Review

Time-dependent spectral analysis of epidemiological time-series with wavelets

Bernard Cazelles[1,2,*], Mario Chavez[3], Guillaume Constantin de Magny[4], Jean-Francois Guégan[5] and Simon Hales[6]

[1]*CNRS UMR 7625, Ecole Normale Supérieure, 46 rue d'Ulm, 75230 Paris, France*
[2]*IRD UR GEODES, 93143 Bondy, France*
[3]*LENA-CNRS UPR 640, Hôpital Pitié-Salpêtrière, 4 boulevard de l'Hôpital, 75651 Paris cedex 13, France*
[4]*Institute for Advanced Computer Studies, University of Maryland, College Park, MD 20742, USA*
[5]*IRD-CNRS UMR 2724, 911 avenue Agropolis, BP 64501, 34394 Montpellier cedex 05, France*
[6]*Wellington School of Medicine and Health Sciences, University of Otago, Mein Street, Newton, Wellington South, New Zealand*

In the current context of global infectious disease risks, a better understanding of the dynamics of major epidemics is urgently needed. Time-series analysis has appeared as an interesting approach to explore the dynamics of numerous diseases. Classical time-series methods can only be used for stationary time-series (in which the statistical properties do not vary with time). However, epidemiological time-series are typically noisy, complex and strongly non-stationary. Given this specific nature, *wavelet analysis* appears particularly attractive because it is well suited to the analysis of non-stationary signals. Here, we review the basic properties of the wavelet approach as an appropriate and elegant method for time-series analysis in epidemiological studies. The wavelet decomposition offers several advantages that are discussed in this paper based on epidemiological examples. In particular, the wavelet approach permits analysis of transient relationships between two signals and is especially suitable for gradual change in force by exogenous variables.

Keywords: wavelets analysis; time series; epidemiology; non-stationarity

1. INTRODUCTION

Globalization has altered human infectious disease risks due to land use changes, intensification of agriculture, loss of biodiversity, climate change, population increase, urbanization, and increased travel. Emerging and re-emerging infectious diseases are nowadays an increasing public health challenge. Disease emergences reveal the complex dynamical relationships between humans, pathogens and the global environment (Morens *et al.* 2004; Woolhouse *et al.* 2005). Recently, new concerns about global warming have yielded numerous studies about the role of climate variability and climate change in interannual disease patterns (McCarthy *et al.* 2001; Cazelles *et al.* 2005; Patz *et al.* 2005; McMichael *et al.* 2006).

Within this context, developing effective public health policy requires integrating data that are diverse and highly variable in quality. Mathematical models of disease transmission can provide a framework for improving our understanding of the complex dynamics of infectious disease epidemics (Anderson & May 1991). This is crucial in attempts to design effective intervention and control strategies. Since the early 1900s, sophisticated mathematical and statistical methods have been used (Ross 1911; Greenwood 1924). Nevertheless, the lack of appropriate datasets has impeded the validation of mechanistic mathematical models. More recently, time-series methods have appeared as an interesting alternative and have been used to explore the dynamics of numerous epidemics (e.g. Helfenstein 1986; Catalano & Serxner 1987).

Time-series are often used in short-term analyses of air pollution and human health (Dominici *et al.* 2002; Bell *et al.* 2004). To take into account the time dependences, trends and cycles in epidemiological time-series, spectral analysis has also been used. For instance, Bishop (1977) has presented a brief review
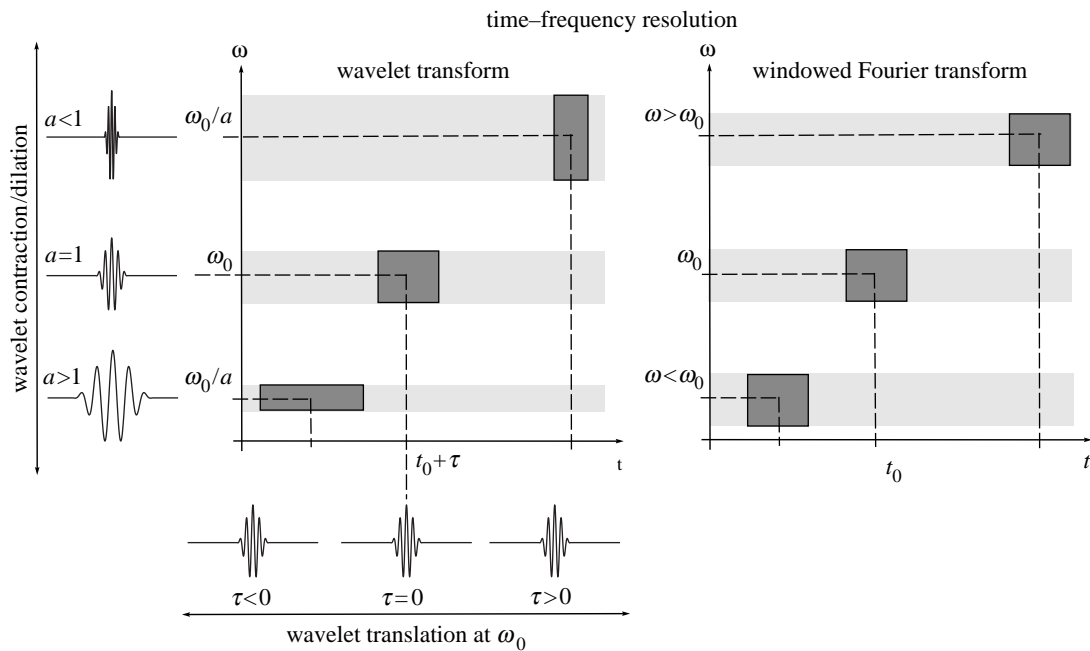
*Author for correspondence (cazelles@biologie.ens.fr).

time–frequency resolution

Figure 1. Time–frequency resolution of the wavelet approach. (*a*) Examples of wavelets and their time–frequency boxes representing the corresponding variance (energy) distribution. When the scale *a* decreases, the time resolution improves but the frequency resolution gets poorer and is shifted towards high frequencies. Conversely, if *a* increases the boxes shift towards the region of low frequencies and the height of the boxes becomes shorter (with a better frequency resolution) but their widths are longer (with a poor time resolution). (*b*) In contrast with wavelets, all the boxes of the windowed Fourier transform are obtained by a time- or frequency shift of a unique function, which yields the same variance spreads over the entire time–frequency plane.

illustrated by epidemiological examples. Bloomfield (1976) has investigated the use of Fourier analysis as a tool for determining whether aggravation of asthma symptoms is related to daily minimum temperature and/or to atmospheric pollutants. Spectral techniques have also been used to remove seasonal components before using regression models for analysing the relation between levels of particulate air pollution and daily mortality (Schwartz 1993; Dominici *et al.* 2002). Pascual and co-authors have used spectral techniques coupled with other approaches (singular spectrum analysis and non-parametric regression models) to study the influence of large-scale climatic oscillations on cholera epidemics (Pascual *et al.* 2000; Rodó *et al.* 2002).

