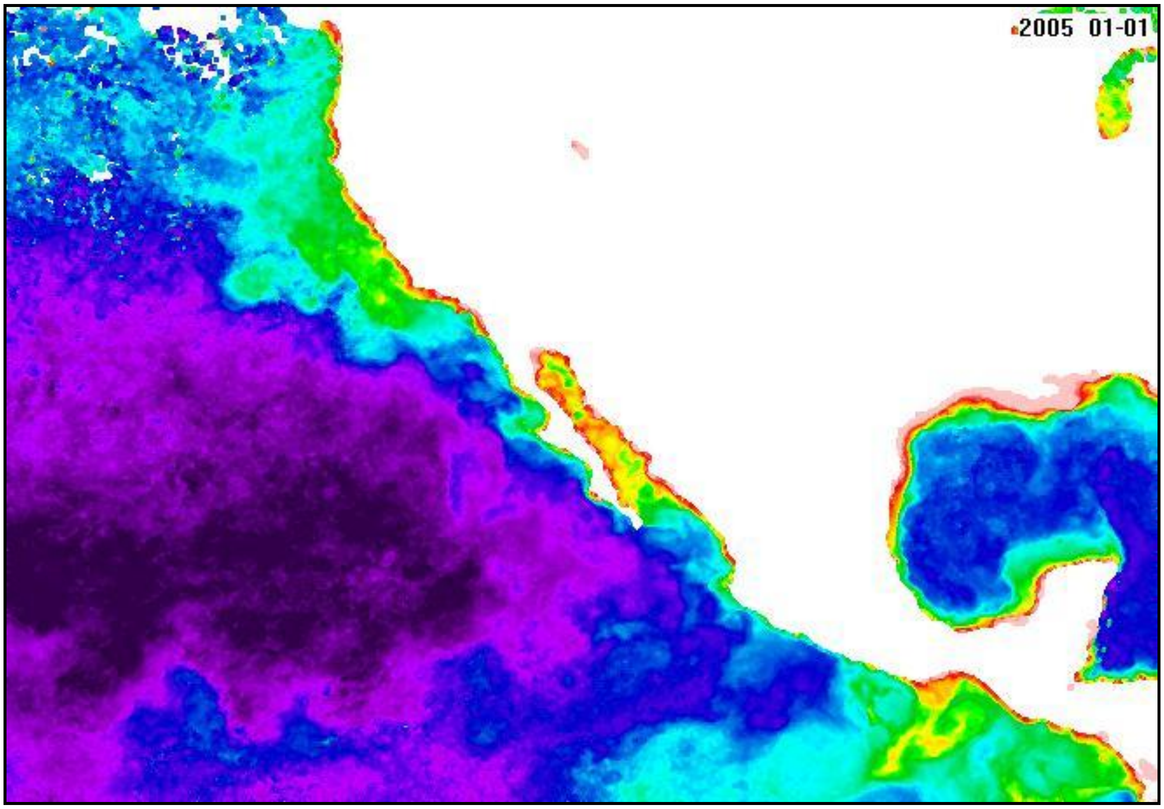


Fig. 2. Sample chlorophyll image subset from *wam_series* for January 2005.

You can see the effects of smoothing and filling. There are still areas of no data (white patches) but that is not crucial as we will be using anomalies and if there is no data we assume that there is no anomaly either (i.e. we have the expected mean). Also note that we have the annotation (image date) written into the image. We will exclude these altered values from the computations as we define a mask area that we are going to use. In order to generate a mask, read one of the created HDF images and create the coastlines with *Geo-Get Map Overlay*. I use the *coast_full.b* coastline database and the pixel value 255. Now I use *Edge-Dilat*e to expand the coastlines 1 times. Then I use *Edit-Draw-Fill* with pixel value 1 and fill the Pacific Ocean. The output is shown below. Please remember to **stretch the colors from 0** – otherwise all pixel values below 47 (as set in *wam_series*) will be shown as black!

