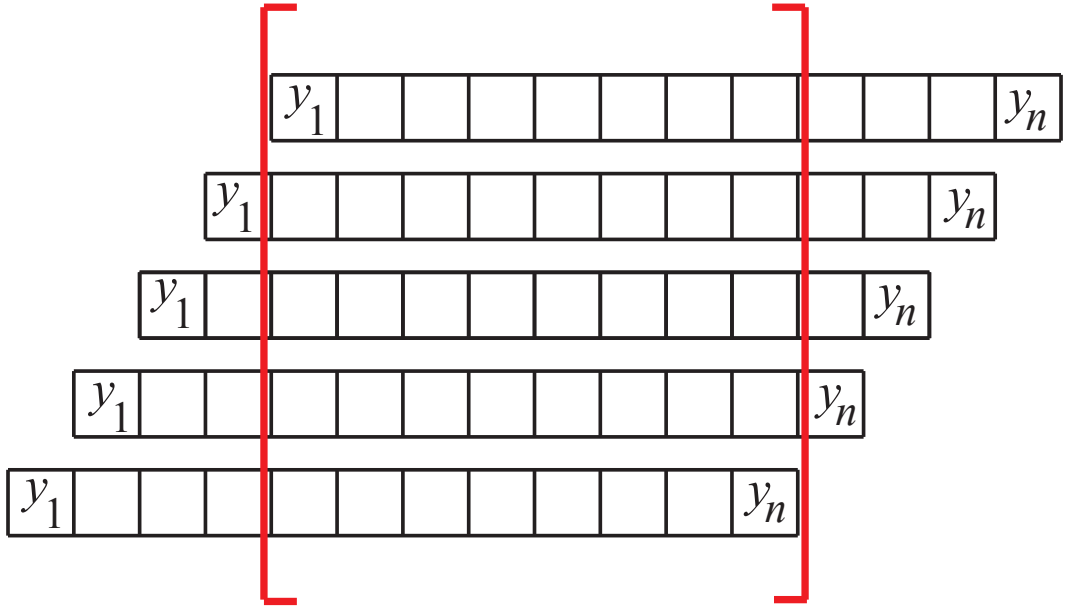
Time series analysis [Book, Ch.3]

4.4 Singular spectrum analysis [Book, Sect. 3.4]

So far, our PCA involves finding eigenvectors containing spatial information. It is possible to use the PCA approach to incorporate *time* information into the eigenvectors. This method is known as *singular spectrum analysis* (SSA), or *time-PCA* (T-PCA) (Ghil et al., 2002).

Given a time series $y_j = y(t_j)$ $(j = 1, \ldots, n)$, lagged copies of the time series are stacked to form the *augmented data matrix* $\mathbf{Y}$,



$$\mathbf{Y} = \begin{bmatrix} y_1 & y_2 & \cdots & y_{n-L+1} \\ y_2 & y_3 & \cdots & y_{n-L+2} \\ \vdots & \vdots & \vdots & \vdots \\ y_L & y_{L+1} & \cdots & y_n \end{bmatrix}. \tag{1}$$

This matrix has the same form as the data matrix produced by L variables, each being a time series of length $n - L + 1$. $\mathbf{Y}$ can also be viewed as composed of its column vectors $\mathbf{y}^{(l)}$, i.e.

$$\mathbf{Y} \equiv \left[\mathbf{y}^{(1)} | \mathbf{y}^{(2)} | \cdots | \mathbf{y}^{(n-L+1)} \right], \tag{2}$$

where the delay coordinate vector

$$\mathbf{y}^{(l)} = \begin{bmatrix} y_l \\ y_{l+1} \\ \vdots \\ y_{l+L-1} \end{bmatrix}. \tag{3}$$

The vector space spanned by $\mathbf{y}^{(l)}$ is called the *delay coordinate space*. The number of lags L is usually taken to be at most $1/4$ of the total record length.

