A wavelet-based mode decomposition applied to the ENSO index

Adrien DELIÈGE, Samuel NICOLAY

Department of mathematics, University of Liège
Allée de la découverte 12, 4000 Liège, Belgium
adrien.deliege@ulg.ac.be, s.nicolay@ulg.ac.be

Résumé – Ce papier présente la description et l'application d'une décomposition modale basée sur la transformée en ondelettes. Une fois la décomposition réalisée, les composantes sont séparément extrapolées, donnant lieu à une prévision sur l'évolution du signal. La qualité de la prévision est évaluée via une procédure de prévisions rétroactives. Pour l'indice climatique ENSO, la décomposition fournit quatre composantes correspondant à des pseudo-périodes d'environ 21, 31, 43 et 61 mois respectivement. La reconstruction détecte 97% des événements El Niño/La Niña des 65 dernières années.

Abstract – This paper consists of a description and an application of a mode decomposition performed through a wavelet transform. Once the decomposition is performed, the components are separately smoothly extrapolated, which leads to a forecast of the signal. The quality of the forecast is assessed through a hindcast (running retroactive probing forecasts) procedure. For the ENSO time series, the decomposition leads to four components corresponding to pseudo-periods of about 21, 31, 43 and 61 months respectively and the reconstruction recovers 97% of the major El Niño/La Niña events of the past 65 years.

1 Introduction

The aim of this work is to provide and apply a wavelet-based mode decomposition (introduced in [13]) and use it to develop a simple yet powerful forecasting method. The basic idea is to extract quasi-periodic oscillating components from the continuous wavelet transform of the original signal but, unlike the Fourier transform, the amplitudes are not constant anymore : they vary smoothly with the abscissa, as in the empirical mode decomposition ([4, 8]). This allows to drastically decrease the number of terms needed to rebuild accurately the signal and to take into account only the terms carrying most of the information. By doing so, the reconstructed signal resolves the large variations of the original one without taking the noise into account. Then, the extracted components are smoothly extrapolated using Lagrange polynomials, such a process leads to an extrapolation of the reconstructed signal which stands for a forecast of the raw signal. The forecasting procedure is finally assessed using hindcasts (running retroactive probing predictions) from which error bars for the predictions are derived.

We use the abovementioned procedure to study real-life data. For that purpose, the El Niño Southern Oscillation (ENSO) index is analyzed. ENSO is a climate pattern induced by sea surface temperature anomalies (SSTA) in the tropical Pacific Ocean. An anomalous warming in the SSTA is known as El Niño, while an anomalous cooling bears the name of La Niña. ENSO is well recognized as the dominant mode of interannual variability in the tropical Pacific Ocean. It affects the atmospheric general circulation which transmits the ENSO signal to the other parts of world ; these remote effects are called "teleconnections" and induce changes in the occurrence of severe weather events, which dramatically affect human activities and ecosystems worldwide [1, 5, 7, 17]. Therefore, short-term ENSO predictions are of first importance for seasonal climate forecasts, in order to help governments and industries to plan actions before the occurrence of these phenomena.

As a result of the analysis, it turns out that the signal can be decomposed into four components of periods of about 21, 31, 43 and 61 months with smooth time-varying amplitudes. The reconstructed signal recovers $30/31$ of the El Niño/La Niña events that occurred from 1950. These components are extrapolated to perform several years forecasts of the ENSO index, in the same spirit as in [16]. The prediction skills of this method are tested through a hindcast procedure. It appears that three years hindcasts allow to recover up to 78% of the El Niño/La Niña events of the last 50 years. Compared to other methodologies, our forecasting method is particularly competitive for predictions exceeding 6 months, since the variability of ENSO is mainly driven by the four mentioned periods. The next La Niña event should start early in 2018 and should be followed by a strong El Niño event in the second semester of 2019.