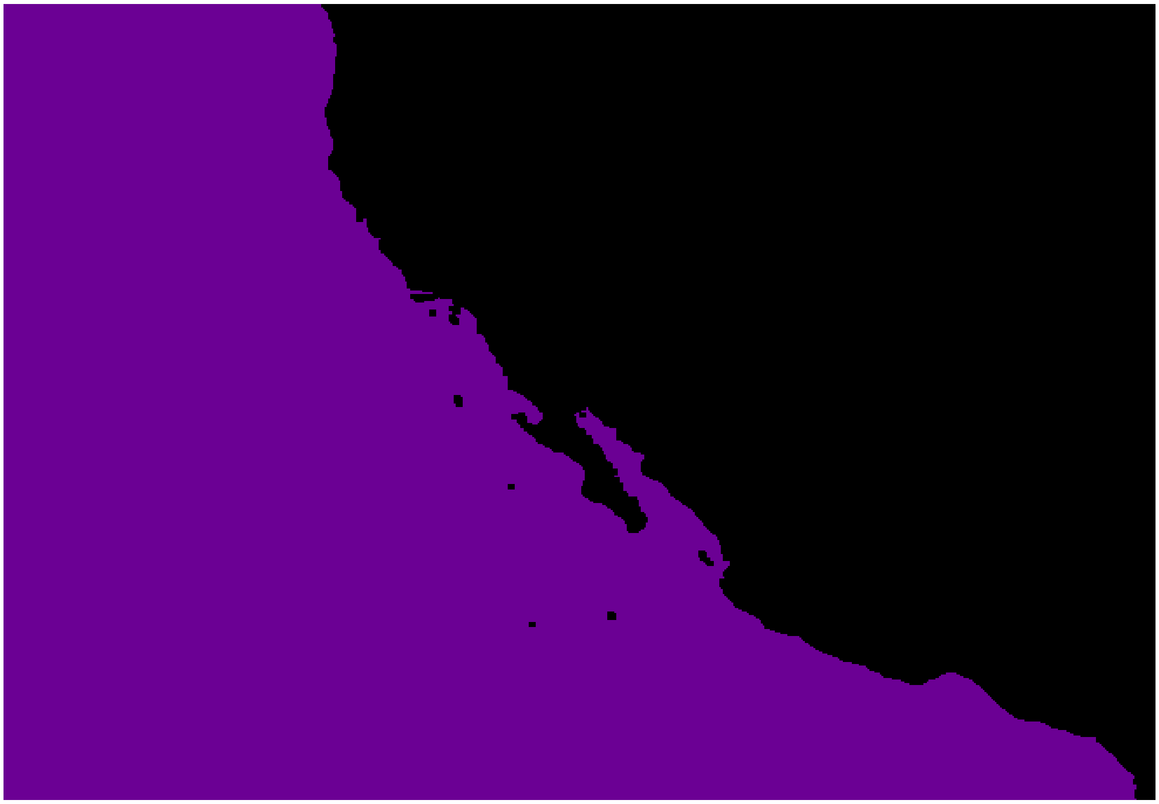


Fig. 3. Sample **sea** mask (the pixel value for land is 0 and for ocean is nonzero, e.g. 1).

The idea of the mask above is that pixel values different from 0 will be used in the subsequent analysis. We now get rid of the coastlines with *Transf-Replace values* from 255 to 255 with 0 and save the output as our mask in HDF format. We choose to save the mask under the name *mask_epac_sea.hdf* as it is a **sea mask**, i.e. ocean and sea pixels have values different from zero. We now use *Transf-Replace Values* a few times and create a **land mask**, i.e. mask where land pixels are different from 0. You have to replace pixels with value 0 to 255 and pixels with value 1 to 0. Figure out the exact sequence yourself! The output should be something like the image below.