Standard PCA can be performed on $\mathbf{Y}$, resulting in

$$\mathbf{y}^{(l)} = \mathbf{y}(t_l) = \sum_j a_j(t_l)\,\mathbf{e}_j\,, \tag{4}$$

where a_j is the j^{th} principal component (PC), a time series of length $n - L + 1$, and $\mathbf{e}_j$ is the j^{th} eigenvector (or loading vector) of length L. Together, a_j and $\mathbf{e}_j$, represent the j^{th} SSA mode.

This method is called singular spectrum analysis, as it studies the ordered set (spectrum) of singular values (the square roots of the eigenvalues).

SSA has become popular in the field of dynamical systems (including chaos theory), where delay coordinates are commonly used.

By lagging a time series, one is providing info on the first-order differencing of the discrete time series, with the first-order difference $\approx$ the derivative.

Repeated lags means higher-order differences (derivatives) are provided.

The first SSA reconstructed component (RC) is the approximation of the original time series $y(t)$ by the first SSA mode.

As the eigenvector $\mathbf{e}_1$ contains the loading over a range of lags, the first SSA mode, i.e. $a_1(t_l)\,\mathbf{e}_1$, provides an estimate for the y values over a range of lags starting from the time t_l. E.g., at time t_L, estimates of $y(t_L)$ can be obtained from any one of the delay coordinate vectors $\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(L)}$.

Each value in the reconstructed RC time series $\tilde{y}$ at time t_i involves averaging over the contributions at t_i from the L delay coordinate vectors which provide estimates of y at time t_i.

Comparison with Fourier spectral analysis:
Unlike Fourier, SSA does not in general assume the time series to be periodic; hence, there is no need to taper the ends of the time series as commonly done in Fourier spectral analysis.

As the wave forms extracted from the SSA eigenvectors are not restricted to sinusoidal shapes, the SSA can in principal capture an anharmonic wave more efficiently than the Fourier method. However in many cases, the SSA eigenvectors may turn out to be not be very different from sinusoidal-shaped functions.

