SOI Demo

This project file includes analysis of a **Southern Oscillation Index** dataset. The 690 month series `soi' was obtained from monthly mean sea-level pressures at Tahiti and Darwin, Australia, by removing their seasonal cycles, dividing the resulting anomalies by the corresponding standard deviations, and then taking the Tahiti-minus-Darwin difference. The SOI series considered here is for the time interval from January 1942 through June 1999, during which no observations are missing. **(**for more information see http://www.atmos.ucla.edu/tcd/ssa/guide/users.guide4.start.html#soi) Here we illustrate how **Toolkit** can be used to identify and reconstruct the low-frequency quasi-quadrennial (QQ 4yr) and quasi-biennial (QB 2yr) components of El Niño (Rasmusson et al., 1990).

What we will learn:

- Read data from files
- Use Toolkit analysis tools (**SSA, BT-FFT, MEM and MTM**) to identify and reconstruct oscillatory components in the noisy dataset.

Getting Started:

Double click **Spectra** and the main Spectra window will appear:

There are five pull-down menus, plus a `Help' button for a brief description of each.

File/Data: used for reading in the dataset to be analyzed **('Read Vector',and 'Read Matrix' functions),** and for writing out results created by the toolkit **('Write Vector' and Write Matrix' functions).** It also enables internal data management operations, such as selecting a column vector from a matrix **('Matrix/Vector'),** and selecting a submatrix from a matrix **('Matrix/Matrix').**

Analysis Tools:

Blackman-Tukey Correlogram
Maximum-Entropy spectrum estimation
Multi-Taper Method spectrum estimation
Singular Spectrum Analysis (SSA)
Multi-Channel Singular Spectrum Analysis (MSSA)

Plot: provides useful plotting functions.

User can plot a selected vector vs. its index using **'Vector'.**
'VectorList' creates a plot of a list of vectors of equal length. **'MultiVector'** creates a plot of two pairs of vectors **(for x and y axis). 'Matrix'** creates a plot of selected matrix columns. **Using 'Plot Options'** user can change the graphics output for ALL plotting functions inside the **Toolkit.** Currently the **Toolkit** supports **IDL, Grace (default!), and ACE/gr (Grace is descendant of ACE/gr)** plotting packages**.**

Logfile: opens the internal logfile containing the detailed output from all tools. It forms an integral part of the Toolkit's output. The Logfile can be saved to a local file 'Logfile' on disk via 'Save' button.

Before applying any of the Toolkit tools, the file containing the time series to be analyzed must be read in. Data should be in the form of ascii columns, and can be either a column vector of a single time series, or a matrix of several columns, one for

each time series. The **'Read Vector'** or **'Read Matrix'** functions from the **File/Data** menu are used to read the data. Values in the time series must be equally spaced in time.

Now go to **File/Data->Read Vector.**

This window allows you to browse a list of files, select a file and then read it and store in a named vector. SSA-MTM Toolkit allows you to handle many time series at once; these might typically be a raw input time series and a filtered series derived using the Toolkit. These time series and results of the computation are stored internally as vectors or matrices, each of which needs to be given a name.

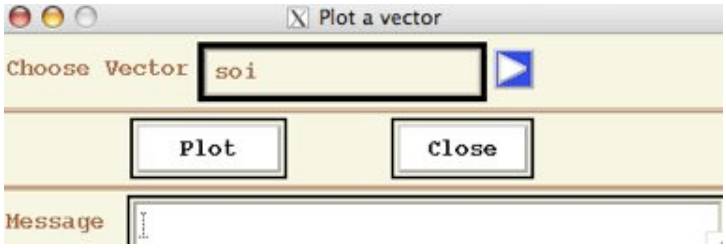
To choose the input time series for analysis, either:

scroll down to the name of the input time-series file, select it with the mouse and press the **`Read File'** button, or else type the input file name in **`Selection'** and press the **`Read File'** button. The input vector is given name `data', by default. If additional time series are read, the user needs to change this name, or the vector with the name 'data' will be overwritten.

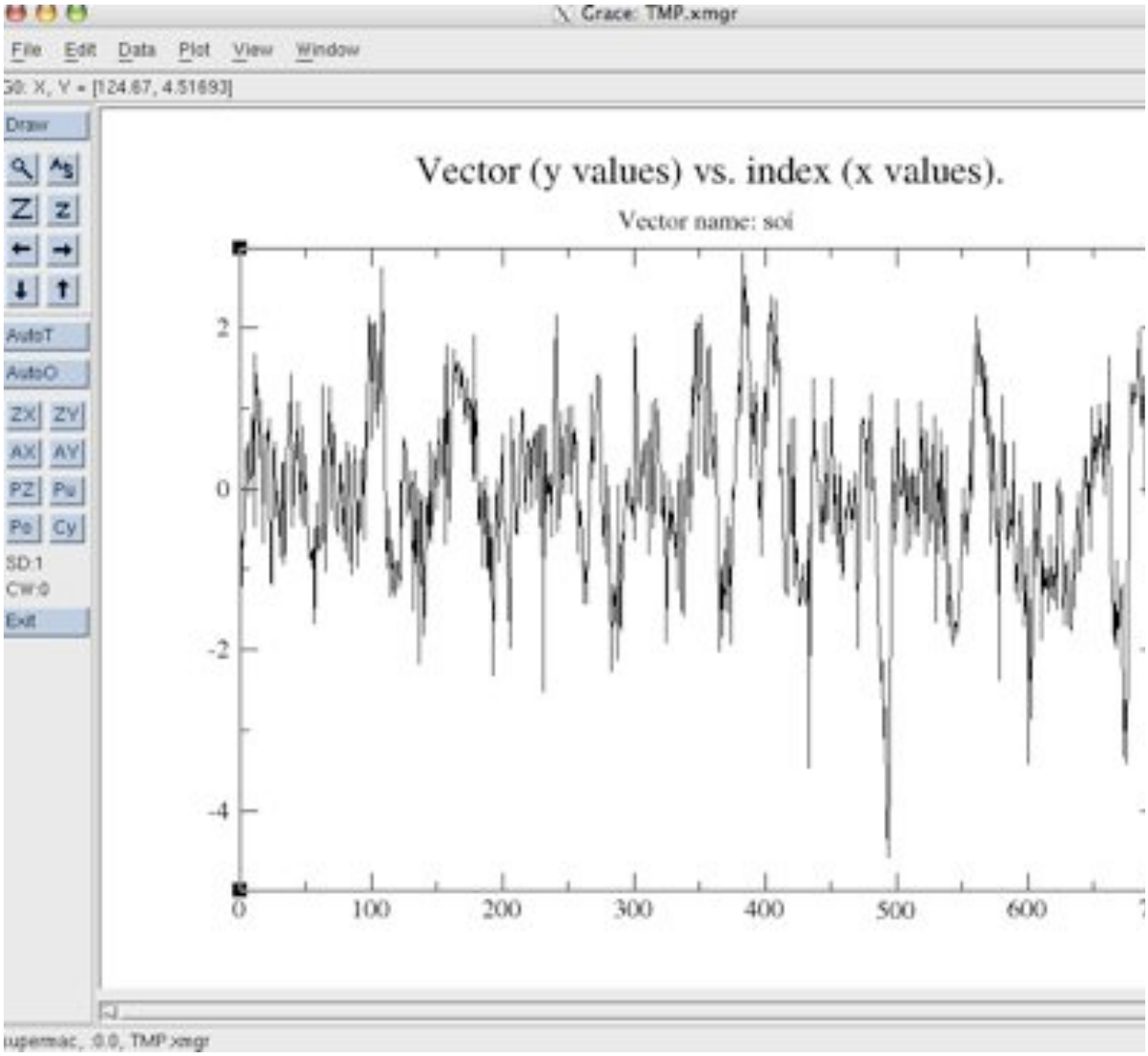
Navigate to the folder S**ummer_School/SOI** folder, select file soi in **Files** panel, change name in **"Store data in vector"** field to soi, and click **Read file**.

The user has access to vectors and matrices via window lists, which show the names and dimensions i.e. the number of elements in a vector and number of rows and columns for a matrix. List windows are activated by pressing on the arrow buttons located next to the boxes with the name of the vector or matrix:

To make a plot of the time-series data, we go to the **'Vector'** function in a **Plot** menu, which opens the following window:



This window allows the vector elements to be plotted against their index. The user has to choose the name of the vector from the vector list by clicking an arrow button. Select **soi** vector, and by pressing on the **'Plot'** button, a new graphics window opens with a plot of SOI time series:



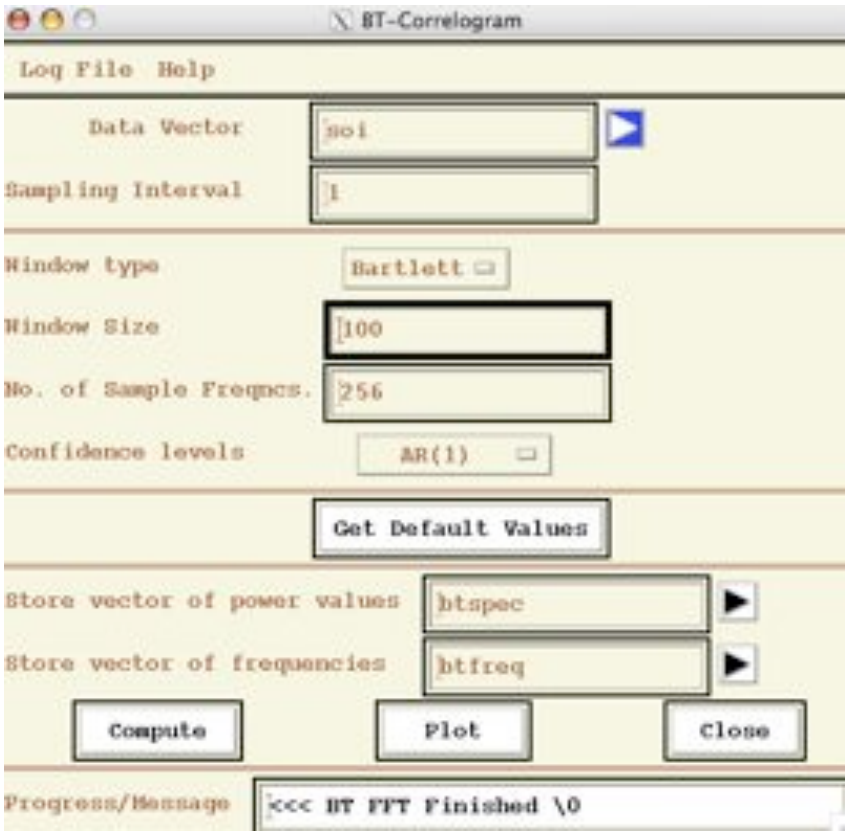
Go to the **Plot->Plot Appearance** to change the title of the plot.