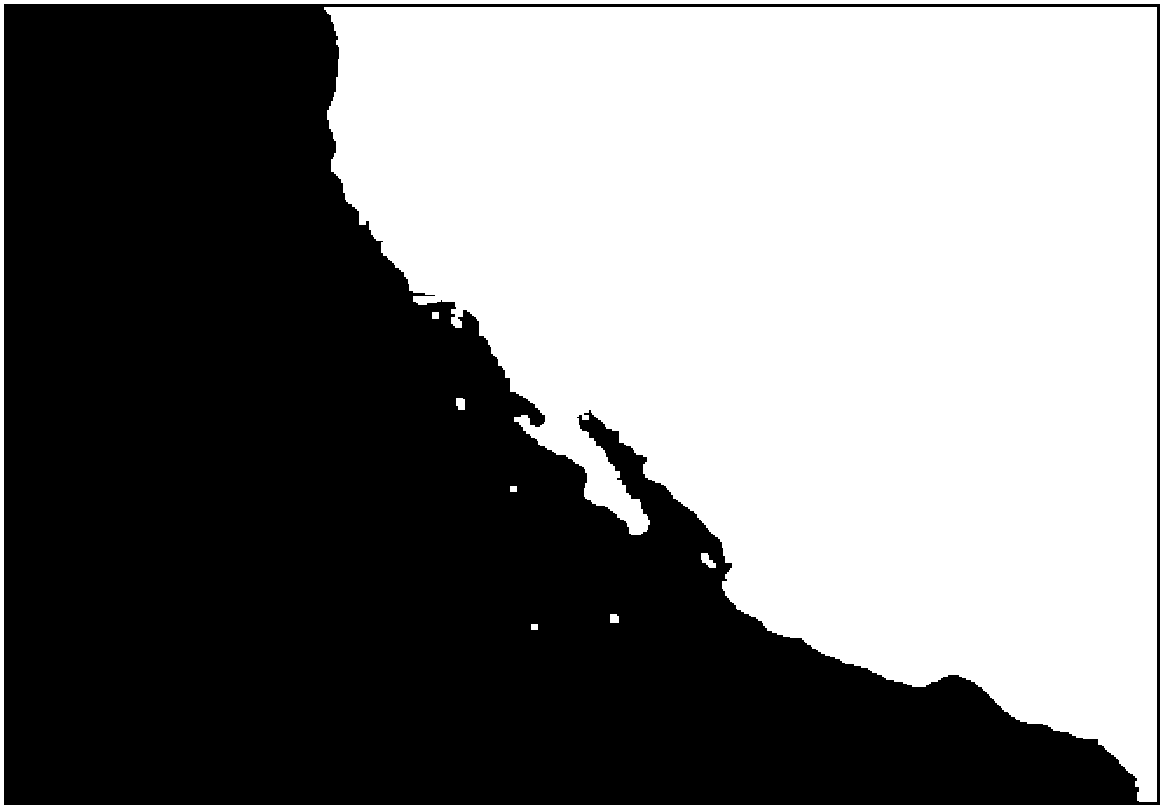


Fig. 4. Sample **land** mask (the pixel value for land is 255 and for ocean is 0).

As you can see, we have expanded the land areas in order to exclude ocean pixels immediately next to the coast. Now we are ready to create the monthly anomaly series.

You can see the syntax (all possible arguments) by invoking *wam_anomaly* without arguments. In short, we use number 12 (twelve) to calculate monthly anomalies (use number 1 (one) for yearly anomalies), *false* to not show missing values. Now type:

```
wam_anomaly Chlo\*.hdf 12 anomaly5.lut false mask_epac_land.hdf
```

Note that this is all in one line. Please note that the full path of the anomaly LUT file cannot include spaces. For example, it you have *anomaly5.lut* under *C:\Program Files\Wimsoft\LUT* then you need to **copy** (not move!) it in from there to the folder with no spaces in the full path. If you want the anomaly images have the date written into the, you should modify the above command slightly and issue the following command:

```
wam_anomaly Chlo\*.hdf 12 anomaly5.lut  false mask_epac_land.hdf no 555 17
```

The "no" here means that we are not going to use a previously calculated mean image but calculate the mean as part of this analysis (first scan of all the files). The numbers 555 and 17 mean, respectively the X and Y coordinates in pixels where the date annotation will be written. You can select the best location for the annotation by opening a sample image and right-clicking at a suitable pixel and then reading the X and Y from the window header.

Wam_series runs through all the matching files twice: first time to calculate the **means** and **valid counts** and the second time to calculate the **anomalies**. After it finishes you can observe that it has created the corresponding *Mean* and *ValidCounts* files as well as the **_anomaly.jpg* and **_anomaly.hdf* files. It is convenient to sort the files by Date and move all the anomaly files as well as the *Mean* and *ValidCounts* files to a new folder, e.g. *Chlo_Anomaly*. You may also want to slightly change the *Mean* and the *ValidCounts* filenames by getting rid of the "*_.hdf_*".