E.g., analyze the sawtooth wave

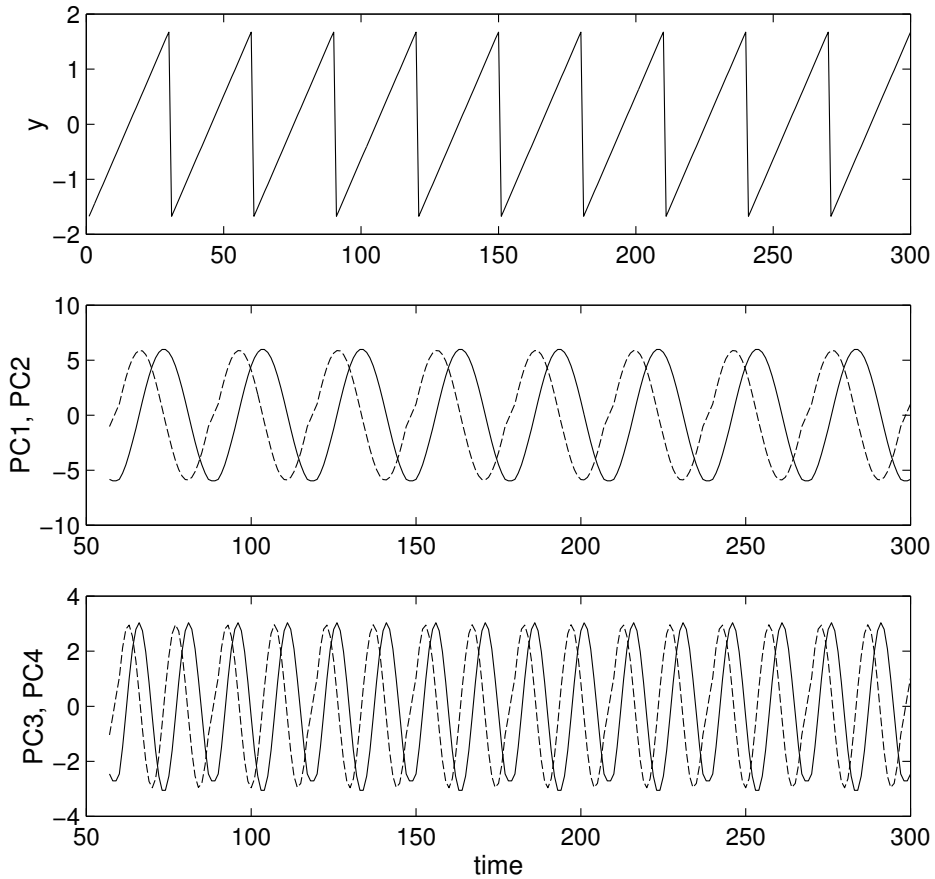


Figure : Top panel shows the sawtooth wave signal used for the SSA analysis. Middle panel shows the SSA PC1 (solid) and PC2 (dashed), while bottom panels shows SSA PC3 (solid) and PC4 (dashed).

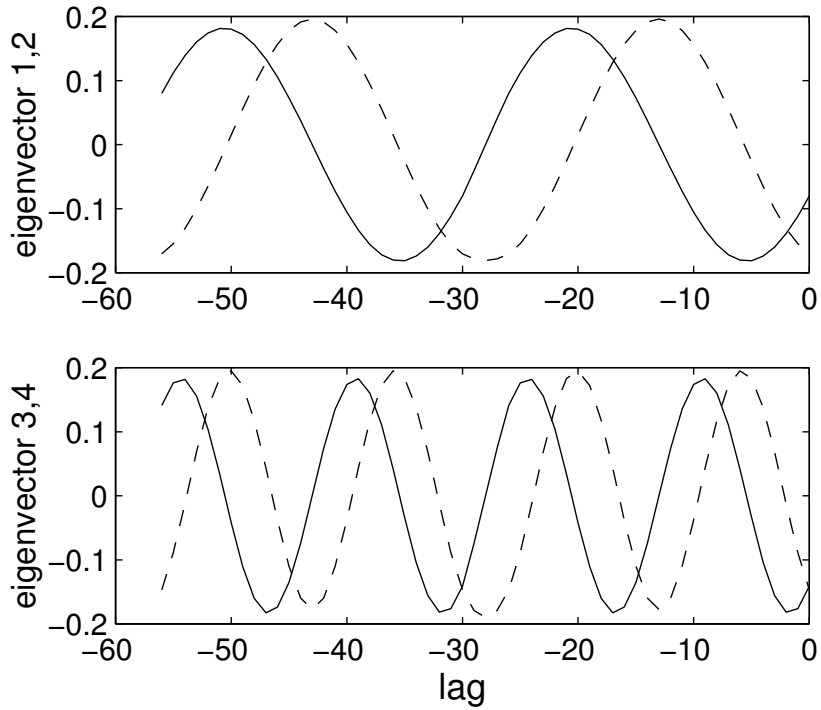


Figure : Top panel shows the SSA eigenvector 1 (solid) and eigenvector 2 (dashed), while bottom panel shows eigenvector 3 (solid) and 4 (dashed) for the sawtooth wave signal.