2 Method

2.1 Wavelet-based mode decomposition

The wavelet-based mode decomposition provides a simple yet powerful method for decomposing a signal into a finite number of components [13]. Each component is associated to a mean frequency but, unlike the Fourier transform, this decomposition does not lead to pure cosines. A component is

not identified with a fixed amplitude or frequency ; they slowly evolve through time. A Fourier series decomposes a periodic signal into an infinite sum of sine and cosine functions ; from a practical point of view, a signal f is approximated by a finite sum of K cosines, $f(t) \approx \sum_{k=1}^{K} c_k \cos(\omega_k t + \phi_k)$, where the coefficients c_k are amplitudes and ω_k frequencies. However, such a decomposition often leads to a sum with too many terms. The idea underlying the wavelet-based mode decomposition is to decrease the number of terms by considering the amplitudes c_k as functions of t. This property relies on the fact that wavelets provide a good local time-frequency resolution. One thus tries to have the following decomposition : $f(t) \approx \sum_{j=1}^{J} C_j(t) \cos(\omega_j t + \phi_j)$, where J should be much smaller than K (see equality (3) for the exact formula). For a detailed description of the theory of continuous wavelet transforms, the reader is referred to [3, 10].

The wavelet used in this study is the one used in [13], which is similar to the Morlet wavelet and is well-suited for frequency analysis. With such a wavelet, the complex wavelet transform of a real signal f (belonging to the space $L^2(\mathbb{R})$) is defined as

$$Wf(a,t) = \int f(x)\bar{\psi}(\frac{x-t}{a})\,\frac{dx}{a}, \qquad (1)$$

where $\bar{\psi}$ is the complex conjugate of ψ, $t \in \mathbb{R}$ stands for the location parameter and $a > 0$ denotes the scale parameter. Let us note that t is also referred as the time parameter in the following.

The decomposition used in this analysis can be summed up as follows. First, a wavelet spectrum Λ is associated to the signal f [14] :

$$\Lambda(a) = E|Wf(a,\cdot)|,$$

where E denotes the mean with respect to the time parameter. The scales $a_1, \ldots, a_J$ for which Λ reaches a maximum are kept (let us recall that a scale (or a period) behaves as the inverse of a frequency, see e.g. [3, 10]). The signal f is then decomposed in the following way,

$$f(t) = \hat{f}_0(t) + \sum_{j=1}^{J} \hat{f}_j(t), \qquad (2)$$

with

$$\hat{f}_j(t) = |Wf(a_j,t)| \cos\left(\arg Wf(a_j,t)\right) \qquad (3)$$

if $j \geq 1$ and $\hat{f}_0(t) = f(t) - \sum_{j=1}^{J} \hat{f}_j(t)$.

2.2 Forecasting procedure

Once the decomposition of f has been performed using equation (2), the signal $\sum_{j=1}^{J} \hat{f}_j$ can be seen as an approximation of f in terms of oscillatory components. The reconstructed signal $\hat{f}$ used below has the following form (see [13]),

$$f(t) \approx \hat{f}(t) = \sum_{j=1}^{J} C_j(t) \cos(\omega_j t + \phi_j), \qquad (4)$$

where ω_j is the mean frequency of $\cos\left(\arg Wf(a_j,t)\right)$ (and is proportional to $1/a_j$). This reconstruction $\hat{f}$ depends on J amplitudes, J frequencies and J phases. Let us remark that the amplitude C_j, corresponding to a given frequency, is a function of t ; this allows to drastically decrease the number of components compared to the Fourier transform. Note that the terms "period" and "frequency" are used interchangeably in the following. Also, the periods detected are also called "pseudo-periods" due to the way the ω_j's are computed.

Each component can be extrapolated in the following way. Under the hypothesis stated above, if T is the abscissa of the last available data, the only unknown parameter is the amplitude $C_j(t)$ ($1 \leq j \leq J$), which is supposed to vary smoothly on $[0,T]$; its time-varying values can therefore be extrapolated (for $t > T$) using Lagrange polynomials. The computation was based on the last 4 available values of the signal (i.e. corresponding to $t = T-1,\ T-2,\ T-3$ and $T-4$). This number 4 was chosen in order to minimize the root mean square error between the real signal and the hindcasts (see below) ; however, altering this parameter (e.g. 2 or 3) does not significantly modify the results and excessively increasing it leads to unstable extrapolations. Summing the extrapolated values related to each component defines the forecast of f ; more precisely, it corresponds to a forecast of the reconstructed signal $\hat{f}$.

Let us remark that in the case of the ENSO time series, the mean frequencies are well-known, as discussed in section 3 (see [12, 14, 9]).