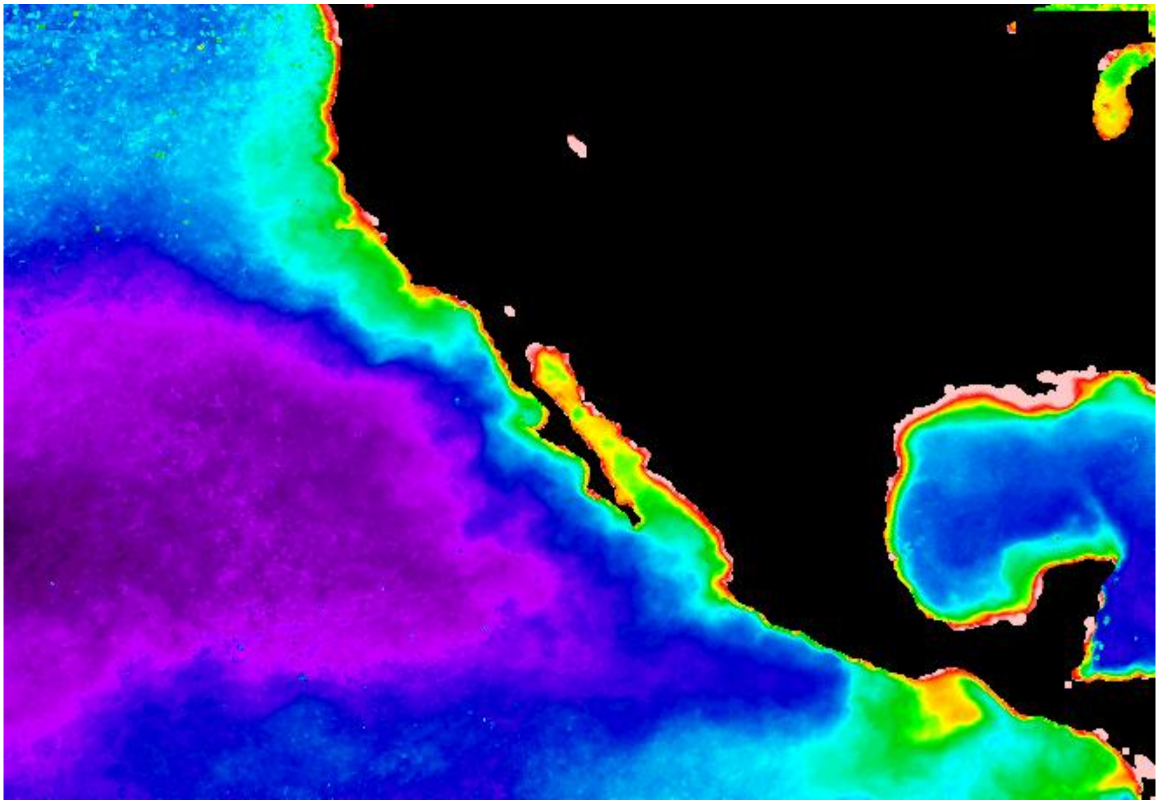


Fig. 5. Mean Chl-a for January.

Now take a look at the content of the *Mean* and *ValidCounts* files. Explain the meaning of these images. As you can observe, **this is a very powerful method to create monthly mean images from a large series of data with a single command.** All monthly means (January to December) are in a single HDF file. You may want to better visualize the monthly means by overlaying coastlines and other masks. You can do all that for all 12 monthly images in a single command. You need to create the overlay image first. For example, create a simple overlay with the coastlines (*Geo – Get Map Overlay*) with pixel value 1 and save it as *coast.hdf* in the same folder. You can then issue a command

```
wam_overlay Mean.hdf coast.hdf
```

This will create 12 monthly HDF files and 12 monthly JPEG files. The month number is also added to the image.

Please note that we used the **land mask** (and **not the sea mask**) here. All pixels marked with the land mask as not zero will not be used in the anomaly analysis. This also includes ocean areas that we want to exclude. For example, in this analysis we don't want to include the Atlantic Ocean and the Great Lakes. They are excluded as they are marked in the land mask. We have now produced 2 anomaly files for each input image: the HDF file with the data values and the JPEG file for quick visualization. If the HDF image has its horizontal dimension larger than 650 pixels then the JPEG image will be reduced. Below is a sample anomaly for January 1998.