These classical techniques can only be used for time-series in which the statistical properties do not vary with time, i.e. are stationary. However, epidemiological time-series are typically noisy, complex and strongly non-stationary. Long-term changes in climate, human demography and/or social features of human populations have generated non-stationarities in numerous epidemics datasets as underlined by the analyses of measles and whooping cough recordings (Duncan *et al.* 1996; Rohani *et al.* 1999). Recently, a robust relationship between El Niño oscillations and cholera prevalence in Bangladesh has been clearly demonstrated though the association was transient over time (Rodó *et al.* 2002).

To overcome the problems of analysing non-stationary time-series, it has recently been proposed to apply wavelet analysis to characterize them and also to estimate dependencies among non-stationary signals. Wavelet analysis performs a time-scale decomposition of the signal, which means the estimation of its spectral characteristics as a function of time (Lau & Weng 1995;

Torrence & Compo 1998). This approach reveals how the different scales (i.e. the periodic components) of the time-series change over time. Cross-wavelets and wavelet coherency generalize these methods, allowing the analyses of scale dependencies between the two signals.

Wavelet analysis appears particularly attractive given the specific nature of epidemiological and environmental time-series and the relationships between them. In 2001, Grenfell and collaborators published a paper that analysed synchrony patterns of measles in the UK (Grenfell *et al.* 2001). They used wavelets to show for the first time a progressive increase in epidemic phase with time, accompanying the increasing trend in vaccination rates. Since this work, several applications of wavelet analysis have been published. Broutin *et al.* (2005) performed a wavelet-based analysis of pertussis time-series in 12 countries to detect and quantify periodicity and synchrony between them. They showed a clear 3–4-year cycle in all countries, but the main finding was that this periodicity was transient and no global pattern in the effect of vaccination on pertussis dynamics emerged. Some of these papers were based on the characterization of time-series and the analysis of their possible association with environmental signals. Cazelles *et al.* (2005) used this approach to demonstrate a highly significant but discontinuous association between El Niño, precipitation and dengue epidemics in Thailand. This transient association has important consequences for the dynamics of this epidemic: when association with El Niño is strong, high synchrony of dengue epidemics over Thailand is observed. When this association is absent, the seasonal dynamics become dominant and the synchrony initiated in Bangkok
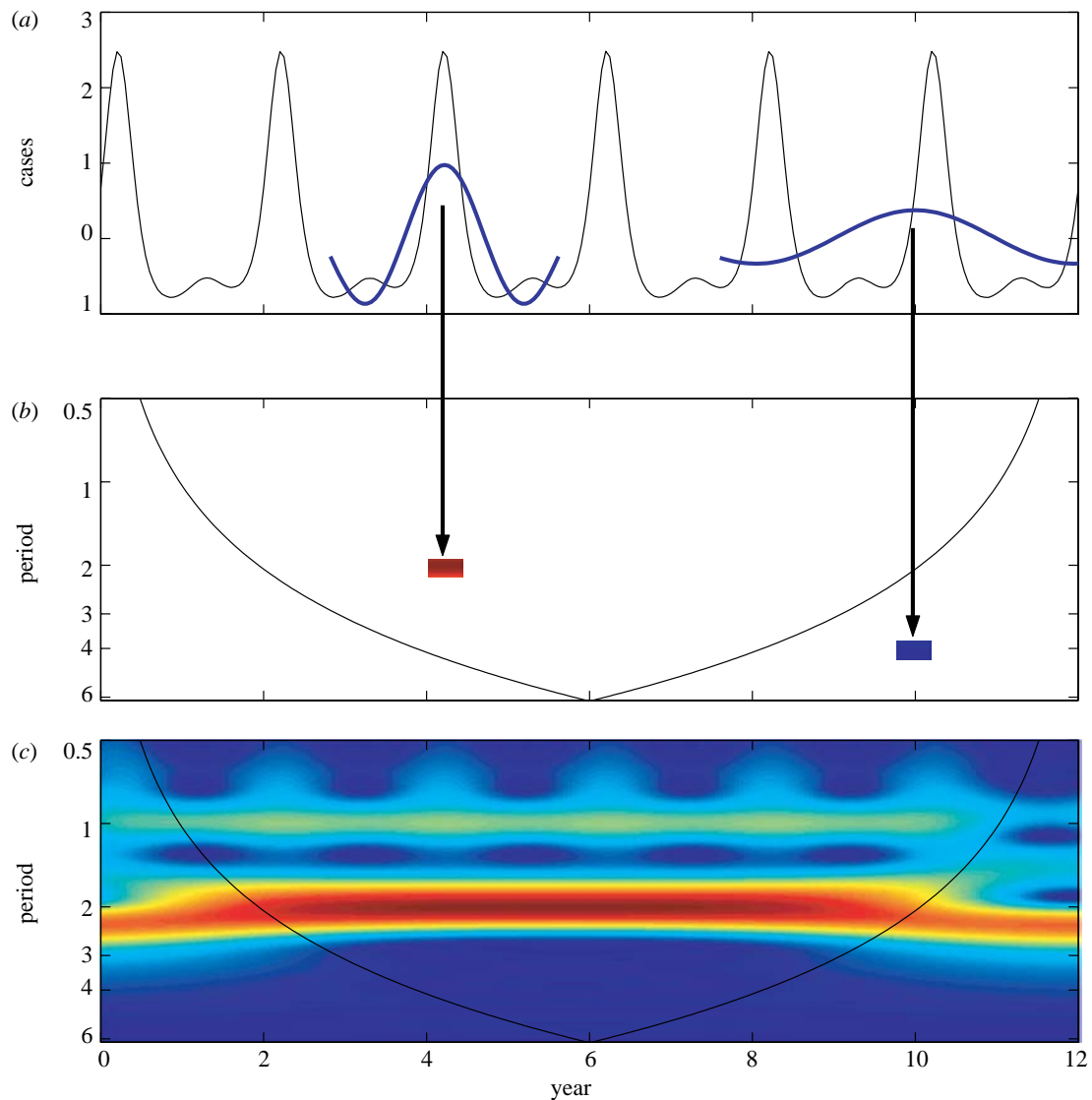
Figure 2. Wavelet power spectrum of an epidemiological time-series. (*a*) A time-series of the infectious generated by the classical SEIR model (Aron & Schwartz 1984). On this time-series Morlet wavelets with a scale (period) $a=2$-year and $a=4$-year are superimposed at time position $\tau_1=4.2$ and $\tau_2=10$, respectively. (*b*) Wavelet power spectrum ($S_x(a,\tau)$) are plotted as function of time and period in a two-dimensional graph. As an example, the graph shows the value of $S_x(a,\tau_i)$ for $a=2$-year at $\tau_1=4.2$ year and for $a=4$-year at $\tau_2=10$ year. At τ_1 for $a=2$-years, the matching between the time-series and the wavelet is high, this gives a high positive value of the wavelet transform and a high value for the wavelet power spectrum ($S_x(a,\tau)$) at this time position for this periodic component. This high value of $S_x(a,\tau_1)$ is shown in dark red in the two-dimensional time-period plot (*b*). Similarly, when the matching is weak as at τ_2 for $a=4$-years, the low value for the wavelet power spectrum is shown in dark blue (*b*). (*c*) The complete two-dimensional plot is obtained simply by computing wavelet transforms and wavelet power spectrum for a given range of a and τ values. The colours code for power values from dark blue, low values, to dark red, high values. The SEIR model used is given by: $\mathrm{d}S/\mathrm{d}t=\mu-\beta(t)SI-\mu S$; $\mathrm{d}E/\mathrm{d}t=\beta(t)SI-(\mu+\alpha)E$; $\mathrm{d}I/\mathrm{d}t=\alpha E-(\mu+\gamma)I$; $\mathrm{d}R/\mathrm{d}t=(\gamma I-\mu)R$; with $S+E+I+R=1$, μ the birth and death rates, $1/\alpha$ the duration of the latency period, $1/\gamma$ the effective infectious period and β the contact rate with $\beta(t)=\beta_0(1+\beta_1\cos 2\pi t)$ (Aron & Schwartz 1984). The parameter values used are: $\alpha=35.84$, $\mu=0.02$, $\gamma=100$, $\beta_0=1800$, $\beta_1=0.10$.