2.3 Hindcasts and estimation of errors

Such a forecast needs to be properly cross-validated. An intuitive idea is to look at the first t_0 ($t_0 < T$) data elements, then produce a retroactive forecast and compare the so-obtained prediction to the original signal. One can then perform the same method again, but with t_0 increased by 1, and so on until the end of the time series. It is important to notice that, in this way, no information past the date from which a prediction is issued is used.

More precisely, the efficiency of the forecasts was tested with probing hindcasts, computed as follows.

1) One considers f_{t_0} (the time series f up to the time point t_0), computes its wavelet spectrum and looks for the scales for which it reaches a maximum. The reconstructed signal $\hat{f}_{t_0}$ is computed using equation (4).
2) Starting from the last four values of $\hat{f}_{t_0}$, the retroactive forecast time series $F_{t_0}(t_0+1), \ldots, F_{t_0}(t_0+k)$, ranging from the abscissa t_0+1 to t_0+k are obtained using the method described in the previous section. In particular, the a priori unknown local future values (abscissa greater than t_0) are not used to obtain the values of F_{t_0}.
3) The initial condition is increased by 1 and steps $1-3$ are repeated.
4) The k-months lead hindcast associated to the initial condition t_0, denoted by $F_{t_0,k}$, is the signal $F_{t_0}(t_0+k), F_{t_0+1}(t_0+1+k), \ldots, F_{T-k}(T)$ where T is the abscissa of the last available data in the original signal.

In order to measure the efficiency of our method, we procee-

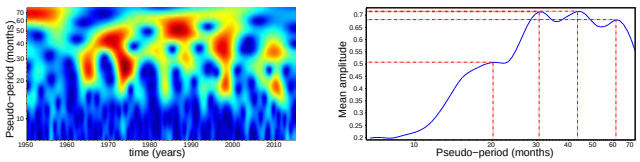

FIGURE 1 – Left : modulus of the wavelet transform (values range from 0 (dark blue) to 1.2 (dark red)) and wavelet spectrum. Four peaks are detected.

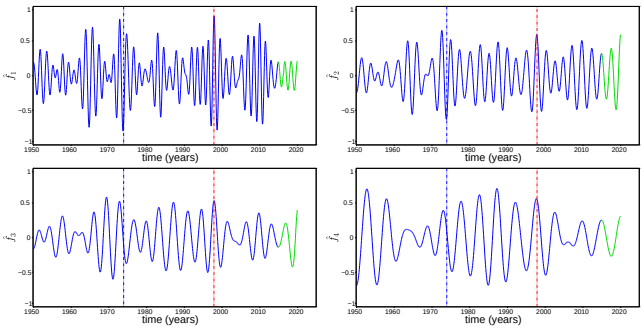

FIGURE 2 – The four components extracted from the Niño 3.4 time series, corresponding to the periods of 20.9, 30.6, 43.28 and 61.21 months respectively. The red (resp. blue) line indicates when the strongest El Niño (resp. La Niña) event reached its peak. The green parts are the forecasts of the components for the next five years.

ded in the following way. Since the difference between $f(t_0 + k)$ and $F_{t_0}(t_0 + k)$ is the error of the k-months prediction at time t_0+k issued at time t_0, the hindcasts were used in a cross-validation procedure. The root mean square error (RMSE) between f and the hindcast $F_{t_0,k}$, was used as a measure of the accuracy of the forecast at k data as lead time and led to error bars. Finally, the Pearson correlation coefficient between these signals was also computed.

3 Application to the ENSO index

The procedure described above is now applied to a real-life signal. The example presented here is the decomposition and forecast of the Niño 3.4 index (ERSST.V3B SST Niño 3.4 time series provided by the Climate Prediction Center), which consists of monthly-sampled SSTA (in Celsius degrees) in the Eastern Pacific Ocean recorded from Jan 1950 to Dec 2014. Over the last two decades, many models have been proposed for forecasting the ENSO phenomenon (see *http ://iri.columbia.edu-/climate/ENSO/currentinfo/QuickLook.html*) by focusing on the Niño 3.4 index [11, 18, 15, 19], generally with lead times of up to one year.