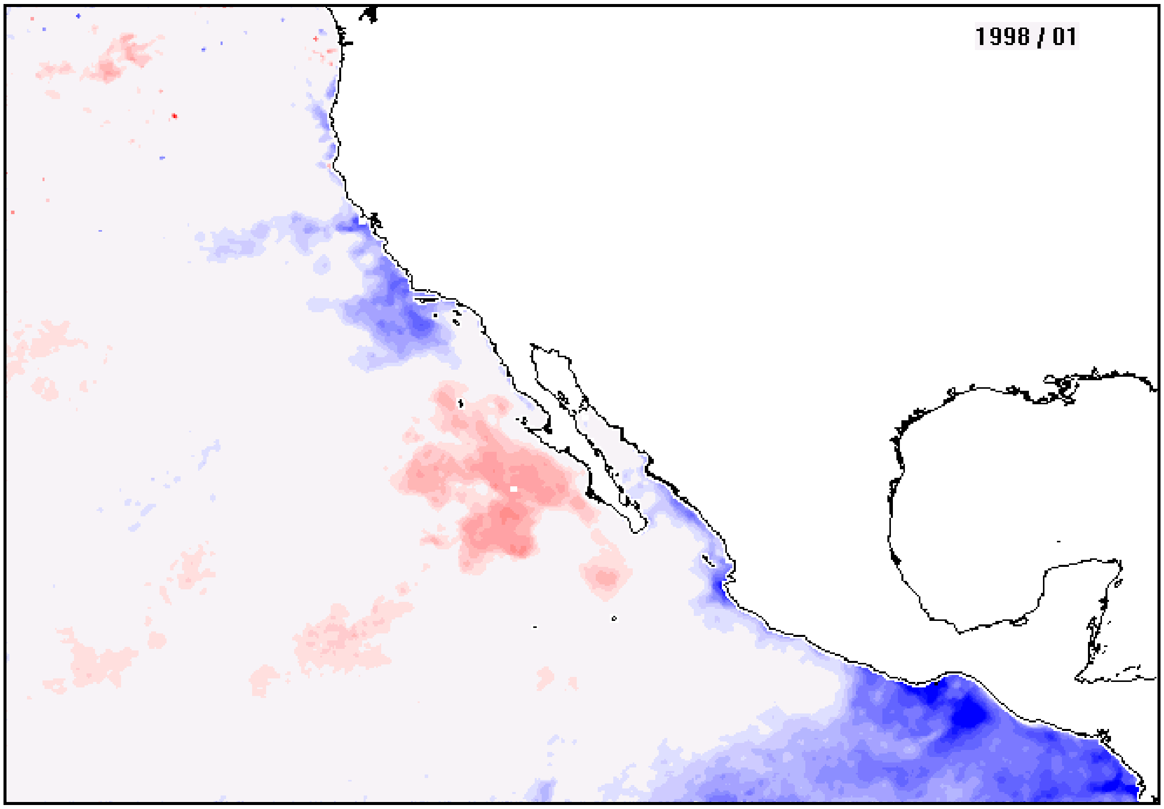


Fig. 6. A sample anomaly image for January 1998 showing the El Niño conditions, i.e. positive anomaly (red) off Baja California peninsula and negative anomaly (blue) in many other areas.

Blue areas are negative anomalies and red areas positive anomalies. However, the color depends on the actual lookup (LUT) file used. The blue and red colors above can be seen after loading the anomaly5.lut LUT (with *File-Lookup Table-Load LUT*). For chlorophyll the anomalies are calculated as the ratio to the mean whereas for SST the anomalies are the difference from the mean. Open one of the anomaly HDF files to look up the values (with right-click). For example, a Chl anomaly value of 0.2 (blue) means that it is 0.2 times the normal value and anomaly value of 4.0 (red) means that it is 4 times the expected mean value. Percent anomaly can be calculated as 100*(Anomaly – 1). For example, these two examples would be -80% and +300% anomalies. Depending on the palette file used, values close to the mean are light-gray. A useful exercise is to create a movie-loop (animated GIF) from the JPEG quick-looks (e.g. using the *Babarosa GIF Animator*).

The anomaly analysis so far has been a powerful method to understand the spatio-temporal dynamics of complex series of images. For example, the sample anomaly image above shows that in January, 1998 during the 1997-1998 El Niño large areas had negative chlorophyll anomalies but in certain regions (e.g. off Baja California peninsula) chlorophyll was significantly higher. Please note that by using a specific anomaly LUT file we hide all small and possibly insignificant anomalies with the light-gray color and all areas colored with red or blue have strong anomalies that are probably highly significant.

4 EOF analysis

We assume that you have all the anomaly files (as HDF) in a folder *Chlo_Anomaly* and that you have the sea mask in the current folder. We can now perform the EOF analysis with the following command:

```
wam_eof Chlo_Anomaly\*anomaly.hdf mask_epac_sea.hdf
```

You can see the syntax (all possible arguments) by invoking *wam_eof* without arguments. The best is to use anomaly images (as in this example) – in that case we don't have to subtract any mean as demeaning is already done in the anomalies. In case you want to apply *wam_eof* directly to a series of images (not anomaly images) you should use either the *Pixel* or *Image* option. In *Pixel* option the mean for each pixel value over time (multiple images) is subtracted. In *Image* option the mean for each image (spatial mean) is subtracted from the corresponding image. As mentioned, we recommend **to use anomalies and not to subtract any mean**.

Bear in mind that EOF analysis involves inverting a huge matrix. While the operation is very fast due to the use of an advanced matrix library, it does require a lot of RAM in your computer. For example, the sample images used here have the size of 660 x 456 = 300,960 pixels. As we have 118 monthly images (November, 1996 to November, 2006) we are inverting a matrix of 300,060 x 118 = 35,513,280 pixels. This size works fine and quite fast with 1 MB of RAM. If you have much larger images then you can get the "out of memory" error. A quick solution for that is to reduce the size of the source images, redo the anomalies and the masks and try again with *wam_eof*.

As output from *wam_anomaly* we get 3 files: the first 6 principal components, the eigenvectors (EOFs) and eigenvalues. The eigenvalues file also shows the mean, SD, eigenvalue size, percentage and the cumulative percentage of the first N principal components. The beginning of the *eigenvalues.csv* file is shown below:

#	Mean	SD	EVal	EVal%	EvalCum%
1	131.989	14.089	12.827	10.9	10.9
2	128.278	15.033	6.909	5.9	16.7
3	127.43	14.889	4.798	4.1	20.8
4	127.6	17.008	4.164	3.5	24.3
5	130.62	18.001	3.688	3.1	27.4
6	127.64	19.073	3.51	3	30.4