collapses. Chaves & Pascual (2006) recently described the oscillating dynamics of cutaneous leishmaniasis incidence in Costa Rica using several methods (including wavelets) and provide evidence for their association to climate variability. Constantin de Magny *et al.* (2006) documented an association between cholera incidence in Ghana and some climatic proxies (local precipitations and surface temperature) in the 1980s. In a subsequent work, they also demonstrated a significant synchrony of cholera epidemics across countries of the Gulf of Guinea, synchrony that would be driven by the association between global and local climatic forcing

(Constantin de Magny *et al.* submitted). Researchers have used the wavelet approach to compare the frequency features of simulated and observed data (e.g. Koelle & Pascual 2004). Wavelet analysis of epidemiological time-series has also been compared with other classical spectral techniques (e.g. José & Bishop 2003). Other analyses investigated the phenomenon of population synchrony where wavelets are employed to extract the phase of the time-series (e.g. Rohani *et al.* 2003 or Xia *et al.* 2004).

In the following, we present the basic ideas of the wavelet analysis and its key features illustrated with

both synthetic and observed epidemiological examples. Our main goal is to emphasize the advantages of these techniques in the context of time-series analyses in epidemiology.

2. WAVELET ANALYSIS

Although Fourier analysis is well suited to quantifying constant periodic components in a time-series, it is not able to characterize signals whose frequency content changes with time. On the other hand, a Fourier decomposition may determine all the spectral components embedded in a signal and does not provide any information about when they are present. To overcome this problem, several solutions have been developed to simultaneously decompose a signal as a function of both time t and frequency f (period or scale a).

Gabor (1946) introduced a windowed Fourier decomposition to quantify the time–frequency content of signals (figure 1). The time–frequency localization of this approach is, however, inefficient because the frequency resolution is the same for all the frequencies. A transient (with higher frequencies) needs a high time resolution to be well localized in time (figure 1). In contrast, a low-frequency structure might need a small time resolution (figure 1).

The wavelet transform decomposes a signal using functions (*wavelets*) that narrow when high-frequency features are present and widen on low-frequency structures (Daubechies 1992; Lau & Weng 1995). This decomposition yields a good localization in both time and frequency (figure 1), which is well suited for investigating the temporal evolution of aperiodic and transient signals. Indeed, wavelet analysis is the time–frequency decomposition with the optimal trade-off between time and frequency resolution (Lau & Weng 1995; Mallat 1998).

2.1. Wavelet transform

The wavelet transform of a signal $x(t)$ (a time-series) is defined as

$$W_x(a,\tau) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} x(t)\psi^*\left(\frac{t-\tau}{a}\right) \mathrm{d}t = \int_{-\infty}^{\infty} x(t)\psi_{a,\tau}^*(t)\mathrm{d}t,$$
(2.1)

where '*' denotes the complex conjugate form. In the definition, parameters a and τ denote the dilation (scale factor) and translation (time shift), respectively. The wavelet transform can be thought as a cross-correlation of a signal $x(t)$ with a set of wavelets $\psi((t-\tau)/a)$ of various 'widths' or scales a, at different time position τ (figure 2a). The wavelet coefficients, $W_x(a,\tau)$, represent thus the contribution of the scales a at different time positions τ. In equation (2.1), the factor $1/\sqrt{a}$ normalizes the wavelets so that they have unit variance and hence are comparable for all scales a.

Note that the choice of the wavelet function $\psi(t)$ is not arbitrary. This function is normalized to have unitary variance $(\int \|\psi(t)\|^2 \mathrm{d}t = 1)$ and it verifies $\int \psi(t)\mathrm{d}t = 0$. The wavelet decomposition is therefore a linear representation of the signal where the variance is preserved (Daubechies 1992). This implies that the

original signal can be recovered by means of the inverse wavelet transform,

$$x(t) = \frac{1}{C_g} \int_{-\infty}^{\infty} \int_0^{\infty} \frac{1}{a^2} W_x(a,\tau)\psi_{a,\tau}(t)\mathrm{d}\tau \, \mathrm{d}a,$$
(2.2)

where $C_g = \int_0^{\infty} \|\hat{\psi}(f)\|^2/f \, \mathrm{d}f$ and $\hat{\psi}(f)$ denotes the Fourier transform of $\psi(t)$.

The wavelet transform appears basically as a linear filter whose response function is given by the wavelet function. By means of the inverse transform, the original signal can be recovered by integrating over all scales a and locations τ, respectively. Nevertheless, one can also limit the integration over a chosen range of scales, a_1 to a_2, to perform a band-pass filtering of the original time-series in this chosen range of scales.

2.2. Different wavelets

There are several considerations in making the choice of a wavelet, for example, real versus complex wavelets, continuous or discrete wavelets, orthogonal versus redundant decompositions. Briefly, the continuous wavelets often yield a redundant decomposition (the information extracted from a given scale band slightly overlaps that extracted from neighbour scales) but they are more robust to noise as compared with other decomposition schemes. Discrete wavelets have the advantage of fast implementation but generally the number of scales and the time invariant property (a filter is time invariant if shifting the input in time correspondingly shifts the output) strongly depend on the data length. If quantitative information about phase interactions between two time-series is required, continuous and complex wavelets provide the best choice (further details can be found in Mallat 1998). However, all the wavelets share a general feature: low oscillations have good frequency and poor time resolution, whereas fast oscillations have good time resolution but a lower frequency resolution.