To save it to the postscript file go to **File->Print Setup->Print to File.**

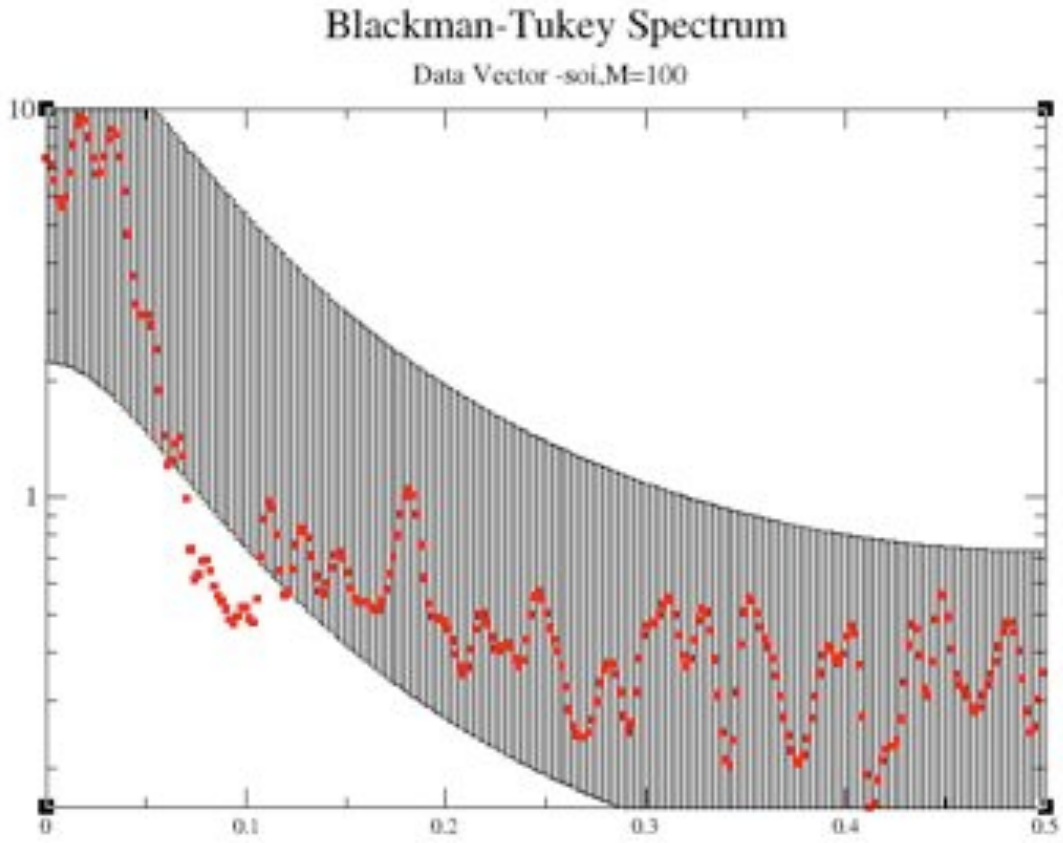
Go to **B-T Correlogram** in **Analysis Tools,** select **soi** from **Data Vector** field by clicking to the arrow next to it, click **Default** to set default parameter values, set **Confidence levels** to **AR(1),** change the **Window** value to 100.

The sampling interval is assumed to be unity by default. If it is not unity, each tool needs to be told individually what the sampling interval is using the "Sampling Interval". The resulting spectra are then plotted accordingly. If, for example, the data are sampled every 2 months instead of monthly, the Nyquist interval on the frequency axis will be labled from 0 to 0.25 cycles/month.

BT-Correlogram:



Now click **Compute** and **Plot.** The power spectrum displayed in the graphics window shows peak power between frequencies of zero and about 0.05 cycles per month. These results indicate presence of oscillatory peaks in low-frequency part of the spectrum, but we have to use other tools in order to confirm this conclusively.



The smaller the window size, the more independent samples are obtained for estimation purposes and therefore the smaller will be the variance of the spectral estimate. However, the smaller the window size, the lower the spectral resolution of the estimate will be. Note that the spectral resolution is independent of the number of sample frequencies.

Robustness of results to changes in window type and length is the simplest test of their validity.

1st Task:

Repeat the **BT Correlogram** computation with the **smaller window size (40 months)** and save results with different name (i.e. **btspec40** for the power values). Compare results by going to **Plot->Multivector** plot:

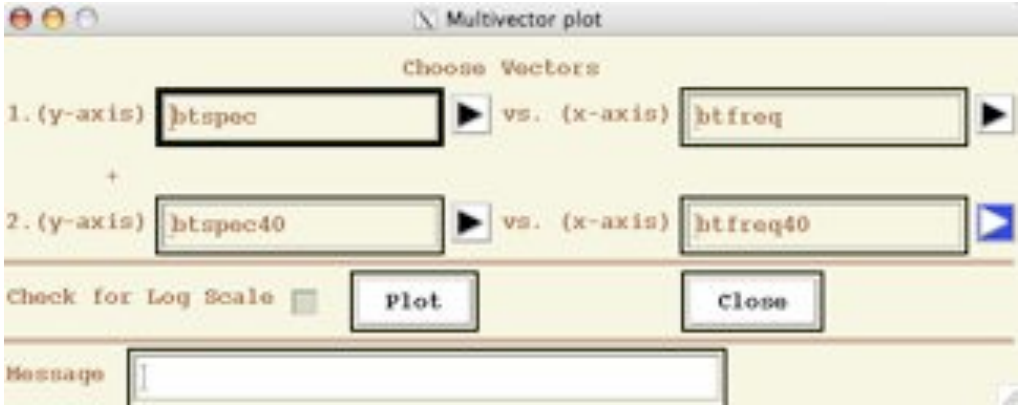
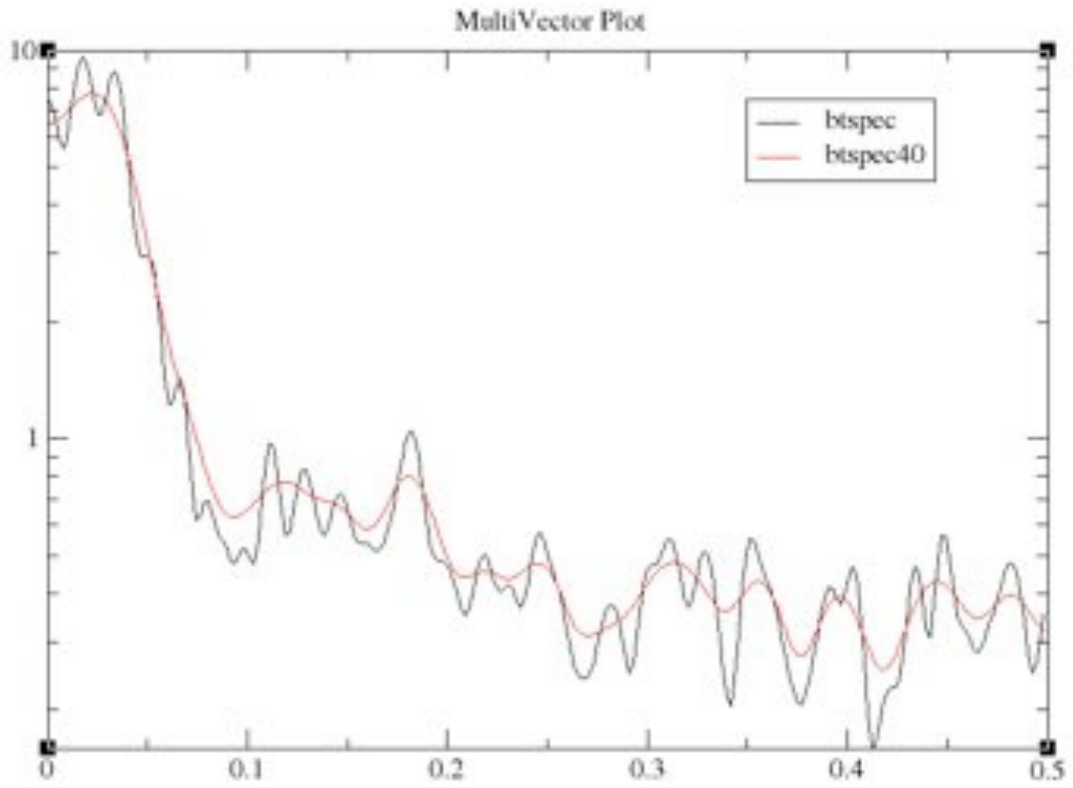
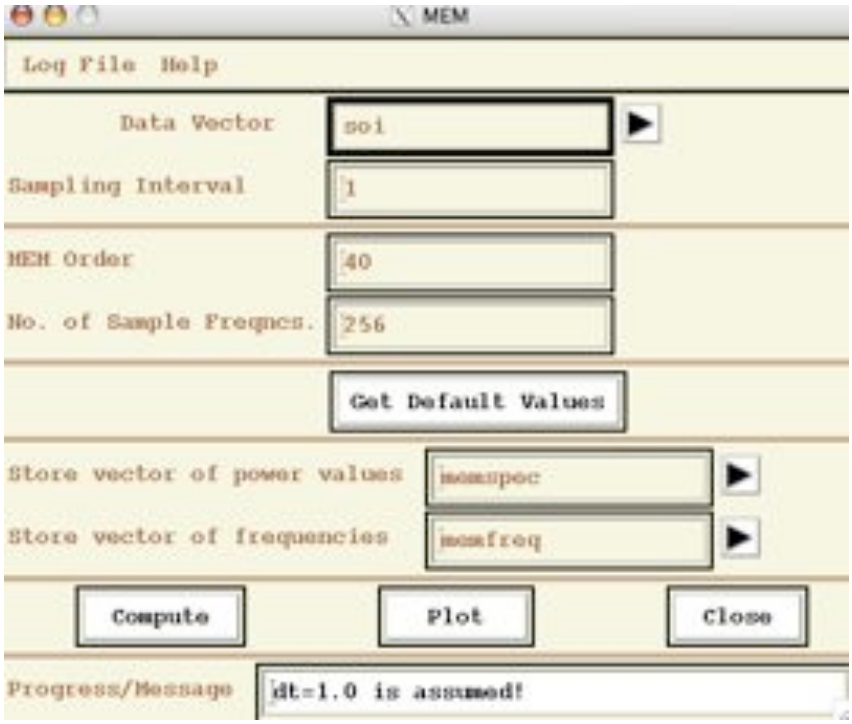