This shows the first 6 lines showing the sequence number (*i*), the mean pixel value of the *i*-th image, its standard deviation, the eigenvalue, the percent size of the eigenvalue and the cumulative percent of the eigenvalues. The anomaly images are scaled so that the **no anomaly** (i.e. equal to the mean value) corresponds to pixel value 128 (in the middle of the 0-255 possible values of a Byte image). Therefore values above 128 mean above average and values below 128 mean below average mean for the whole image. This output shows that the first eigenvalue and its corresponding principal component describe 10.9% of the total variance. Each following EOF and the corresponding PC describes a smaller part of the total variance. The first 10 explain 40% of the total variance. The full table is saved in **Eigenvalues.csv* file. The **Eigenvectors.csv* file contains all the eigenvectors, i.e. time components of the corresponding eigenvalues. Please load it into MS *Excel* and check it out. As we were using files from multiple sensors (OCTS, SeaWiFS) the output is sorted by the filename and not by time. If you are using Aqua files starting with A, you need to manually sort the lines in Excel. Now save as *Excel* workbook (**.xls*). The first column shows the start time of the image in decimal years and the subsequent columns show the intensity of the corresponding mode (eigenvector). The spatial modes (principal components) corresponding to the eigenvectors are calculated as projections of the data matrix to the corresponding eigenvectors. The first 5 of the PCs are save in a file **PC.hdf*. You can load it with WIM. The image below shows the first principal component (spatial mode).

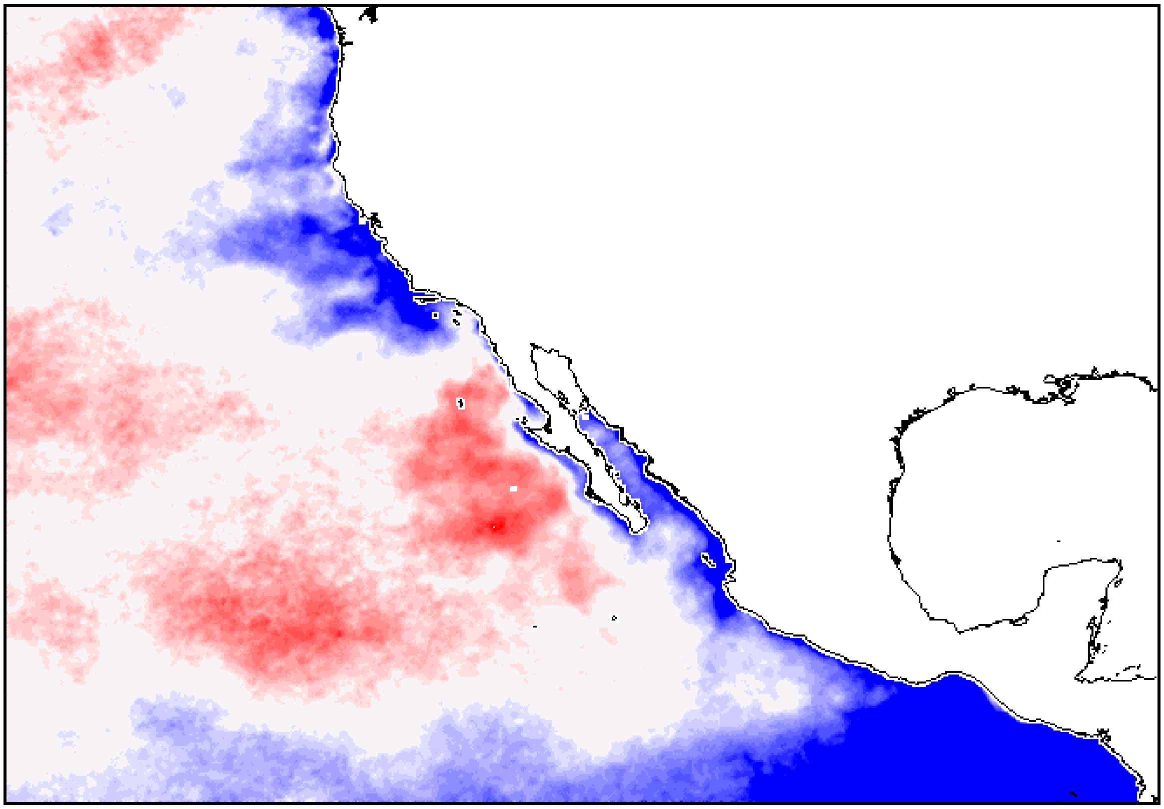


Fig. 7. The first principal component (spatial mode) of Chl variability in Eastern Pacific.

All the PCs have linear scaling and are stretched from min to max. The exact numerical values are not important as they can be multiplied by any constant, e.g. -1 in order to mirror it . It is convenient to annotate all PC images with the *wam_overlay* command. Please note that the spatial structure of the 1st principal component is similar to the anomaly structure during the peak of the El Niño in January, 1998 (Fig. 5). Positive values in PC1 image correspond to negative anomalies and vice versa, negative values of PC1 correspond to positive anomalies during the El Niño. Also, if you construct a time series plot of the first eigenvector (in file **Eigenvectors.xls* that you just saved) you can see resemblance to any of the El Niño-ENSO indices. Of course, the resemblance is not absolute as some features may show a different response compared to some indices. For example, when performing similar EOF analysis for the whole Northern Pacific we obtained the following EOF1 and Northern Oscillation Index plot.