The wavelet transform of the signal and its wavelet spectrum are plotted in Fig. 1. The spectrum displays four peaks corresponding to pseudo-periods of 20.9, 30.6, 43.28 and 61.21 months respectively (note that a weak peak is also detected around 7 months but is considered as noise). Then, the four corresponding components are computed and rescaled to minimize the RMSE between the reconstructed signal and the original series (see Fig. 2). It can clearly be seen that the strongest El Niño and La Niña events occurred when all the components simultaneously reached a peak.

The reconstruction of the Niño 3.4 signal, denoted by m-Niño 3.4, is computed as the sum of the four components mentioned above and is plotted in Fig. 3. The Pearson correlation between these two signals is 0.894 and the RMSE is $0.366°$C, which confirms that m-Niño 3.4 allows to recover most of the variability of the original signal. Let us note that if the component associated to the 7 months period is added, the results remain almost unchanged (the RMSE is $0.336°$C and the correlation is 0.912). Using the definition of an El Niño/La Niña event of the Climate Prediction Center (threshold of $+0.5°$C (El Niño) or $-0.5°$C (La Niña) reached by at least five consecutive months), $30/31$ major events are recovered by m-Niño

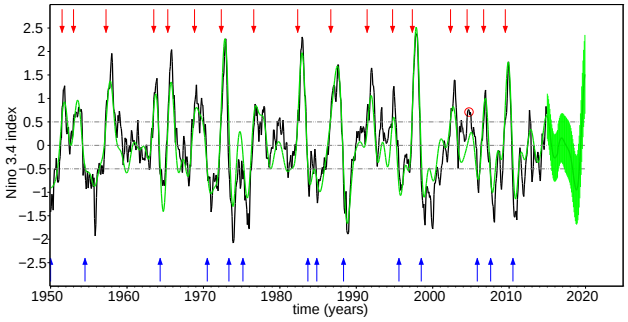

FIGURE 3 – Niño 3.4 (black) and m-Niño 3.4 (green) and the forecast of m-Niño 3.4 for the next five years. The error bars were estimated using the RMSE related to the hindcasts. The beginning of the El Niño and La Niña episodes in the Niño 3.4 index are indicated with red and blue arrows respectively. The undetected El Niño event is pointed by a red circle.

3.4 (see Fig. 3).

As shown in Fig. 3, a forecast of the Niño 3.4 time series is issued, where the errors bars are computed through the RMSE of the hindcasts of the corresponding lead time. Let us note that 92% of the extreme events that occurred in the last 50 years can be detected 1 year in advance, and still 87% (resp. 78%) can be recovered 2 (resp. 3) years in advance.

In comparison with other Niño 3.4 forecasting methodologies [2, 6, 18, 19], our approach becomes competitive for retroactive predictions exceeding 6 months. Indeed, for most of the other methodologies, the correlation between predicted and observed Niño 3.4 drops below 0.7 and the RMSE rises above the standard deviation (i.e. the monthly variability $\sim 0.82°$C) for retroactive predictions exceeding 12 months (see e.g. [2, 19]). That is why most of them limit their retroactive predictions to 12 months. In our case, the correlation (from 0.80 to 0.68) and the RSME (from 0.50 to 0.62) barely change for

lead times between 12 and 36 months. However, for retroactive predictions shorter than 6 months, our methodology is worse because the predictions are obtained by extrapolating reconstructed signals and not the Niño 3.4 signal itself. Our approach cannot resolve the short term (1–6 months) variability of Niño 3.4 as the other forecasting methodologies, but is able to successfully predict the large variations in Niño 3.4 several years in advance. Taking more modes into account (such as the 7 months period component) would slightly increase the accuracy of the reconstruction and short-term forecasts but would barely change mid-term predictions.

4 Conclusions

We apply the wavelet-based methodology described in [13] to decompose the Niño 3.4 index into quasi-periodic oscillating components (periods of about 21, 31, 43 and 61 months) with smooth time-varying amplitudes. The components obtained allow to rebuild accurately the original signal (strong correlation, low RMSE). More than 95% of the observed major El Niño/La Niña events are detected in the reconstructed signal. This shows that the main part of the variability found in ENSO is explained by a superposition of these four quasi-periodic signals. Extrapolating the amplitude of each of the four components (using Lagrange polynomials) allows us to extrapolate the rebuilt signal and therefore to predict the next occurrence of a major event. Moreover the hindcast procedure also leads to error bars for these predictions. The next La Niña event should start early in 2018 and should be followed soon after by a strong El Niño event in the second semester of 2019.