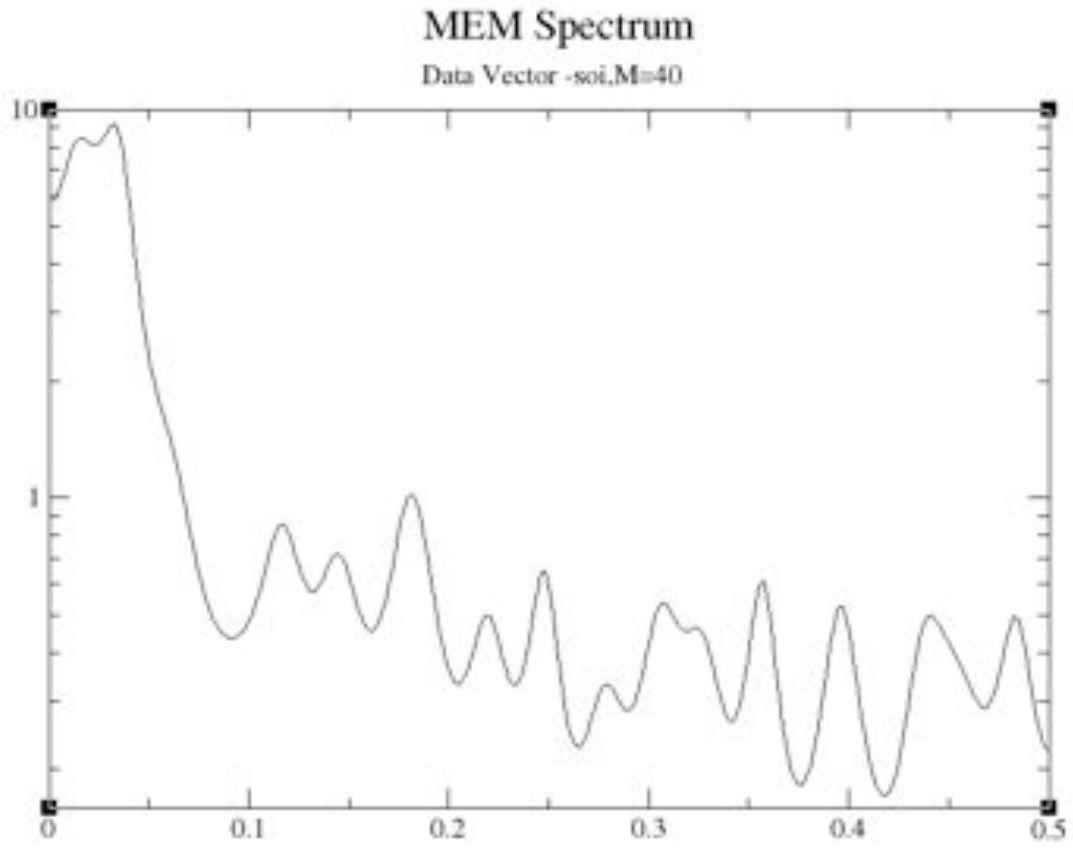
Figure 1 (print to a file):



Maximum Entropy Method:



Click **Default** to set default parameter values, then **Compute** and **Plot.**



Like the B-T correlogram , the MEM spectrum for the SOI shows strong peak between 0 and 0.03 cycles/month with hints of two separate peaks. It is also accompanied by a large number of much smaller peaks at higher frequencies, that are spurious. The order of the MEM is the number of AR components (or poles) to be included in the analysis, and the number of spurious peaks usually grows with the MEM order. Robustness of results to MEM order is the simplest test of their validity.

2nd Task:

Repeat the **MEM** computation with the **MEM order equal to 100** and save results with different name (i.e. **memspec100** for the power values). Compare results by going to **Plot->Multivector** plot:

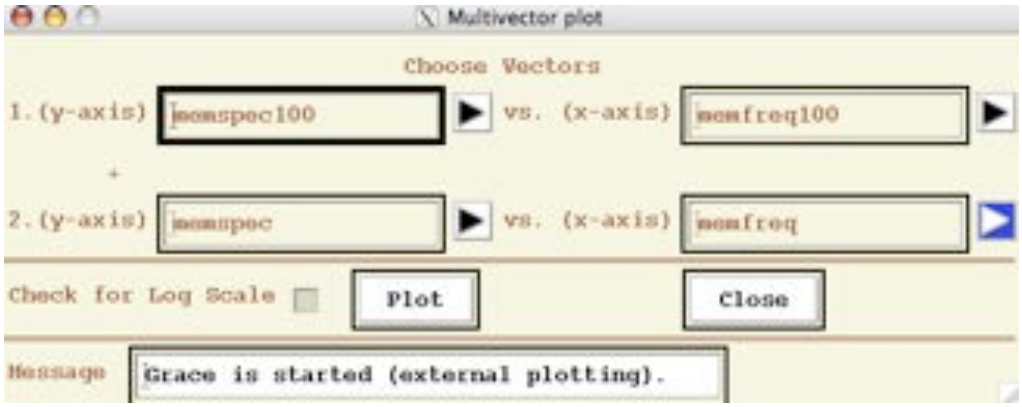
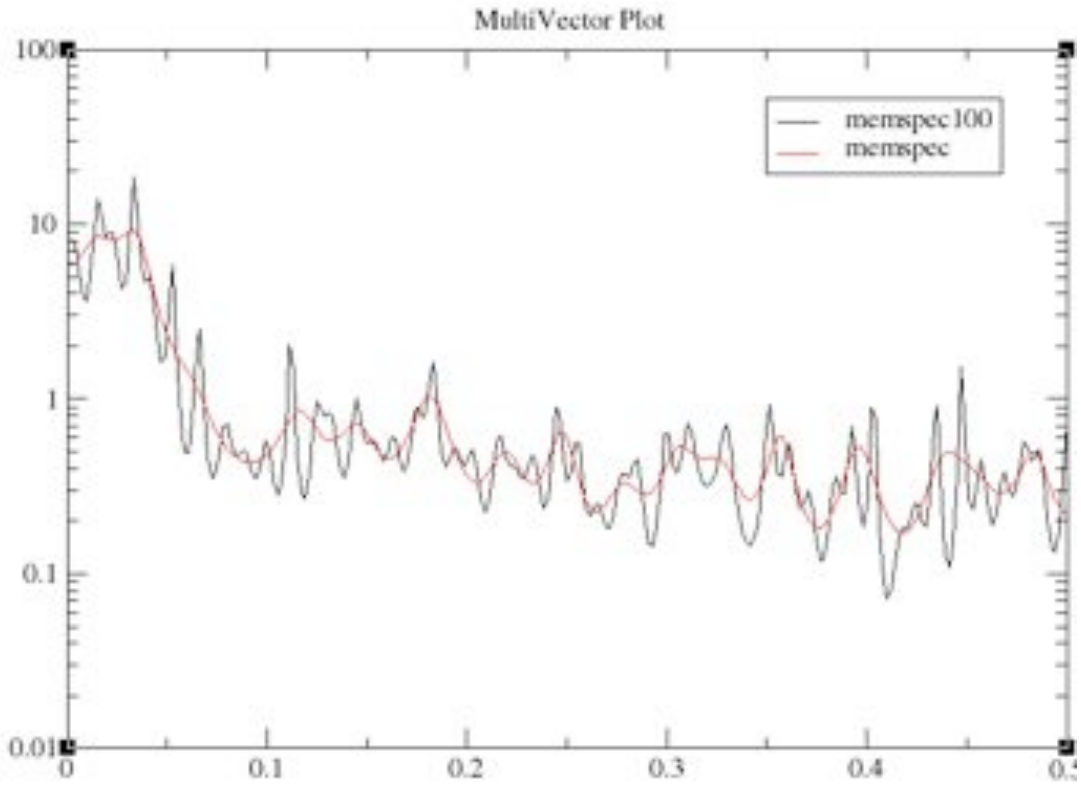


Figure 2 (print to a file):