One particular complex continuous wavelet, the Morlet wavelet, is defined as

$$\psi(t) = \pi^{-1/4} \exp(-\mathrm{i}2\omega_0 t) \exp(-t^2/2).$$
(2.3)

This wavelet is the product of a complex sinusoidal $\exp(-\mathrm{i}2\omega_0 t)$ by a Gaussian envelope $(\exp(-t^2/2))$ where ω_0 is the central angular frequency of the wavelet. The term $\pi^{-1/4}$ is a normalization factor to ensure unit variance. For our applications we have chosen this wavelet.

For the Morlet wavelet, the relation between frequencies and wavelet scales is given by $(1/f) = ((4\pi a)/(\omega_0 + \sqrt{2+\omega_0^2}))$. When $\omega_0 \approx 2\pi$, the wavelet scale a is inversely related to the frequency, i.e. $f \approx 1/a$. This greatly simplifies the interpretation of the wavelet analysis and one can replace, on all equations, the scale a by the period $1/f$ (or wavelength).

2.3. Wavelet power spectrum

In some sense, the wavelet transform can be regarded as a generalization of the Fourier transform and by analogy with spectral approaches, one can compute the *local wavelet power spectrum* by $S_x(f,\tau) = \| W_x(f,\tau)\|^2$.

The Fourier spectrum of a signal can be compared with the global wavelet power spectrum which is defined as the averaged energy (the averaged variance) contained in all wavelet coefficients of the same scale (period) a.

$$\bar{S}_x(a) = \frac{\sigma_x^2}{T} \int_0^T \| W_x(a, \tau) \|^2 \mathrm{d}\tau, \qquad (2.4)$$

with σ_x^2 the variance of the time-series x and T the duration of the time-series. Another interesting computation is the mean variance at each time location, obtained by averaging the scale components,

$$\bar{s}_x(\tau) = \frac{\sigma_x^2 \pi^{1/4} \tau^{1/2}}{C_g} \int_0^\infty a^{1/2} \| W_x(a, \tau) \|^2 \mathrm{d}a. \qquad (2.5)$$

Figure 2 attempts to visualize the underlying processes associated with the computation of the wavelet transform and the wavelet power spectrum. In figure 2a, Morlet wavelets of scales (periods) $a = 2$ and 4-year centred at two different time locations τ_i are shown superimposed on an epidemiological time-series. The time-series used is the number of infectious cases generated by a susceptible–exposed–infectious–recovered model (SEIR) described in caption of figure 2. In the case of good matching between the time-series of the infectious $x(t)$ and the wavelets $\psi((t-\tau)/a)$, as shown for $a = 2$-year at location $\tau_1 = 4.2$ year, the integral of the product of the time-series with this wavelet (2.1) produces a large value for the wavelet transform, and then a large value for the wavelet power spectrum $S_x(a, \tau_1)$ at this position τ_1 (figure 2b). When the matching is low, as for the Morlet wavelet with $a = 4$-year at location τ_2 (figure 1a), $S_x(a, \tau_2)$ takes low values (figure 2b). By moving the wavelet along the time-series (by increasing the τ parameter), structures relating to a specific scale (period) a can be identified by high value of $S_x(a, \tau)$. This process is repeated over continuous range of a and τ for identifying all the coherent structures within the time-series producing a two-dimensional surface of $S_x(a, \tau)$ (figure 2c). In this example, with the wavelet power spectrum, we are able to recover the two dominant periodic components of the SEIR model with periods 1- and 2-year in accordance with the classical approaches (Aron & Schwartz 1984).

2.4. Wavelet coherency and phase difference

In many applications, it is desirable to quantify statistical relationships between two non-stationary signals. In Fourier analysis, the coherency is used to determine the association between two signals, $x(t)$ and $y(t)$. The coherence function is a direct measure of the correlation between the spectra of two time-series (Chatfield 1989). To quantify the relationships between two non-stationary signals, the following quantities can be computed: the *wavelet cross-spectrum* and the *wavelet coherence*.

The wavelet cross-spectrum is given by $W_{x,y}(a, \tau) = W_x(a, \tau) W_y^*(a, \tau)$ with '*' denoting the complex conjugate. As in the Fourier spectral approaches, the wavelet coherency is defined as the cross-spectrum normalized by the spectrum of each signal,

$$R_{x,y}(a, \tau) = \frac{\| \langle W_{x,y}(a, \tau) \rangle \|}{\| \langle W_{x,x}(a, \tau) \rangle \|^{1/2} \| \langle W_{y,y}(a, \tau) \rangle \|^{1/2}}, \quad (2.6)$$

where '$\langle\ \rangle$' denotes a smoothing operator in both time and scale. The smoothing can be obtained by a convolution with a constant-length window function both in the time and scale axis: $\langle W_{u,z}(a, \tau) \rangle = \int_{a-\Delta/2}^{a+\Delta/2} \int_{\tau-\delta/2}^{\tau+\delta/2} W_u(\alpha, t) W_z^*(\alpha, t) U_{\delta,\Delta}(\alpha, t) \mathrm{d}t\, \mathrm{d}\alpha$ denotes a smoothing operation with the weight function $U_{\delta,\Delta}(\alpha, t)$ that satisfies $\iint U_{\delta,\Delta}(\alpha, t) \mathrm{d}t\, \mathrm{d}\alpha = 1$ (Chatfield 1989). The values of $R_{x,y}(a, \tau)$ are thus bounded by $0 \leq R_{x,y}(a, \tau) \leq 1$. The wavelet coherency is equal to 1 when there is a perfect linear relation at particular time and scale between the two signals, and equal to 0 if $x(t)$ and $y(t)$ are independent. The advantage of these 'wavelet-based' quantities is that they may vary in time and can detect transient associations between studied time-series (Liu 1994).

As with the Morlet wavelet the $W_x(a, \tau)$ is a complex number, one can write $W_x(a, \tau)$ in terms of its phase $\phi_x(a, \tau)$ and modulus $\| W_x(a, \tau) \|$ (Le Van Quyen *et al.* 2001). The local phase of the Morlet wavelet transform is proportional to the ratio between the imaginary part ($\Im$) and the real part ($\Re$) of the wavelet transform,

$$\phi_x(a, \tau) = \tan^{-1} \frac{\Im(W_x(a, \tau))}{\Re(W_x(a, \tau))}. \qquad (2.7)$$

The phase of a given time-series $x(t)$ can be viewed as the position in the pseudo-cycle of the series and it is parameterized in radian ranging from $-\pi$ to π. The phases can then be useful to characterize possible phase relationships between $x(t)$ and $y(t)$ by computing the phase difference $\phi_{x,y}(a, \tau) = \phi_x(a, \tau) - \phi_y(a, \tau)$ or

$$\phi_{x,y}(a, \tau) = \tan^{-1} \frac{\Im(\langle W_{x,y}(a, \tau) \rangle)}{\Re(\langle W_{x,y}(a, \tau) \rangle)}. \qquad (2.8)$$

A unimodal distribution of the phase difference (for the chosen range of scales or periods) indicates that there is a preferred value of $\phi_{x,y}(a, \tau)$ and thus a statistical tendency for the two time-series to be phase locked. Conversely, the lack of association between the phase of $x(t)$ and $y(t)$ is characterized by a broad and uniform distribution. To quantify the spread of the phase difference distribution, one can use circular statistics or quantities derived from the Shannon entropy (Pikovsky *et al.* 2001; Cazelles & Stone 2003).