This forecasting method is particularly well-suited for predictions exceeding 6 months and allows to predict the large variations in Niño 3.4 (which induce major El Niño/La Niña events) up to four years in advance. As future developments, understanding the underlying mechanisms governing ENSO variability (i.e. the origin of the "pseudo-periodicities" detected in Niño 3.4) is a subject of first importance in order to improve the reliability of the forecasting models. Finally, this methodology could be applied to other time series (e.g. the North Atlantic Oscillation index) where periods of about 30, 40 and 60 months have also been found [14].

Références

[1] K. Ashok, S.K. Behera, S.A. Rao, H. Weng, and T. Yamagata (2007) El Niño Modoki and its possible teleconnection. *J. Geophys. Res*, 112(10.1029).

[2] D. Chen, M.A. Cane, A. Kaplan, S.E. Zebiak, and D. Huang (2004) Predictability of El Niño over the past 148 years. *Nature*, 428(6984) :733–736.

[3] I. Daubechies (1982) *Ten Lectures on Wavelets*. SIAM.

[4] P. Flandrin, G. Rilling, and P. Goncalvès (2003) Empirical Mode Decomposition as a filter bank. *IEEE Sig. Proc. Lett.*

[5] M.H. Glantz (2001) Currents of Change : Impacts of El Niño and La Niña on climate society. Cambridge University Press.

[6] Y.G. Ham, J.S. Kug, and I.S. Kang (2009) Optimal initial perturbations for El Nino ensemble prediction with ensemble Kalman filter. *Climate dynamics*, 33(7) :959–973.

[7] S.M. Hsiang, K.C. Meng, and M.A. Cane (2011) Civil conflicts are associated with the global climate. *Nature*, 476(7361) :438–441.

[8] N.E. Huang, Z. Shen, S.R. Long, M.C. Wu, H.H. Shih, Q. Zheng, N.-C. Yen, C.C. Tung, and H.H. Liu (1998) The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. *Proc. R. Soc. London A*, 454 :903–995.

[9] G. Mabille and S. Nicolay (2009) Multi-year cycles observed in air temperature data and proxy series. *The European Physical Journal-Special Topics*, 174(1) :135–145.

[10] S. Mallat (1999) *A Wavelet Tour of Signal Processing*. Academic Press.

[11] S.J. Mason and G.M. Mimmack (2002) Comparison of some statistical methods of probabilistic forecasting of ENSO. *Journal of Climate*, 15(1) :8–29.

[12] V. Moron, R. Vautard, and M. Ghil (1998) Trends, interdecadal and interannual oscillations in global sea-surface temperatures. *Climate Dynamics*, 14(7) :545–569.

[13] S. Nicolay (2011) A wavelet-based mode decomposition. *Eur. Phys. J. B*, 80 :223–232.

[14] S. Nicolay, G. Mabille, X. Fettweis, and M. Erpicum (2009) 30 and 43 months period cycles found in air temperature time series using the Morlet wavelet method. *Climate dynamics*, 33(7) :1117–1129.

[15] M.K. Tippett and A.G. Barnston (2008) Skill of multimodel ENSO probability forecasts. *Monthly Weather Review*, 136(10) :3933–3946.

[16] P.J. Webster and C. Hoyos (2004) Prediction of monsoon rainfall and river discharge on 15-30-day time scales. *Bulletin of the American Meteorological Society*, 85(11) :1745–1765.

[17] S.W. Yeh, J.S. Kug, B. Dewitte, M.H. Kwon, B.P. Kirtman, and F.F. Jin (2009) El Niño in a changing climate. *Nature*, 461(7263) :511–514.

[18] F. Zheng, J. Zhu, R.H. Zhang, and G. Zhou (2006) Improved ENSO forecasts by assimilating sea surface temperature observations into an intermediate coupled model. *Advances in Atmospheric Sciences*, 23(4) :615–624.

[19] J. Zhu, G.Q. Zhou, R.H. Zhang, and Z. Sun (2012) Improving ENSO prediction in a hybrid coupled model with an embedded entrainment temperature parameterisation. *International Journal of Climatology*, online, 2012.