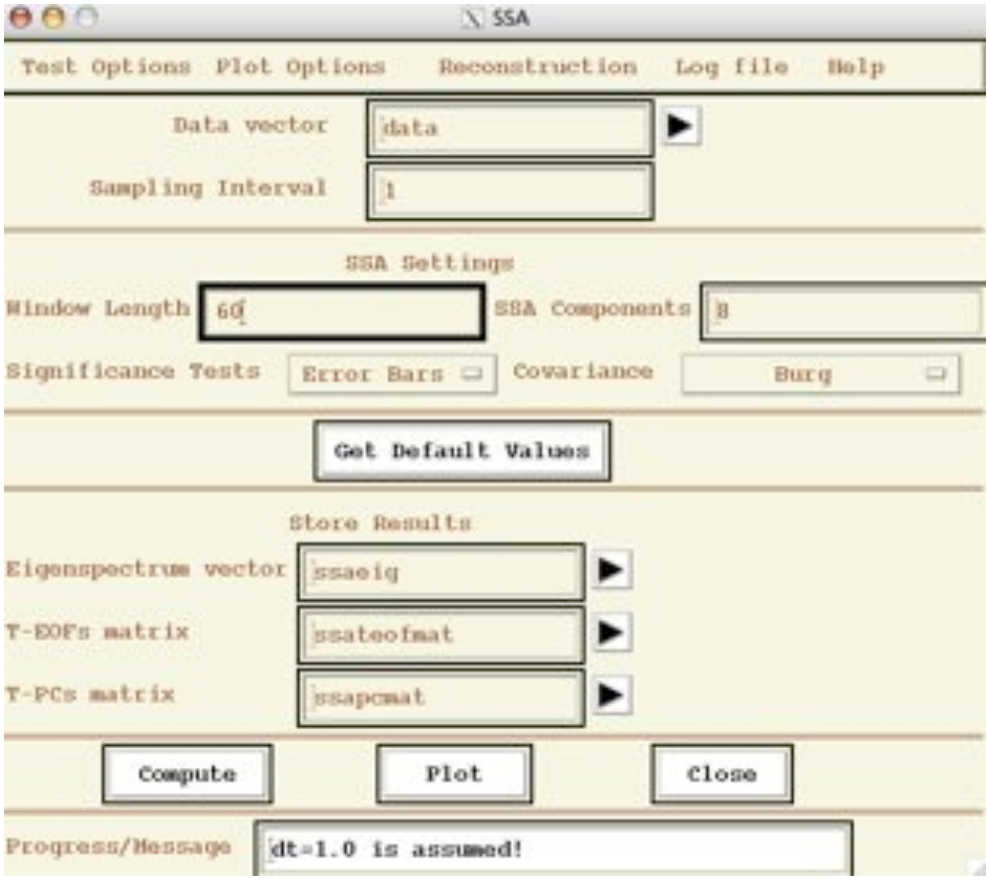
SSA: scree diagram, pairing & trend tests.

Go to **SSA** in **Analysis Tools,** select **soi** from **Data** Pop-up menu, click **Default** to set default parameter values.

As a rule of thumb, the **window length** M should be chosen to be longer than number of data points in the oscillatory periods under investigation, and shorter than number of data points in the spells of an intermittent oscillation. Vautard et al. (1992) recommend that the window length be less than about N/5 where N is the number of points in the timeseries.

Robustness of results to M is an important test of their validity.

The choice of window length sets the dimension of the lag autocorrelation matrix to be constructed and diagonalized by SSA, and thus determines the computational burden of the application. Larger values of M correspond to higher spectral resolution, although there is no direct equivalence between them. We will set the **Window Length** to 60, which is a good choice for out time series (690 data points with a one month sampling rate) and the oscillatory periods (2 and 4 years) under investigation.



The **Covariance Estimation** method for estimating the autocovariance matrix that is decomposed (diagonalized) by SSA is chosen by selecting either **`Burg'** , **`Vautard -Ghil'** , or ` **Broomhead &King** ' from the **`Covariance'** option on the main SSA panel. Both the Burg and Vautard- Ghil methods impose a Toeplitz structure upon the autocovariance matrix whereas the Broomhead & King method does not. The Toeplitz methods also impose symmetries on the EOF shapes whereas the Broomhead & King method does not.

Burg estimation is an iterative process based on fitting an AR modelwith a number of AR components equal to the SSA window length, and in principal should involve less "power leakage" due to the finite length of the time series and should therefore improve resolution (Penland et al., 1991). However, Vautard et al. (1992) found that the Burg estimate can induce significant biases when nonstationarities and very low-frequency variations are present in the series.

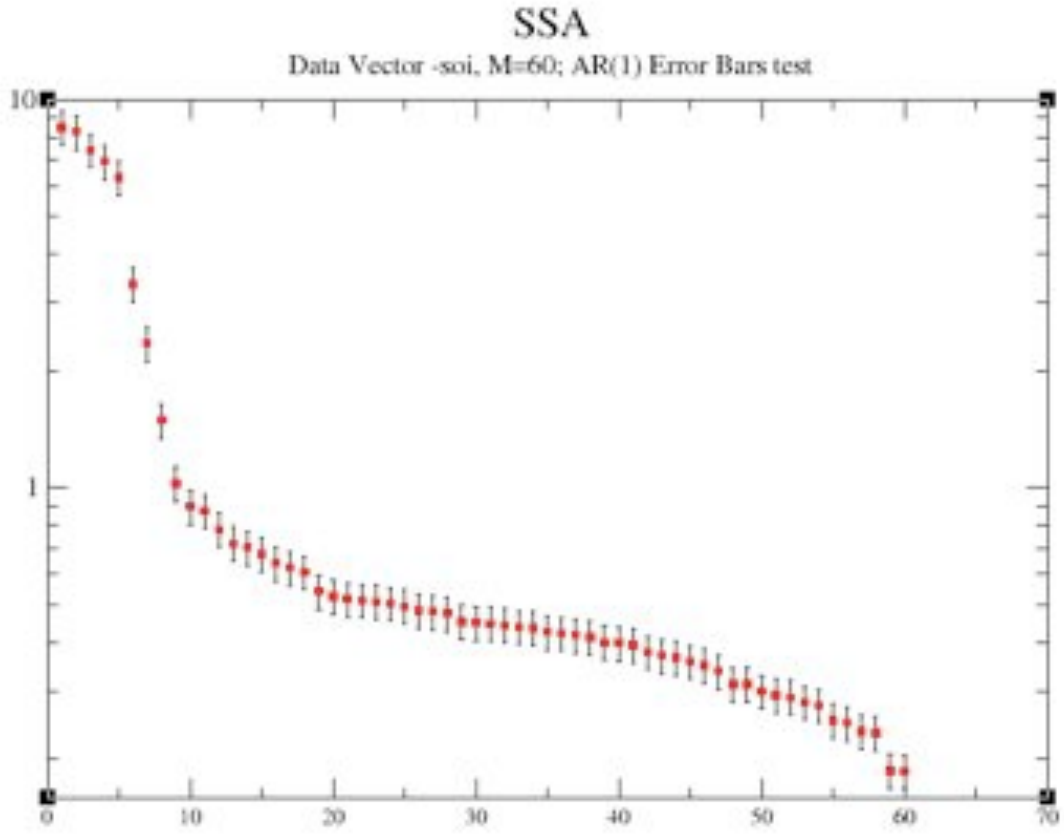
Go ahead select **Vautard&Ghil** covariance method for analyzing SOI time series.

There are three choices for Significance test:

Error Bars
Chi-Squared
Monte Carlo SSA

If **Error Bars** are selected, the eigenspectrum is displayed in order of eigenvalue rank. For the **Chi-Squared** and **Monte Carlo SSA** tests, the spectrum of eigenvalues is plotted against the dominant frequency associated with the respective T-EOF.

Select **Error Bars,** and them click **Compute** and **Plot** buttons to display the eigenvalue spectrum of the specified SSA. Since the window length was set to 60, SSA decomposes the time series into 60 components and thus 60 eigenvalues are plotted.



The significance of the various components can be judged qualitatively by noting which components contribute significantly more variance relative to the noise background. The latter in turn assumed to include the components that lie in the flattish tail of the eigenvalue spectrum, i.e. components from about 10 to 60. The leading 10 components in the plot lie above a distinct break in the eigenvalue spectrum, and thus may be of physical significance. In particular, we are interested in the leading four components which form two pairs of nearly equal eigenvalues.

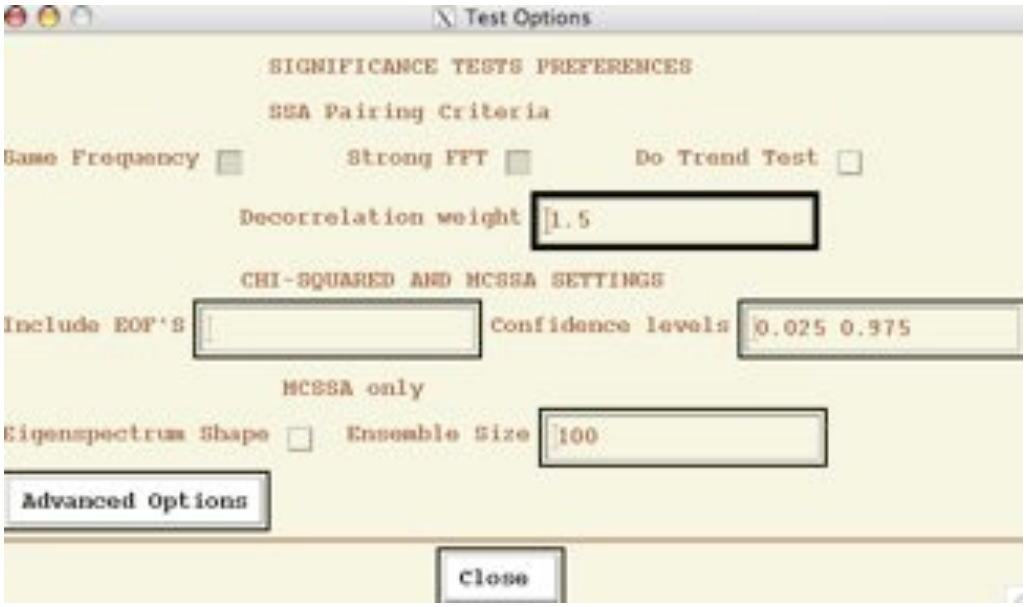
The percentage of variance captured by the SSA components is written out to a temporary file test_pct.out (1st column is the rank, 2nd column is %).

"Simple" SSA significance tests

Pair of nearly equal eigenvalues in SSA is one of the characteristics of an oscillation. However, the eigenvalues are subject to numerical and sampling errors, and mere pairing of eigenvalues is not enough to guarantee that an oscillation has been identified. In the eigenvalue plot above, the error bars show an ad hoc range of the estimation errors. Any of the eigenvalues with significantly overlapping error bars could represent an "oscillatory pair'. Also eigenvalues with error bars that overlap significantly with the error bars of the noise part of the spectrum must also be suspected of being part of that noise.