2.5. Assessment of statistical significance

As with other time-series methods, it is crucial to asses the statistical significance of the patterns exhibited by the wavelet approach. To this end, bootstrap methods have been used to quantify the statistical significance of the computed patterns. The idea is to construct, from observed time-series, control datasets, which share with the original series some properties but are constructed under the following null hypothesis: *the variability of the observed time-series or the association between two time-series is no different to that expected from a purely random process.* The construction of our control datasets was performed by classical bootstrap
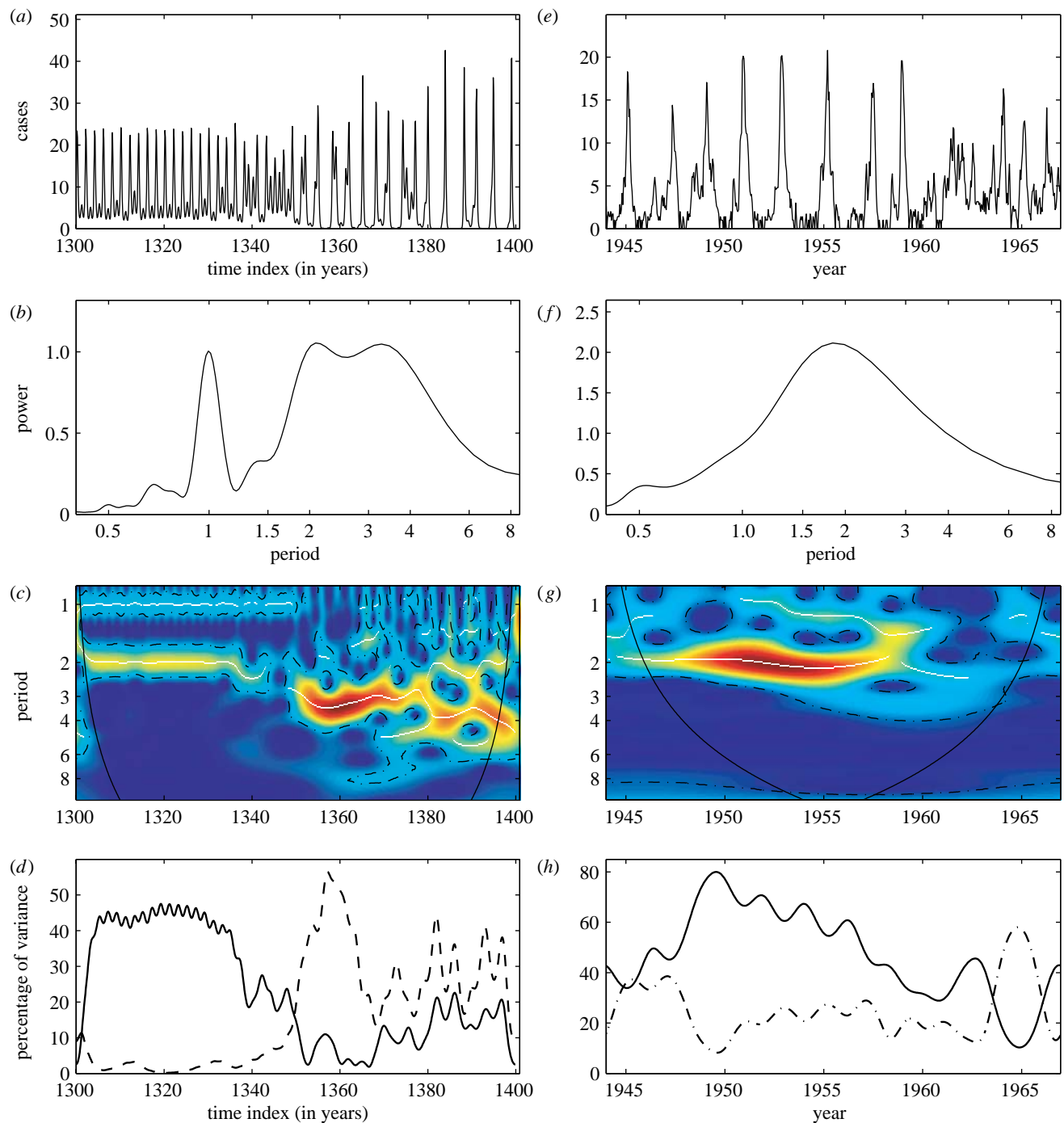
Figure 3. Wavelet analysis for the characterization of evolving epidemics. (*a–c*) Time-series of the infectious population generated by a chaotic SEIR model. The SEIR model is described in the figure 2 caption and details concerning the transients can be found in Tidd *et al.* (1993). The parameter values are identical than those used in figure 2 but $\beta_1 = 0.28$. (*d–f*) Monthly reported measles cases in York (Grenfell *et al.* 2001). (*a*) and (*e*) The time-series have been square root transformed. (*b*) and (*f*) Classical Fourier spectrum of the time-series. (*c*) and (*g*) Wavelet power spectrum ($S_x(a, \tau)$) of the time-series. The colours code for power values from dark blue, low values, to dark red, high values. On these graphs, the dotted white lines show the maxima of the undulations of the wavelet power spectrum and the dotted-dashed lines show the $\alpha = 5\%$ significant levels computed based on 1000 bootstrapped series. On the two-dimensional graphs, the cone of influence which indicates the region not influenced by edge

(Efron & Tibshirani 1993), but other resampling schemes are possible (Cazelles & Stone 2003). To test whether the raw time-series is inconsistent with the null hypothesis, we have computed the wavelet transform and related quantities for each time-series of the control set. Then we can compare the original values computed from raw series with the distribution of the same quantities under the null hypothesis, extracting, for instance, the 99th or the 95th quantiles of these distributions.