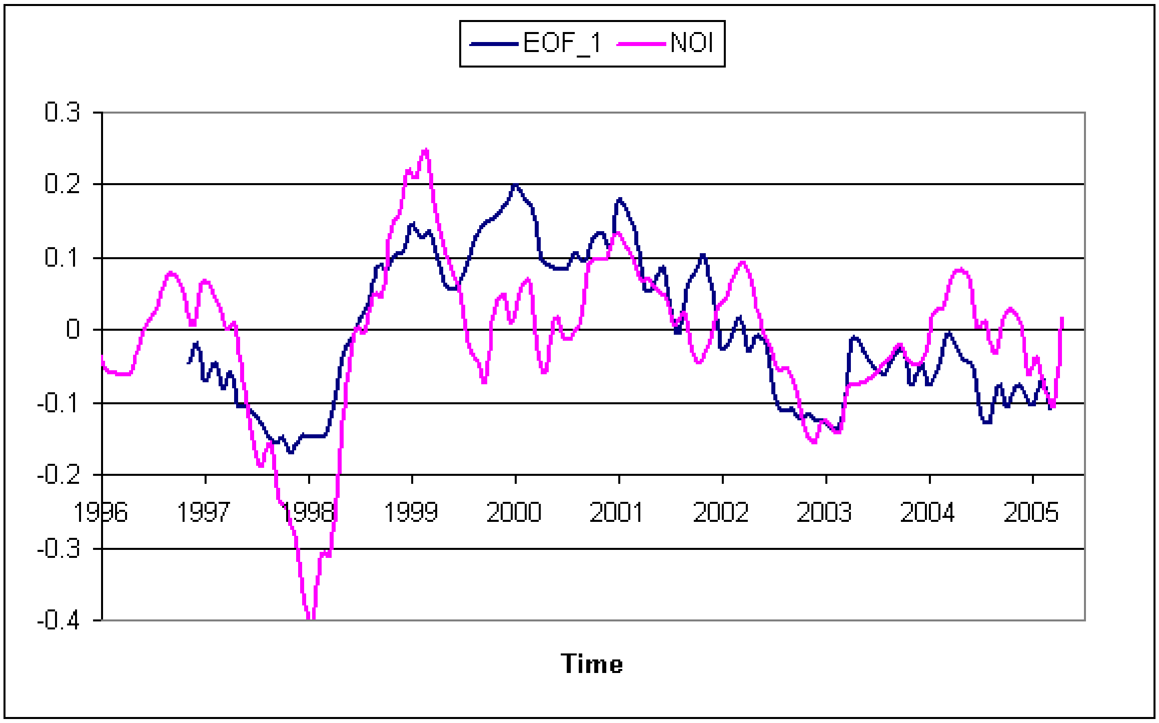
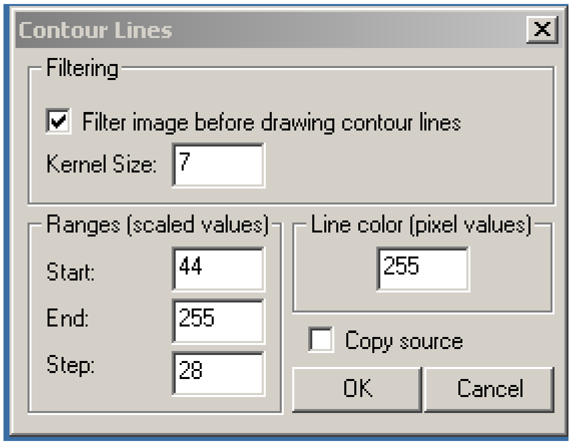


Fig. 8. The temporal evolution of the first principal component (EOF1) shows close resemblance with the Extratropical Northern Oscillation index (NOI, Schwing et al., 2002). Note the strong El Nino event of 1997-1998, and the weak events of 2003 and 2005 corresponding to the minima.

A similar analysis of chlorophyll variability in the North Pacific variability is described at http://spg.ucsd.edu/Satellite_Projects/NPac_variability/Npac_variability.htm.

We can now try to distinguish regions with similar time-scale patterns. For example, using PC1 we can separate regions with a different response to El Niño. Of course, we have to bear in mind the space-time variability is very complex and when using PC1 or any other component we are only looking at a small percentage of the total variability.