The Toolkit identifies potential oscillatory and trending components using a few pairing criteria which can be activated

concurrently using checkboxes in **Test Options.**



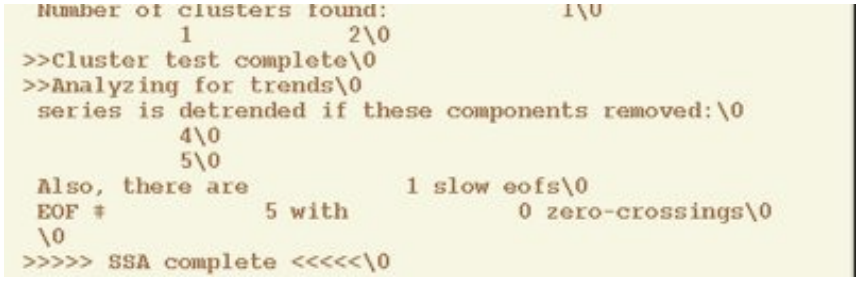
For **`Same Frequency'** test, the T-EOFs associated with potential pairs or clusters are subjected to a simple Fourier transform to identify the their dominant frequency. A pair (or cluster) is identified as an oscillatory one when the associated T-EOFs have the same dominant frequency, within a fraction of the SSA bandwidth.

For **`Strong FFT'** test, the same Fourier transform is used to determine how much of a given signal the potential oscillatory pair accounts for at their dominant frequency. This variance fraction must exceed 95% for a pair (or cluster) to be identified as an oscillatory one.

When Do Trend Test box in SSA Advanced options panel is checked, two tests are performed that help to identify trending SSA components, up to the maximum number set in Components on the main SSA panel:

- Kendall's tau nonparametric trend tests; the component is labeled as "trend" if detected at 95% significance in both T-PCs and RCs.
- The numbers of zero crossings within the T-EOFs are counted; those with 0 or 1 crossings are considered to be trending components.

Go to **Test Options,**check **`Same Frequency'** and **'Strong FFT'** boxes, run the SSA again with the different window sizes (M=40--60) and then check in the **Log file** that pair 1-2 is robust for meeting both criteria, while the component 5 is the trend (for M=60).



Task 3:

For phase-quadrature test to check that **T-EOFs** for the two leading pairs are in phase-quadrature (shifted by a quarter of a period), go to **Plot Options,** input **1 2** or **3 4** in the **Components** field and click **plot T-EOFs.**

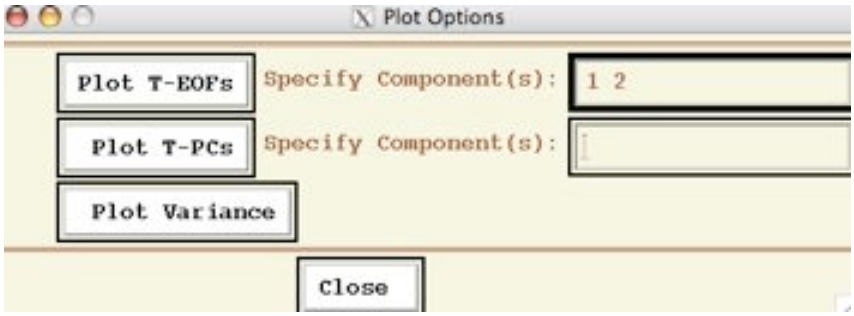
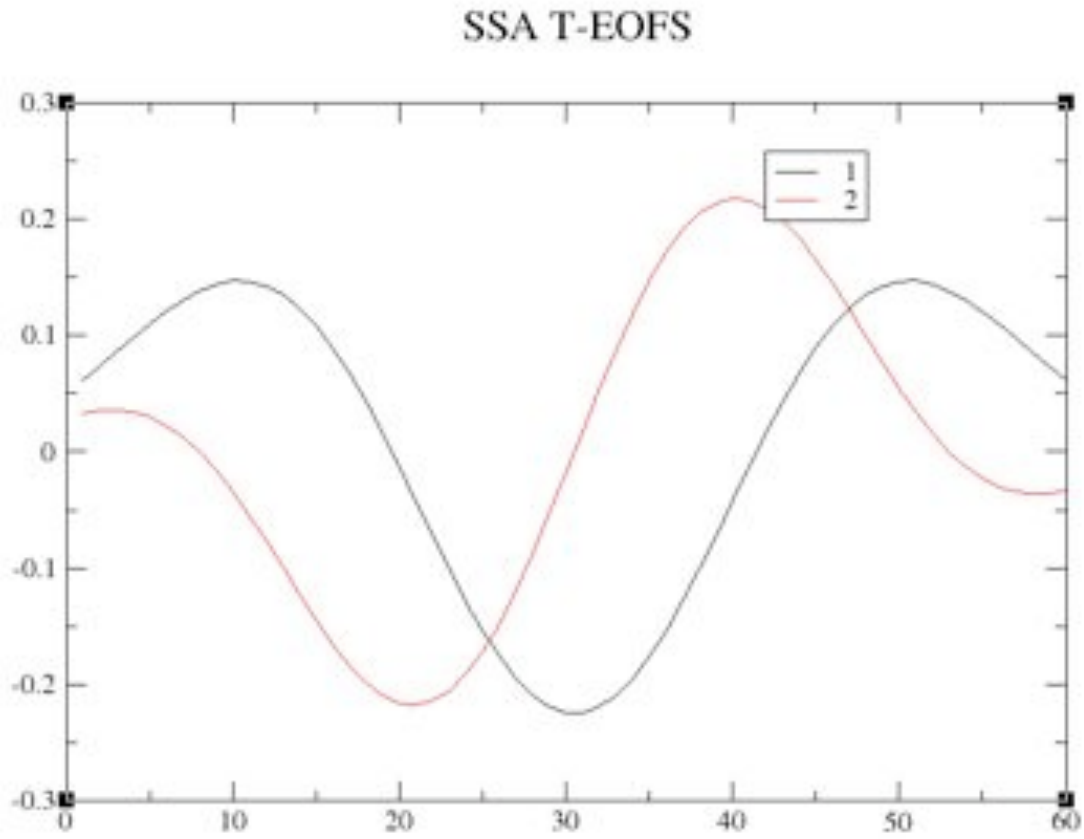


Figure 3 (print to a file):



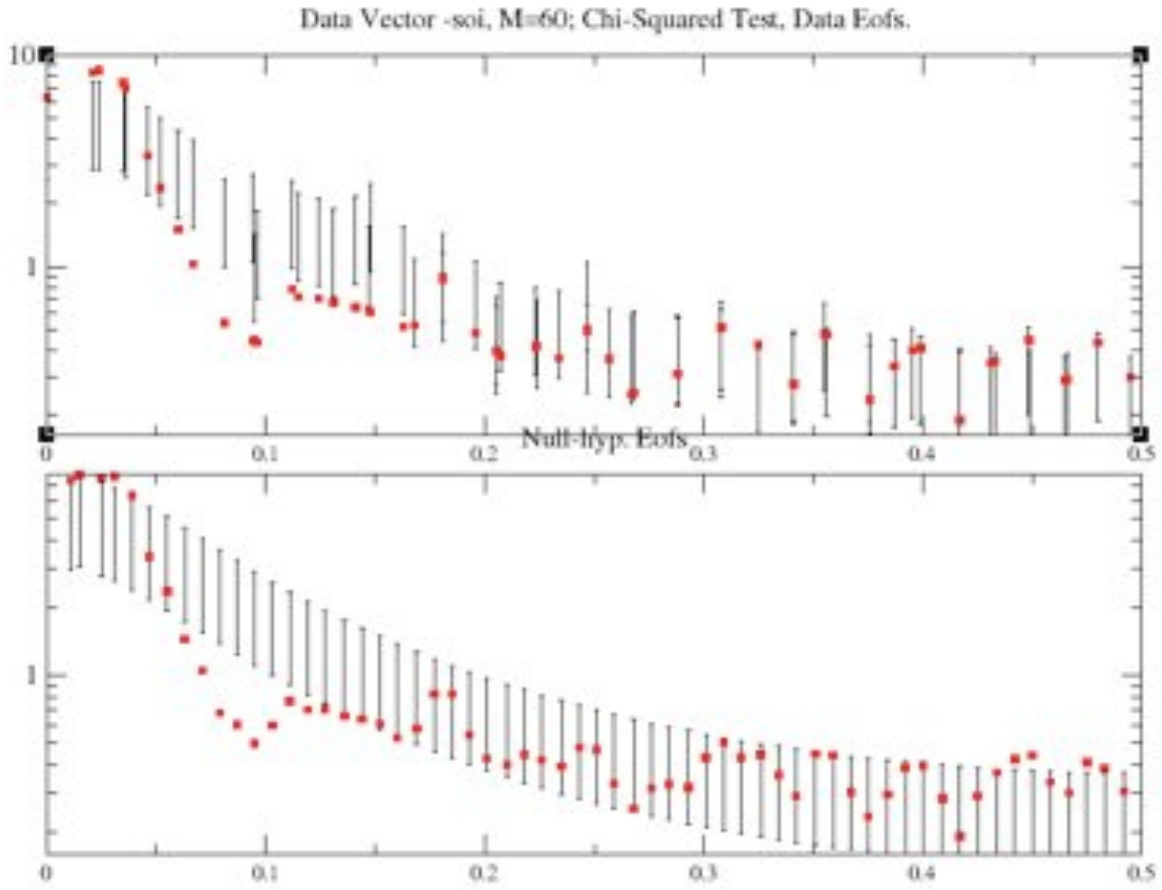
Monte-Carlo SSA significance test

Monte Carlo SSA pairing is an advanced test. Here a large ensemble of **red-noise** surrogate time series is generated, each with the same length and same expected lag-1 autocorrelation as the time series to be tested.

First, the original time series, together with many AR(1) noise realizations are projected onto the EOFs of the data covariance matrix, i.e. the **"data basis"**. Second, the data, together with many AR(1) noise realizations are projected onto the EOFs of the expected covariance matrix of pure noise. This is called the **"null-hypothesis basis"**.