Epidemiological and environmental time-series are very often short and noisy. The values of the wavelet transform are generally corrupted as the wavelet approaches the edges of the time-series, creating a boundary effect. Further, the affected region increases
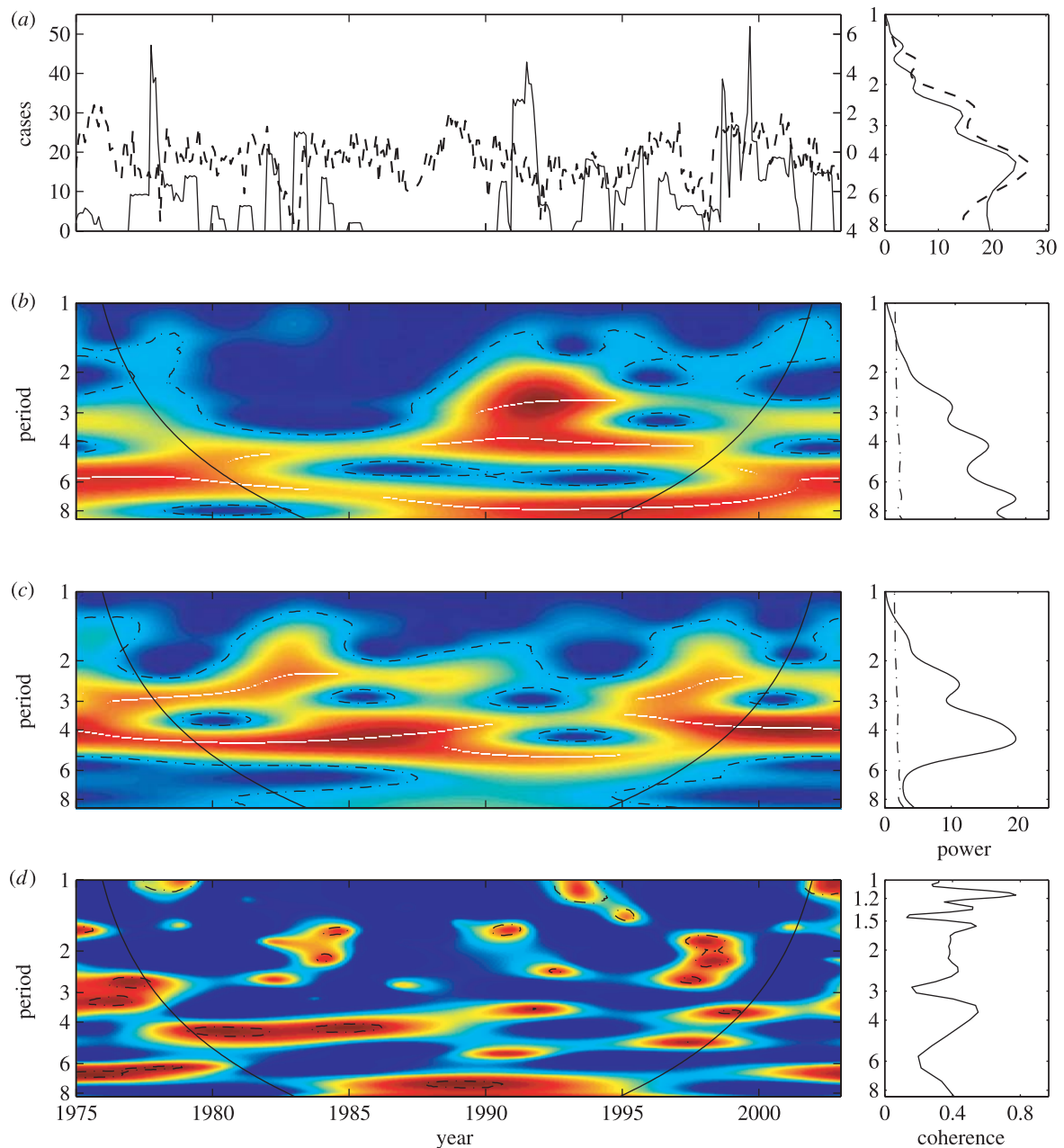
Figure 4. Wavelet analysis of the transient relationship between cholera incidence in Ghana and SOI. (*a*) (i) the analysed time-series: the cholera incidence (solid lines) and the SOI (dashed lines). The incidence series are square root transformed, and all series are filtered and normalized before analyses. (*a*) (ii) the Fourier spectrum for the cholera incidence series (solid line) and for the SOI (dashed line). The y-axis describes period (in year) as for other *y*-axis of (*b*–*d*) graphs. (*b*) (i) wavelet power spectrum ($S_x(a, \tau)$) of the cholera incidence. The colours code for power values from dark blue, low values, to dark red, high values. On this graph, the white line shows the maxima of the undulations of the wavelet power spectrum. (*b*) (ii) the average wavelet spectrum ($\bar{S}_x(a)$) of cholera incidence. (*c*) as in (*b*) but for the SOI. (*d*) (i) Wavelet coherence between cholera incidence in Ghana and the SOI. The colours are coded a dark blue, low coherence and dark red, high coherence. (*d*) (ii) Classical Fourier coherence between cholera incidence and the SOI. On (*b*–*d*), the dotted-dashed lines showed the $\alpha = 5\%$ significant levels computed based on 1000 bootstrapped series and on the two-dimensional graphs the cone of influence, which indicates the region not influenced by edge effects, is also shown.

in extent as the scale (period) parameter *a* increases. This region is known as the cone of influence (Torrence & Compo 1998) and the spectral information within this cone is likely to be less accurate.

2.6. Available wavelet toolboxes

Several wavelet toolboxes are catalogued on the Amara's wavelet page (www.amara.com/current/ wavelet.html). A general time–frequency toolbox developed for GNU Octave and Matlab is available at tftb.nongnu.org/. Another toolbox widely used in epidemiology and geosciences is that of Torrence & Compo (1998) (atoc.colorado.edu/research/wavelets/). Nevertheless, these toolboxes only consider univariate time-series analysis. Other collections of functions have been recently developed to explore bivariate time-series (including wavelet coherence) in geosciences:

Maraun & Kurths (2004) propose a set of R functions (www.agnld.uni-potsdam.de/∼maraun/wavelets/) whereas Grinsted *et al.* (2004) have extended the Matlab functions of Torrence & Compo (www.pol.ac.uk/home/research/waveletcoherence/). The Matlab functions used in our research, as well as some examples of analysis of epidemiological time-series can be downloaded at ecologie.snv.jussieu.fr/cazelles/wavelets/.

3. ANALYSIS OF EPIDEMIOLOGICAL TIME-SERIES

3.1. *Characterization of time-evolving epidemics*

The wavelet approach enables the evolution of the oscillating characteristics of a given time-series to be described. To illustrate this, we use a synthetic time-series generated by a classical SEIR model with chaotic and transient dynamics (as described in the caption of figure 2; Tidd *et al.* 1993; Engbert & Drepper 1994).

The results of wavelet analysis are summarized in figure 3a–d. Figure 3a displays the transient dynamics of the infectious population simulated by the SEIR model. Figure 3b displays the classical Fourier spectrum (Chatfield 1989) of this time-series. There are large peaks at a period of 1, 2 and around 3–4 years. Although these periodic modes are present in the time-series (figure 3a), the Fourier spectrum is not able to specify when these modes are present. Conversely, the wavelet power spectrum can quantify the time evolution of these oscillatory modes and show when they are dominant (figure 3c).

This synthetic time-series is dominated by a 2-year mode (and its 1-year subharmonic) before time index 1340 (figure 3c). After this, the dynamics become more complex (chaotic, see Tidd *et al.* 1993) with numerous significant modes of oscillation, but the 3–4 year mode dominates (figure 3c and also figure 3b). Note that the wavelet power spectrum is coherent with the time-series, with a major shift in the periodicity around time index 1340. We can quantify the shift in the dominant mode by estimating the variance of the time-series explained by different modes using the scale-averaged wavelet power (2.5). Figure 3d shows the time evolution of the percentage of the variance explained by the 2- and the 3–4 year modes. In the first part of the time-series before the time index 1340, the 2-year mode explains more than 40% of the variance. After this, the 3–4 year mode becomes dominant, explaining more than 30% on an average of the variance.