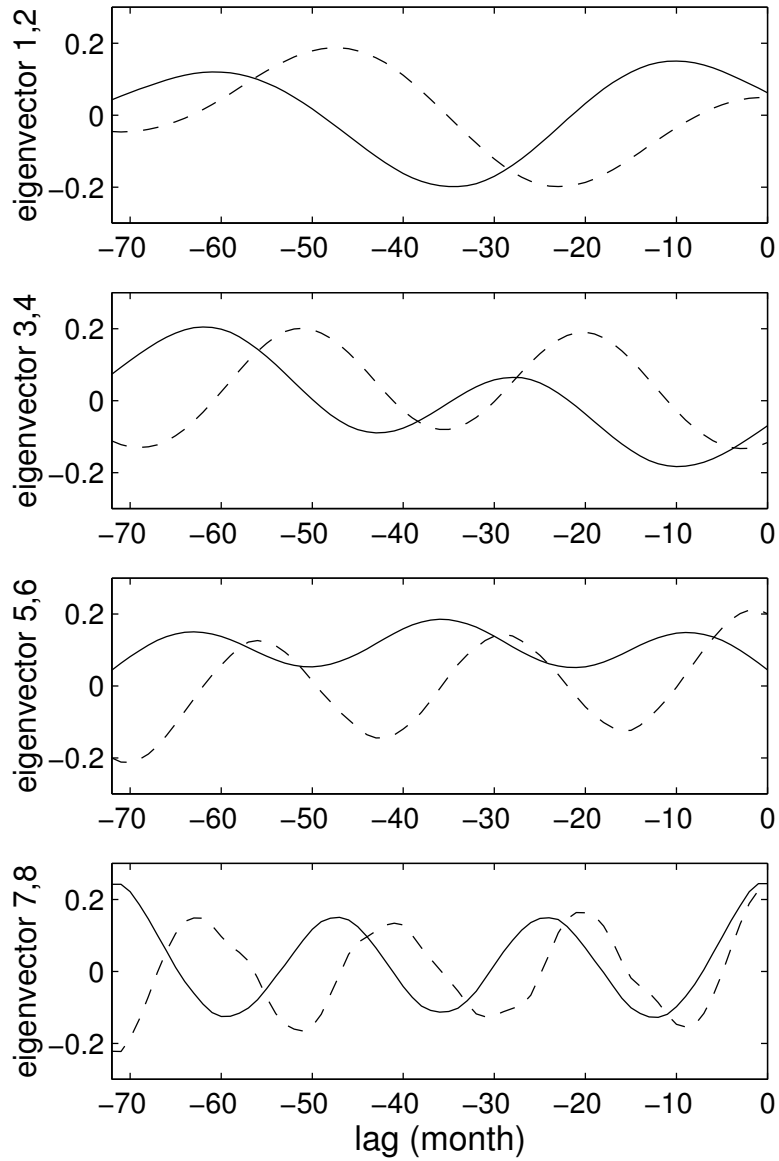
The first pair of SSA modes captured 61.3% of the variance, while the second pair captured 15.4%.
With Fourier spectral analysis: the fundamental frequency band captured 61.1%, while the first harmonic band captured 15.4% of the variance.

In the real world, except for the seasonal cycle and tidal cycles, signals tend not to have a precise frequency like the sawtooth wave. E.g. El Niño-Southern Oscillation (ENSO) .

The *Southern Oscillation Index* (SOI) is defined as the normalized air pressure difference between Tahiti and Darwin.

The SOI is known to have the main spectral peak at a period of about 4-5 years. For SSA, the window L needs to be long enough to accommodate this main spectral period; choose $L = 72$ months.