Chi-Squared test is an approximation of a Monte-Carlo test and is computationally fast.

To test against a red-noise null-hypothesis, we choose **Chi-Squared** as a significance test in the **Test Options**. Then click **Compute** followed by **Plot**.



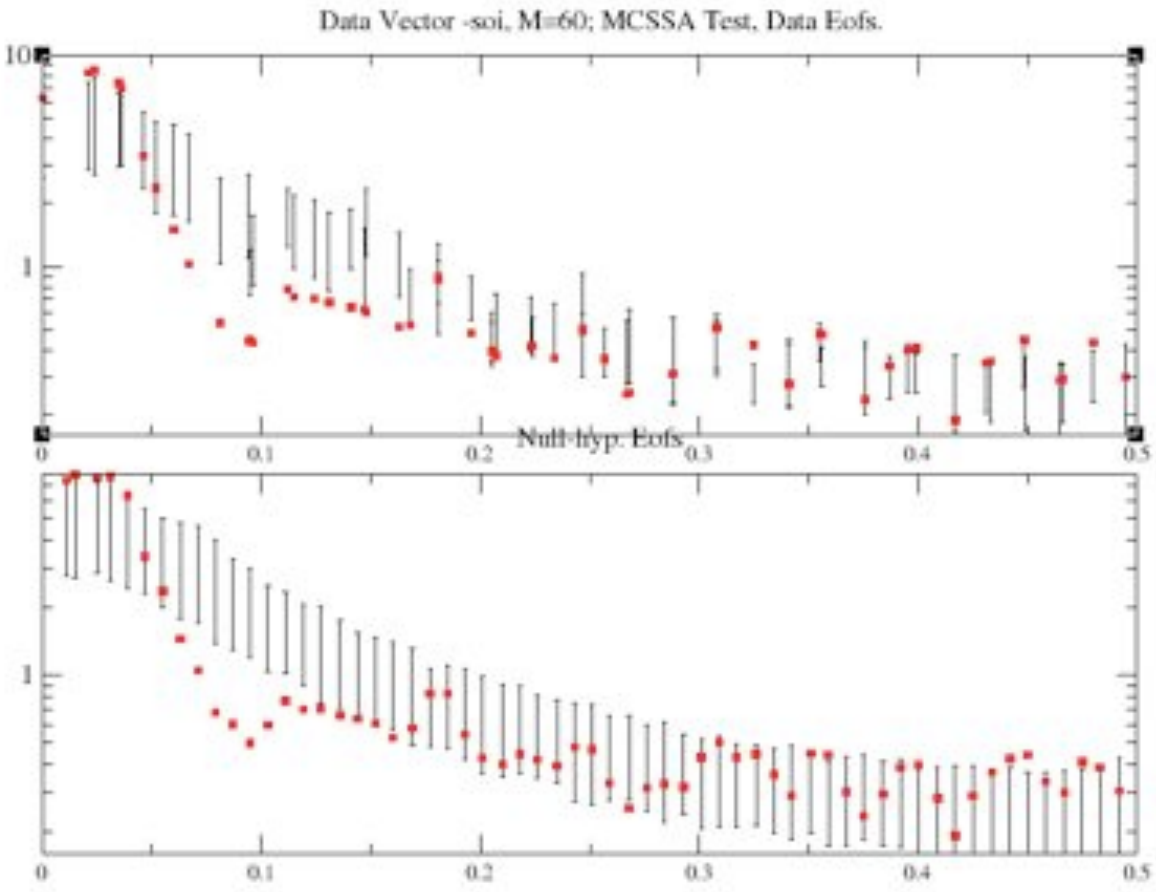
For the **Data Eofs** basis we observe that eigenvalues corresponding to **EOFS 1-2** and **3-4** appear almost superimposed on each other at frequencies equal **to ~ 0.02** and **0.034** cycle/month, and lie outside the null- hypothesis error bars. Thus they are relatively unlikely (at the **95% level**, with settings of **Confidence levels** in **Test Options**) to be merely due to selected null-hypothesis process, and represent two signficant oscillations. These two leading oscillatory pairs correspond to the two low-frequency oscillatory ENSO modes, with periods of 2 and 4 years. **The Null-hypothesis (NH) red-noise basis** (lower graph) confirms the significance of these pairs. Note how the EOFs are almost regularly spaced by frequency. The **NH basis** avoids artificial variance compression inherited in SSA and therefore it has a lower probability of false-positive results, i.e. identifying the noise components as significant.

Task 4:

The results of the **Chi-squared test** should always be checked using the Monte Carlo approach, which is also essential with more complex noise models. Set **Monte Carlo** as **Significance** test.

Click **Test Options** and change **Ensemble size** for the number of Monte-Carlo red-noise realizations to 200, then click **Compute** followed by **Plot**. Check that similar results are obtained as for the **Chi-squared** test.

Figure 4 (print to a file):

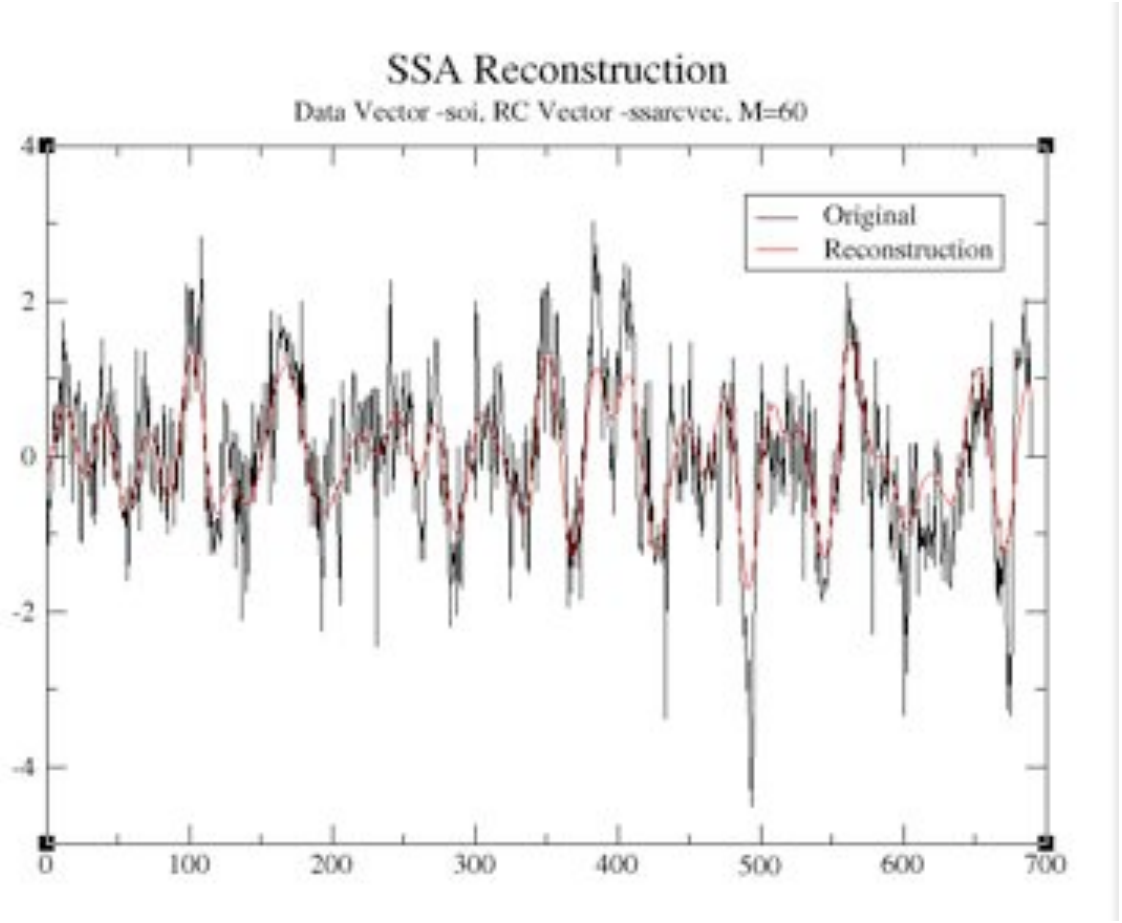


SSA Reconstruction.

Click **Reconstruction on the main SSA panel,** and set **"1 2 3 4"** in **Specify Components.**

Set the name of RC-Vector to **"ssarcvec"**. Then click **Reconstruct**, followed by **Plot**, to obtain figure with the reconstructed

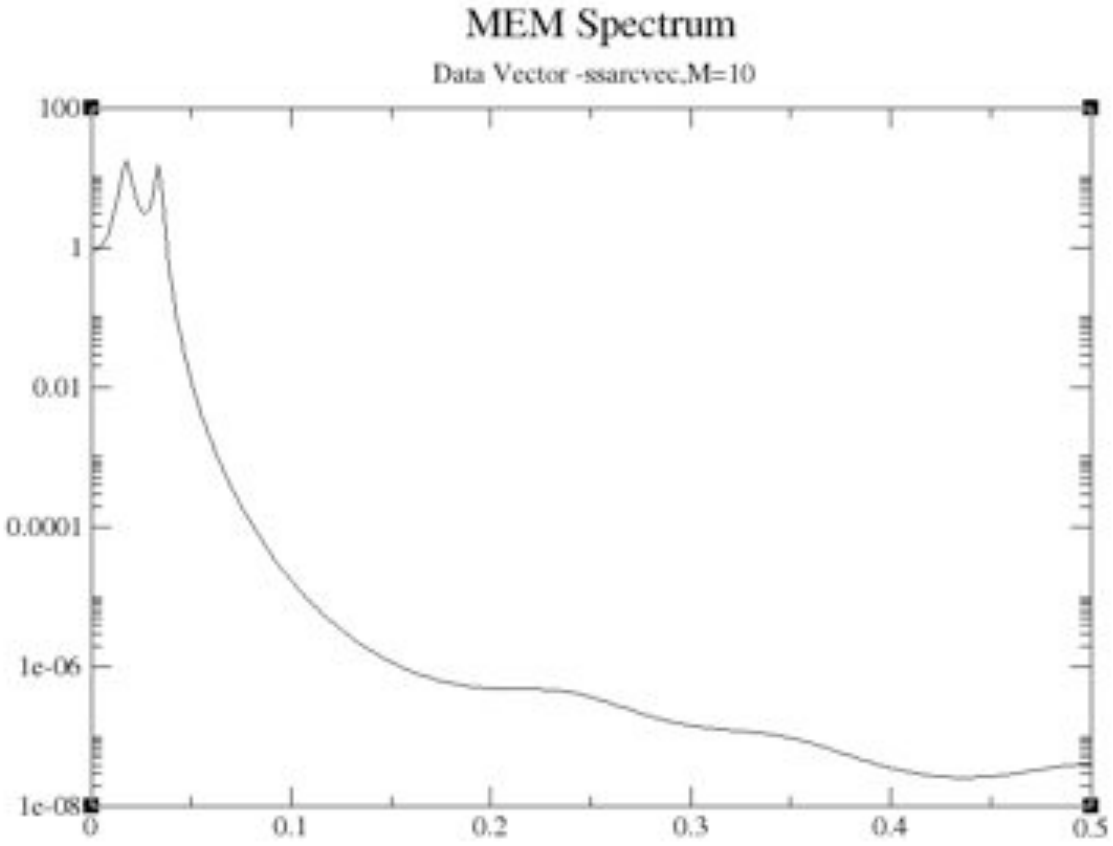
signal plotted against original timeseries.