The above example shows that wavelet analysis allows the detection of shifts in the periodic components of a given time-series. To illustrate the method using an observed time-series, we have analysed weekly measles notifications in the city of York (UK) for the pre-vaccination era, 1944–1966 (figure 3e–g; data from Grenfell *et al.* 2001). This series is dominated by the 2-year periodic mode (figure 3f). As with classical approaches, the wavelet power spectrum reveals not only the importance of this oscillatory mode, but also its time localization. Before 1960, the 2-year mode dominates (figure 3g) as for other cities (Grenfell *et al.*

2001) but after 1960, this periodic mode disappears and is replaced by more aperiodic dynamics (figure 3g). This observation is supported by examining the time evolution of the percentage of the variance explained by the 2-year mode. Figure 3h shows that the 2-year mode explains on average more than 50% of the variance before 1960 and a significant decrease of the dominance of this mode around 1960. This graph (figure 3h) also shows the rapid alternation of the 2-year and the 1-year modes after 1960.

3.2. *Analysis of transient relationship between epidemics and environmental forcing*

Cholera is a highly contagious disease caused by specific strains of the bacterium *Vibrio choleræ* that is naturally present in the environment and is autochthonous to coastal and estuarine ecosystems (Colwell 1996). Cholera is known to be influenced by climate and ocean conditions (Bouma & Pascual 2001; Pascual *et al.* 2002). Previous studies of cholera population dynamics have proposed links between oceanographic environmental conditions and human cases of cholera morbidity and mortality (Epstein 1993; Colwell & Huq 2001). Climate may also affect the dynamics of cholera by shifting pathogen abundance, host species abundance, population dynamics or community interaction. For example, cholera outbreaks in Peru and Bangladesh have been linked to periodic climatic cycles such as the El Niño Southern oscillation (ENSO; Colwell 1996; Salazar-Lindo *et al.* 1997; Pascual *et al.* 2000; Rodó *et al.* 2002). For example, Rodó *et al.* (2002) used singular spectral analysis and Fourier analysis in order to isolate the main interannual variability in long-term Bangladesh cholera time-series. In this study, the authors showed that the strong association between cholera dynamic and ENSO is discontinuous in time owing to the shifts in the ENSO frequency spectrum. They needed to analyse data in three successive time periods, with classical Fourier spectrum, in order to capture these changes. Our initial aim was to determine whether the associations between ENSO and cholera outbreaks in Ghana (Western Africa) are similar to those observed in Peru and/or Bangladesh, but with a method that potentially is able to capture the non-stationary features of the observed dynamics if present. Then we analysed with wavelets reported cholera cases for Ghana for the years 1975–2002, compiled from the weekly epidemiological record (available at the WHO website: www.who.int/wer/en/). As a proxy for ENSO, we used the southern oscillation index (SOI) for the same years (www.cpc.ncep.noaa.gov/data/indices/soi).

Figure 4a (i) displays the observed time-series. Results from the classical Fourier spectral analysis tell us that both time-series have similar oscillating components mainly around the 2–3 year and the 4–5 year modes (figure 4a, ii). This could suggest a possible association between cholera dynamics and ENSO. These associations estimated by the classical coherence (Chatfield 1989) are shown to be high for the 1.2-year, the 2.5-year and the 4-year modes (figure 4d, ii). Nevertheless, the results of the wavelet analysis lead to more restrained conclusions.

For cholera incidence in Ghana, the wavelet power spectrum shows a continuous oscillating mode at both 3–5 year and 6–8 year during the whole time period (figure 4b). Nevertheless, these modes of oscillation vary in strength. The dominant modes are the 3-year mode for 1989–1995, the 4-year mode for 1989–1996, around the 6-year mode for 1975–1985 and the 8-year mode for 1990–1998 (figure 4b, i). Wavelet analysis of SOI detected a significant 4-year periodic mode for the whole time-series but the SOI wavelet power spectrum is characterized by a strong 4-year periodic component during both the 1980–1989 and 1995–2002 time periods (figure 4c, i). During 1989–1995, one observed a shift of the dominant mode to the 5-year periodic band. The right of figure 4c,d display the average wavelet power spectrum (2.4) that are very similar than those obtained with the Fourier method (figure 4a, ii). Although wavelet analysis reveals that these two time-series have similar oscillating components, these components are not always present at the same time. In fact, the wavelet coherence (that quantifies the association between the time-series) is significant only for 1980–1985 around the 4-year mode (figure 4d, i). These results are in contrast to those obtained with the classical approach (figure 4d, ii).

Wavelet analysis suggests that the incidence of cholera cases in Ghana is weakly coherent with ENSO except for 1980–1985 in the 4-year periodic band. This pattern of association appears more complicated than those observed in other regions (Salazar-Lindo *et al.* 1997; Pascual *et al.* 2000). This particular analysis reveals that the timing of association between disease dynamics and environmental forcing must not be disregarded. Wavelet analysis is one of the statistical tools with the power account for such time-evolving relationships.

3.3. Analysis of spatial synchrony of epidemics

Using monthly data for Thailand, Cummings *et al.* (2004) identified travelling waves of dengue hemorrhagic fewer (DHF), initiated in the capital city, Bangkok. In subsequent research, it has been demonstrated that this synchrony is transient and linked to El Niño (Cazelles *et al.* 2005).

The data are monthly reports of DHF in each of 72 provinces of Thailand (Cummings *et al.* 2004). We analysed two incidence time-series from this dataset: the incidence in Bangkok and the averaged incidence for the rest of Thailand (figure 5a). The oscillations of the dengue incidence time-series are dominated by the annual mode, but they also have a statistically significant common mode of oscillation around a period of 2–3 year (Cazelles *et al.* 2005).

We computed the different temporal associations between the DHF incidence in Bangkok and the rest of Thailand, showing three main regions of high and significant coherence (figure 5b). The first one is for the 2–3 year periodic band for the time period 1983–1991, the second is for the 1-year mode for 1983–1985 and for 1991–1997, and the last for the 5-year mode after 1988. We also analysed the phases of the time-series that permit to obtain information about the possible delay in the relationship (i.e. in phase or out of phase

relations; figure 5c). This phase analysis is completed by the computation of the evolution of periodic components in the 2–3 year and the 1-year modes for DHF in Bangkok and in the rest of Thailand (figure 5d,e). The two incidence series are phase locked with a mean delay of three months for the 2–3 year band (figure 5c,d), but mainly within the period 1983–1992 where there is high coherence with El Niño oscillations (Cazelles *et al.* 2005). For 1983–1986 and 1991–1997, the seasonal mode is dominant. During these years, phase locking is seen only for the seasonal mode and the dengue incidence in Bangkok follows the incidence in the remainder of Thailand with an average delay of one month (figure 5e).