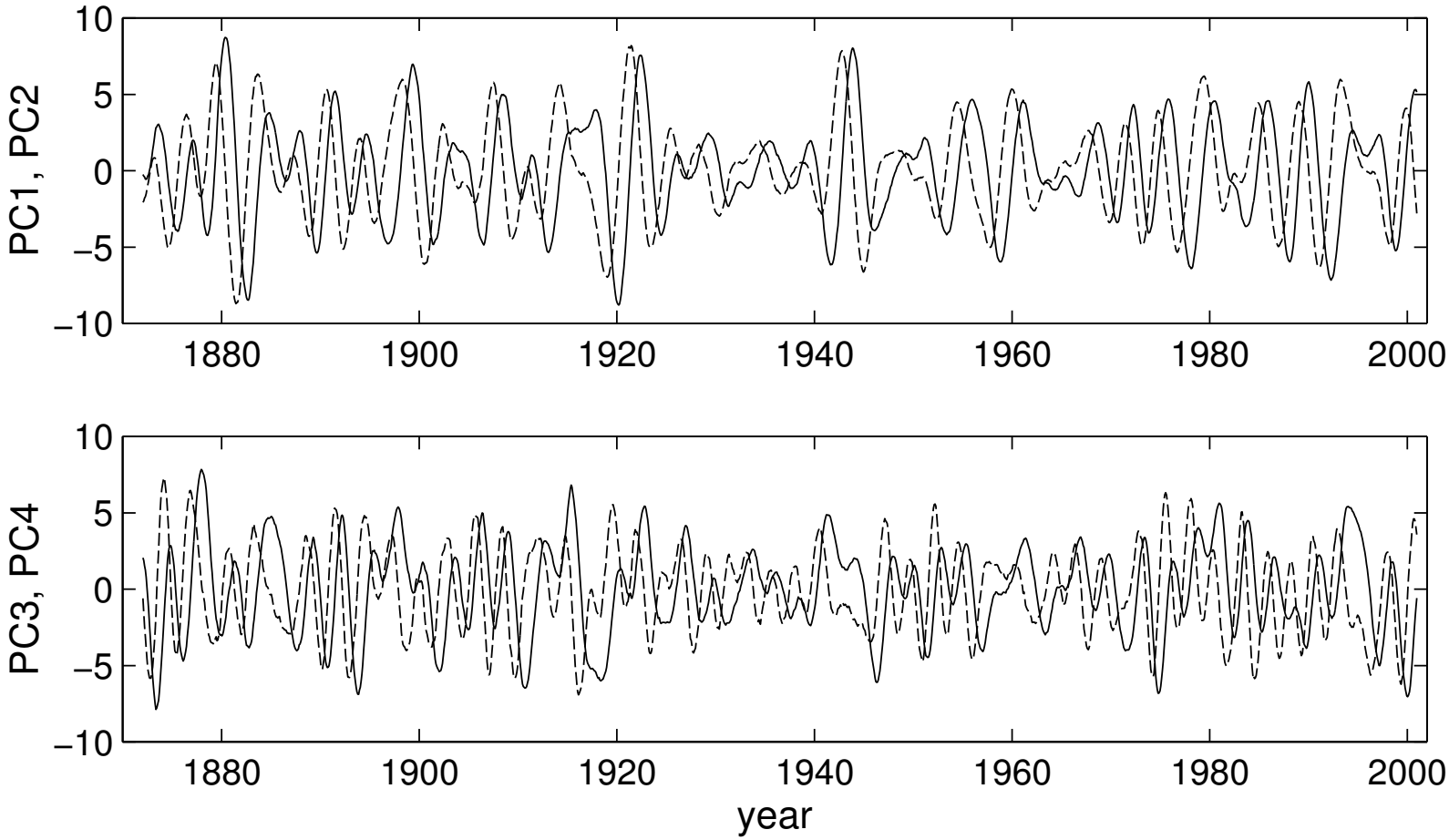
The first 8 SSA eigenvectors of the monthly SOI time series (1866 – 2000) (Hsieh and Wu, 2002) (with modes 1,3,5,7 solid and modes 2,4,6,8 dashed):



These first eight modes account for 11.0, 10.5, 9.0, 8.1, 6.4, 5.6, 3.3 and 2.0 %, respectively, of the variance.

Since the SO phenomenon does not have a precise frequency, Fourier analysis led to energy being spread between many frequency bands, with the strongest band accounting for only 4.0% of the variance (vs. 11.0% of the first SSA mode).

The four leading PCs:



Another advantage of SSA over the Fourier approach lies in the multivariate situation— the Fourier approach does not generalize naturally to large multivariate datasets, whereas the SSA, based on the PCA method, does.

4.5 Multichannel singular spectrum analysis [Book, Sect. 3.5]

There are m variables $y_k(t_j) \equiv y_{kj}, (k = 1, \ldots, m;\ j = 1, \ldots, n)$. The data matrix time lagged by l $(l = 0, 1, 2, \ldots, L-1)$ is

$$\mathbf{Y}_{(l)} = \begin{bmatrix} y_{1,1+l} & \cdots & y_{1,n-L+l+1} \\ \vdots & \ddots & \vdots \\ y_{m,1+l} & \cdots & y_{m,n-L+l+1} \end{bmatrix} . \tag{5}$$