Task 5:

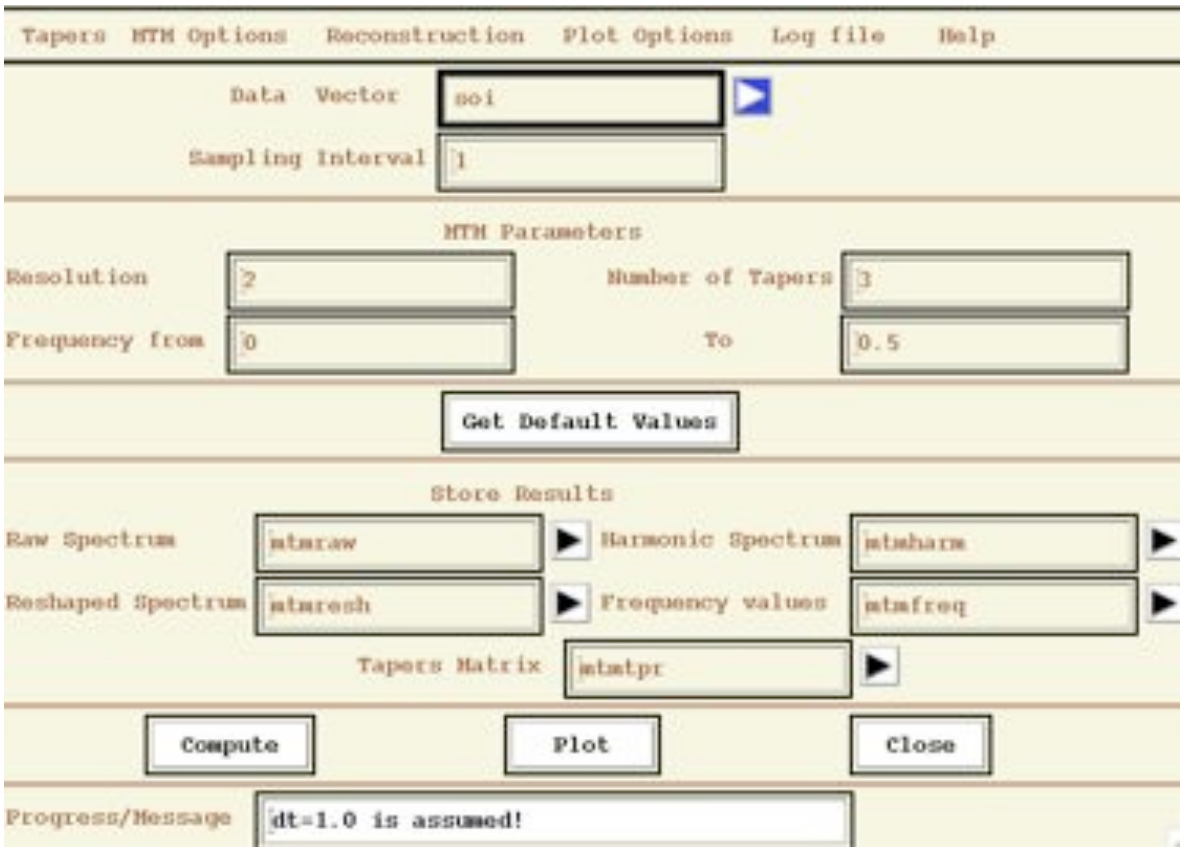
If this SSA-filtered RC series (**'ssarcvec' vector**) is now analyzed by **MEM**, the frequency spectrum of the RCs is much simpler than that of the complete time series, since most of the "noise" has been removed. Go to **Maximum Entropy** of **Analysis Tools**, and use order **M=10** for the **"ssarcvec"** series; MEM spectrum suffices to separate clearly two distinct peaks at 0.018 and 0.034 cycles/month, which correspond to the QQ and QB components of El Niño. There are now no spurious peaks and the signal-to-noise ratio considerably increased.

Figure 5 (print to a file):



Multi-taper Method.

Go to **MTM** in **Tools,** select **soi** from **Data** Pop-up menu, click **Default** to set default parameter values.



Recall that MTM attempts to reduce the variance of spectral estimates by using a small set of tapers. A set of independent estimates of the power spectrum is computed, by pre-multiplying the data by orthogonal tapers which are constructed to minimize the spectral leakage due to the finite length of the data set.

`Resolution'** field resets the resolution half-bandwidth **delta f = p f_Rayleigh** (where f_Rayleigh=1/(N*dt) is the minimum possible spectral resolution) from its default value p=2. Entering a new value in the **`Number of Tapers' field resets the number of windowing functions used in spectral estimation. This value cannot exceed the 2p-1 where p is the integer entered in the `Resolution' box. A lower value can however be set for `Number of Tapers' if the user wants to be especially conservative regarding potential spectral leakage bias.

The selected frequency range (**`Frequency from' .. `To' .. '** values) allow one to change the frequency range from its default value **f=0 to f=f_Nyquist=0.5/d**t, where dt is the sampling interval entered by the user). It will determine the frequency interval over which other MTM options will operate.

Then go to **MTM Options.** The main options for significance tests to set there are **``Null hypothesis''** and **"Signal Assumption":**

There are three possible choices of ``**Null hypothesis**'':

red noise (default)
locally white
white

The choice of ``**red noise**'' assumes a noise background that consists of a temporally integrated Gaussian white noise or ``AR(1)'' noise process. This null hypothesis is strongly motivated for dynamical reasons in the study of geophysical phenomena, and represents the default option of the Toolkit.

The choice of ``**locally white noise**'' assumes a colored noise process that varies slowly but arbitrarily with frequency. This choice is recommended if there is a priori reason to believe that the noise background has a complex structure.

The choice of ``**white noise**'' represents a good null hypothesis if absolutely nothing is known a priori about the physics or dynamics of the process producing the noise background.

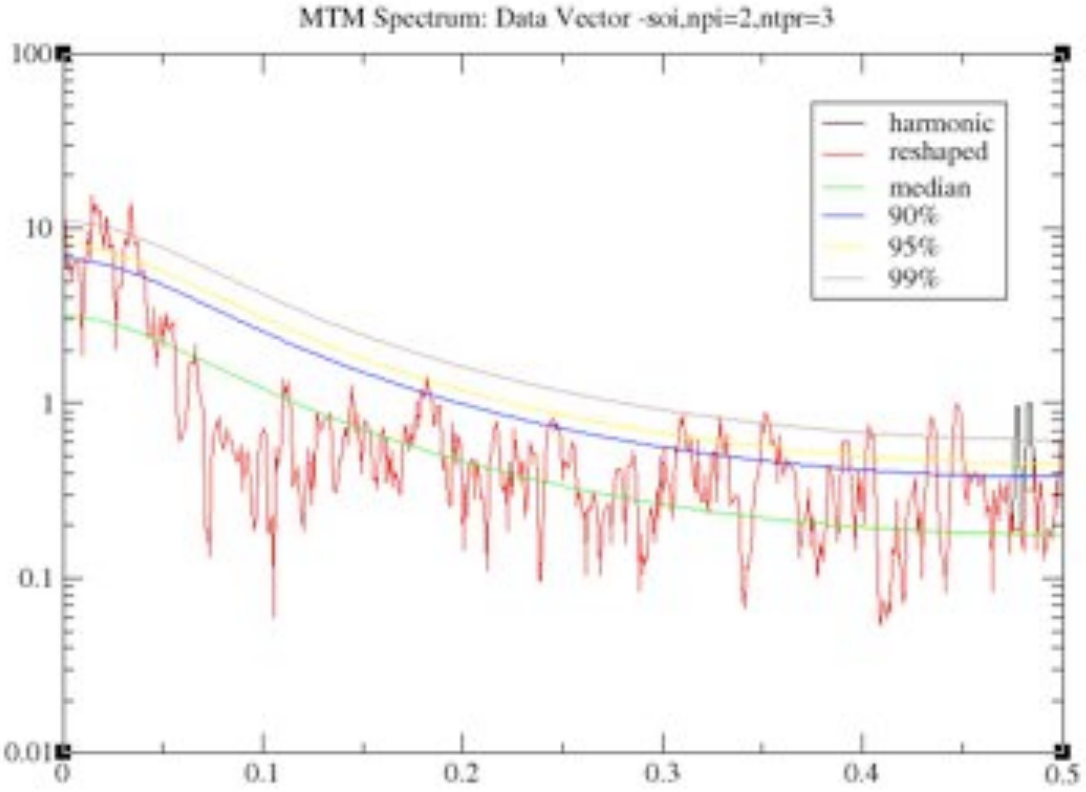
There are three possible ``**Signal Assumptions**'':

The choice ``**either**'' (default choice) indicates that the spectrum should be tested both for the presence of **narrowband** signals whose significance is measured by their amplitude relative to the estimated noise background, and for the presence of ``**harmonic**'' signals which are significant as measured by the Thomson variance ratio test for periodic signals (**F-test** against the white noise). The choice ``**narrowband**'' will test the spectrum only for the presence of narrowband signals (the former), while the choice ``**harmonic**'' will test the spectrum only for the presence of periodic signals (the latter).