For example, in this analysis PC1 explains only about 11% of the total variance. We can use isolines overlaid to the image to visually separate different regions. For example, we can use *Examine-Contour lines* in WIM to create contour lines. Before doing that you may want to set the Value scaling to Pixel value in order to use pixel values and not the real scaled PC values. Pixel values are easier to understand here as they are scaled from 0 to 255 with the zero value corresponding to pixel value of 128 (middle of the 0-255 range). For example, we can use the following settings in the *Examine-Contour lines* dialogue.



☑ Filter image before drawing contour lines

Kernel Size: 7

Start: 44

End: 255

Step: 28

Line color (pixel values): 255

☐ Copy source

Fig. 9. *Examine-Contour lines* dialogue box in WIM.

You can notice with a step of 28 we get isolines for pixel values 44, 72, 100, 128, etc.

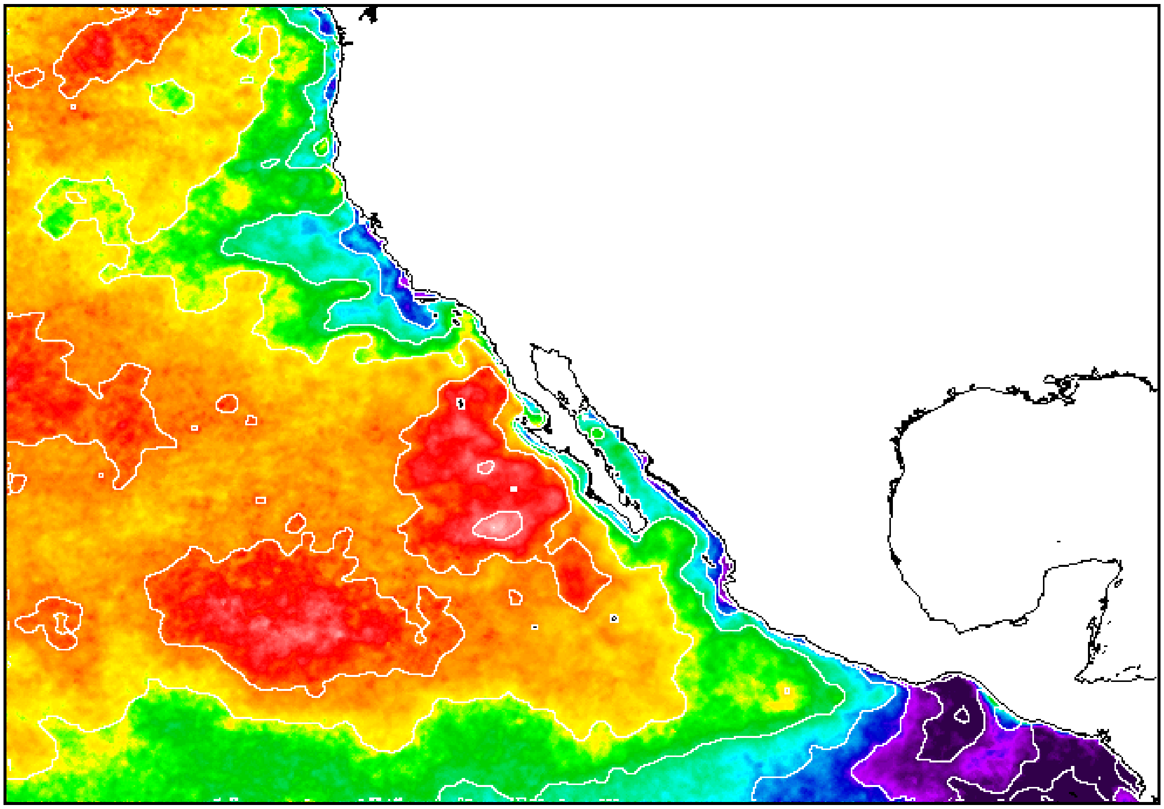


Fig. 10. PC1 with overlaid contour values.

This analysis shows the response of the Eastern Pacific surface chlorophyll to the El Niño-ENSO variability. As known from previous analysis (e.g. Kahru and Mitchell, 2000) El Niño causes suppressed chlorophyll values along the typical upwelling areas of the California coast but increased Chl offshore waters off the Baja California peninsula. As shown by Figs. 6 and 9 the strongest negative response to El Niño is in the Gulf of Tehuantepec and the Costa Rica dome area. Using this kind of EOF analysis we can more objectively map areas with similar variability structure. For example, we can more objectively define areas of interest with similar dynamics and construct time series for these areas (as in http://spg.ucsd.edu/Satellite_Projects/NPac_variability/Npac_variability.htm).

References

Kahru, M., B.G. Mitchell (2000), Influence of the 1997-98 El Niño on the surface chlorophyll in the California Current, *Geophysical Research Letters*, 27, 2937-2940.

Schwing, F.B., T. Murphree, P.M. Green (202), The Northern Oscillation Index (NOI): a new climate index for the northeast Pacific, *Progress in Oceanography*, 53, 115-139.