The augmented data matrix is

$$\mathbf{Y} = \begin{bmatrix} \mathbf{Y}_{(0)} \\ \mathbf{Y}_{(1)} \\ \mathbf{Y}_{(2)} \\ \vdots \\ \mathbf{Y}_{(L-1)} \end{bmatrix}, \tag{6}$$

i.e.

$$\mathbf{Y} = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1,n-L+1} \\ \vdots & \vdots & \vdots & \vdots \\ y_{m1} & y_{m2} & \cdots & y_{m,n-L+1} \\ \vdots & \vdots & \vdots & \vdots \\ y_{1L} & y_{1,L+1} & \cdots & y_{1n} \\ \vdots & \vdots & \vdots & \vdots \\ y_{mL} & y_{m,L+1} & \cdots & y_{mn} \end{bmatrix}. \tag{7}$$

PCA can again be applied to the augmented data matrix $\mathbf{Y}$ to get the SSA modes.

Q3: (a) What is the dimension of the augmented data matrix $\mathbf{Y}$ with lag L, where the original data matrix has m variables and n time observations?
(b) If your spatial domain has 1°lat.$\times 1^\circ$lon. gridded data covering 50° of latitude and 130° of longitude, and you have monthly observations for 60 years, what is the dimension of your matrix $\mathbf{Y}$ with lag $L = 72$ months?

If more than one variable, the method is called the *space-time PCA* (ST-PCA) , or *multichannel singular spectrum analysis* (MSSA). We will, for brevity, use the term SSA to denote both the univariate and the multivariate cases.

The term *extended empirical orthogonal function* (EEOF) analysis is also used in the literature, especially when the number of lags (L) is small.

So far, we have assumed the time series was lagged one time step at a time. To save computational time, larger lag intervals can be used, i.e. lags can be taken over several time steps at a time.

E.g. SSA analysis of the tropical Pacific monthly sea surface temperature anomalies (SSTA) data from 1950-2000, where the climatological seasonal cycle and the linear trend have been removed from the SST data to give the SSTA. A window of 73 months was chosen (Hsieh and Wu, 2002).

With a lag interval of 3 months, the original plus 24 lagged copies of the SSTA data formed the augmented SSTA dataset. (Note that if a lag interval of 1 month were used instead, then to cover the window of 73 months, the original plus 72 copies of the SSTA data would have produced a much bigger augmented data matrix).

The first six SSA modes respectively explain 12.4%, 11.7%, 7.1%, 6.7%, 5.4%, 4.4% of the total variance of the augmented dataset.