The ``**Noise Estimation**'' menu allows two choices of the way in which the noise background is estimated, where the default **robust** option guards against the contamination of noise paramater estimation by narrowband signal and significant trend contributions.

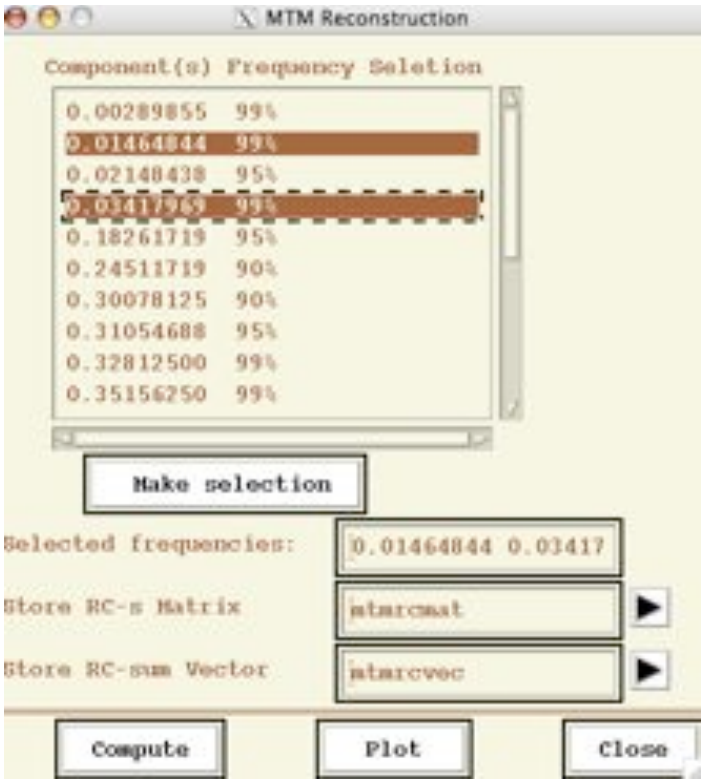
A ``**Reshaping**'' procedure is used to separate the continuous and harmonic portions of the spectrum. In `**Reshape Threshold**'' field the threshold for significance of harmonic peak detection (**90%,95%,99%,99.5% and 99.9%**) in the ``**F-test**'' can be changed from its default setting of 95%. When the **Signal Assumption** option is set to **Either**, the detected **harmonic** peak will be reshaped only if it is **ALSO** signficant to the estimated noise background at a Reshape Threshold level. For the ``**Harmonic**'' option all F-test significant peaks will be reshaped.

To calculate the spectrum of our SOI series, click **Compute** followed by **Plot**. The presence of two significant low-frequency peaks confirms **SSA** findings.

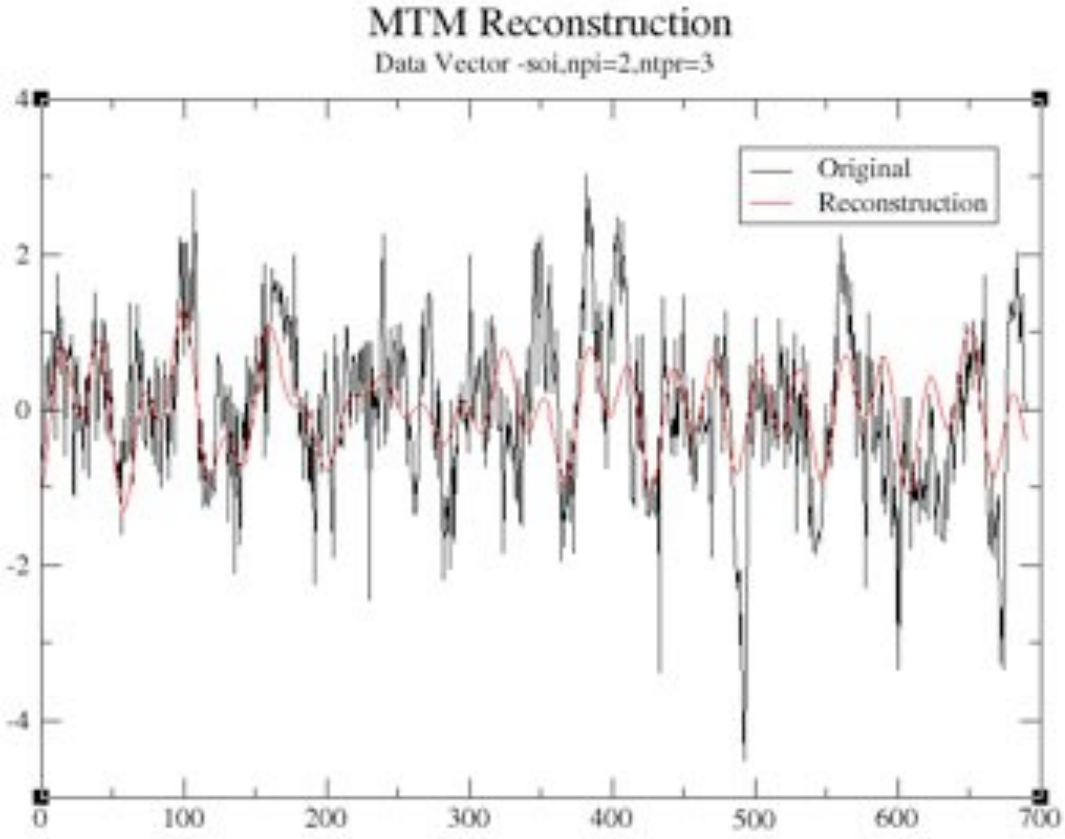


Now go to the **Reconstruction.** If **Signal Assumptions (in Test Options**) is set to **Either or Narrowband**, the **Component(s) Frequency** in **Reconstruction table** will contain a list of the central frequencies of narrowband signals identified as significant at greater than the 90% level relative to the specified null hypothesis (red-noise by default). Our MTM analysis identifies two highly significant peaks, one centered at f=0.0146, and another centered at f=0.0342. The signals are significant at well above the 99% level. We associate thease peaks with the low-frequency LF(band) and high frequency (HF) band ENSO signals, in agreement with SSA results.

Select these signals in the Reconstruction table by the mouse, and then click **Make Selection** button to fill in the **Selected Frequencies** field.



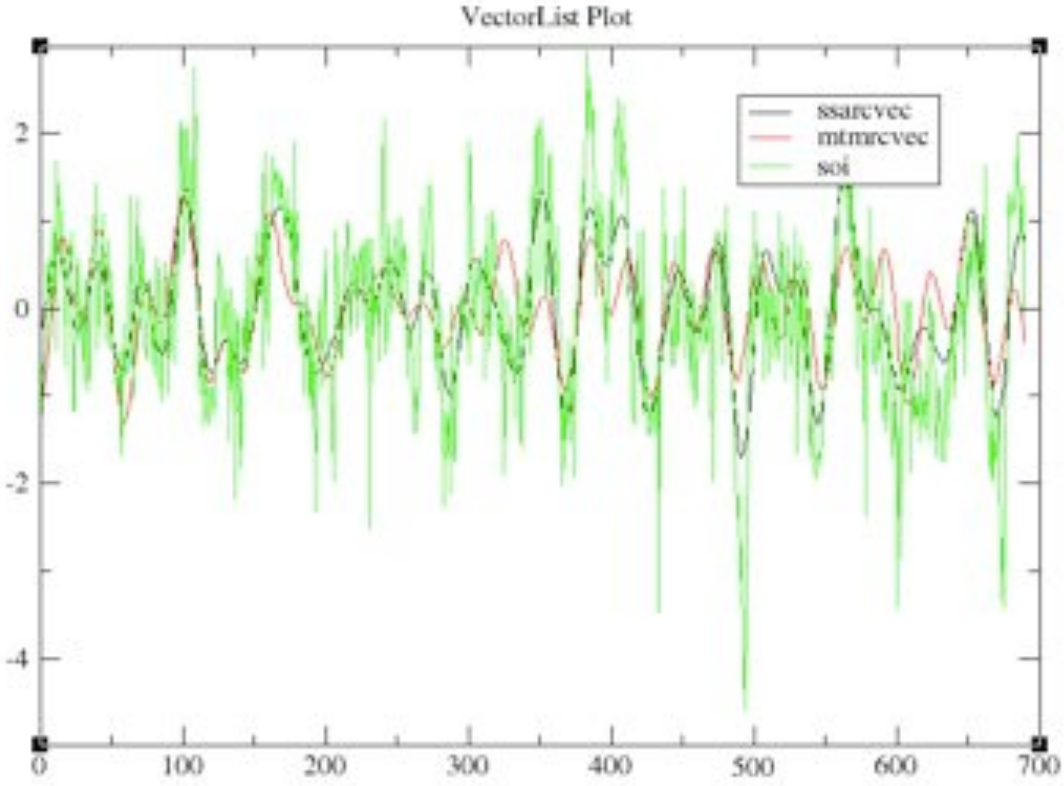
Now click **Compute**. The ``RCs matrix'' will contain the individual RCs, while the ``RC-sum vector'' will contain their sum. The **``Plot''** button plots the RC-sum against the original timeseries, with a mean average removed.



Task 6:

Now compare **SSA reconstruction** with **MTM Reconstruction**, go to **Plot->Vector List,** add **ssarcvec** and **mtmrcvec** in the list

Figure 6 (print a file):


VectorList Plot
ssarcvec
mtmrcvec
soi
2
0
-2
-4
0
100
200
300
400
500
600
700