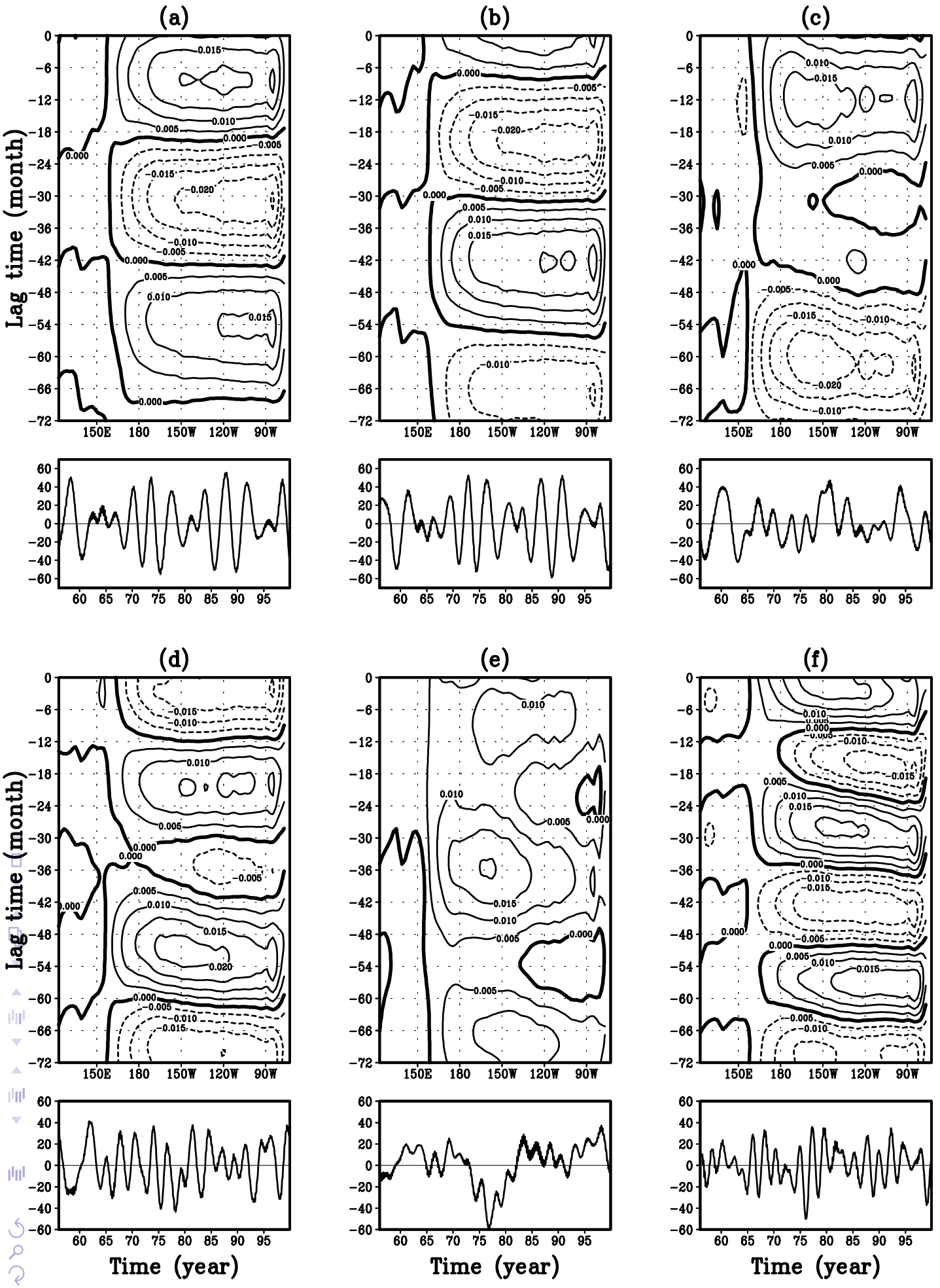

(a)
(b)
(c)
(d)
(e)
(f)
Lag time (month)
Time (year)
150E
180
150W
120W
90W
60
65
70
75
80
85
90
95

Figure: The SSA modes 1-6 for the tropical Pacific SSTA shown in (a)-(f), respectively. The contour plots display the space-time eigenvectors (loading patterns), showing the SSTA along the equator as a function of the lag. Solid contours indicate positive anomalies and dashed contours, negative anomalies, with the zero contour indicated by the thick solid curve. In a separate panel beneath each contour plot, the principal component (PC) of each SSA mode is also plotted as a time series.

The first two modes have space-time eigenvectors (i.e. loading patterns) showing an oscillatory time scale of about 48 months, comparable to the ENSO time scale, with the mode 1 anomaly pattern occurring about 12 months before a very similar mode 2 pattern, i.e. the two patterns are in quadrature. The PC time series also show similar time scales for modes 1 and 2. Modes 3 and 5

show longer time scale fluctuations, while modes 4 and 6 show shorter time scale fluctuations— around the 30-month time scale.

References:

Ghil, M., Allen, M. R., Dettinger, M. D., Ide, K., Kondrashov, D., Mann, M. E., Robertson, A. W., Saunders, A., Tian, Y., Varadi, F., and Yiou, P. (2002). Advanced spectral methods for climatic time series. *Reviews of Geophysics*, 40. 1003, DOI: 10.1029/2000RG000092.

Hsieh, W. W. and Wu, A. (2002). Nonlinear multichannel singular spectrum analysis of the tropical Pacific climate variability using a neural network approach. *Journal of Geophysical Research*, 107(C7). DOI: 10.1029/2001JC000957.