The wavelet allows us to explore the transient nature of the spatial synchrony of dengue incidence in Thailand and the potential influence of El Niño on these complex phenomena. With wavelets, we are able to state that the observed travelling waves in the 2–3 year mode described by Cummings *et al.* (2004) coincides with a period of high coherence between dengue and El Niño, convincingly suggesting that climate may be involved in these travelling waves (Cazelles *et al.* 2005).

4. CONCLUDING REMARKS

Wavelet analysis is an important addition to time-series methods with practical applications in epidemiology. We have reviewed the basic concepts of the wavelet analysis and illustrated the approach using examples, which appear particularly interesting from an epidemiological point of view. Wavelet analysis can help us to interpret multi-scale, non-stationary time-series data and reveals features we could not see otherwise. This is clearly the case for both the complex relationship between cholera incidence in Ghana and SOI and for the transient spatial synchrony of DHF in Thailand.

The major aims of much current epidemiological research are to characterize and understand disease processes and the potential influence of exogenous changes on these processes. In this context, as experiment is usually not possible, the retrospective (historical) approaches are frequently employed. Retrospective approaches use a mode of analysis, which is dependent on the comparative and observational richness of the data. A key requirement is to take into account the major characteristics of the observations that mirror the underlying properties of the system. As stressed in §1, epidemiological and environmental time-series observed are typically non-stationary: epidemiological time-series can change dramatically with time. These characteristics may make use of traditional techniques inappropriate and results from classical approaches must be interpreted with caution.

Using the wavelet approach, we have shown that it is possible to study irregular, non-stationary and noisy time-series, and also to analyse weak and transient interactions between such series. Wavelet power spectra quantify the main periodic component of a given time-series and its time evolution. Wavelet coherence is used to quantify the degree of linear relation between two non-stationary time-series in the time–frequency domain. The main advantage of the wavelet approach
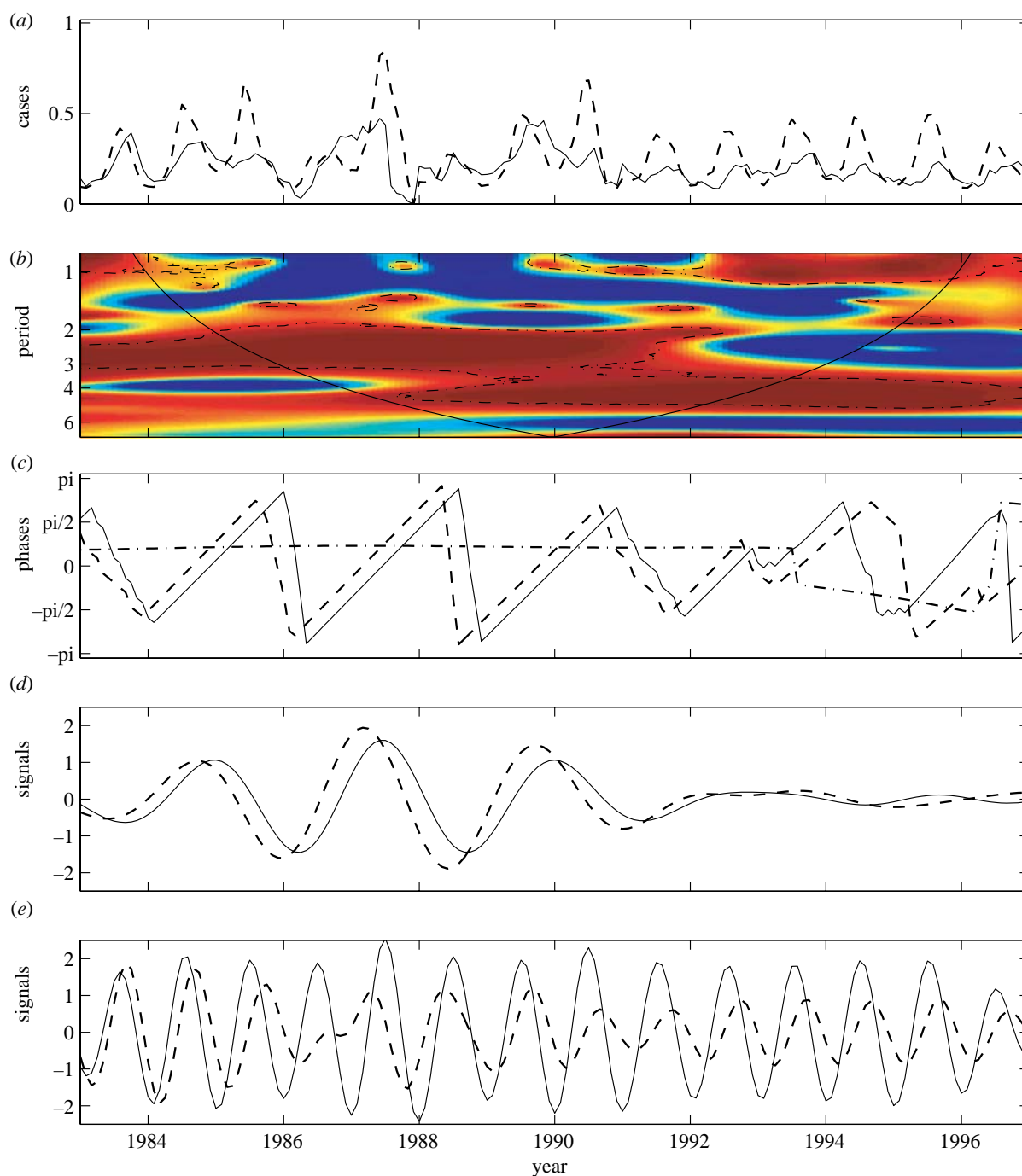
Figure 5. Wavelet analysis of the synchrony between dengue incidence in Bangkok and in the rest of Thailand. The incidence series are square root transformed and all series are normalized. (*a*) Bangkok dengue incidence (solid line) and Thailand dengue incidence (dashed line). (*b*) Wavelet coherence between dengue incidence in Bangkok and in the rest of Thailand. The colours are coded a dark blue, low coherence and dark red, high coherence. The dotted-dashed lines show the $\alpha = 5\%$ significance levels computed based on 1000 bootstrapped series. The cone of influence indicates the region not influenced by edge effects. (*c*) Phase evolutions of the two series computed with the wavelet transform in the 2–3 year periodic band. Line symbols are as in (*a*) and the dotted-dashed line is for the phase difference between the two series. (*d*) Oscillating components computed with the wavelet transform in the 2–3 year periodic band (line symbols as in (*a*)). (*e*) Oscillating components computed with the wavelet transform in the 0.8–1.2 year periodic band (line symbols as in (*a*)).

is the ability to analyse transient dynamics, for one-dimensional signals and for the association between two time-series. A further interesting aspect is the analysis of the phases of the studied time-series (Grenfell *et al.* 2001; Cazelles & Stone 2003). Phase analysis is a nonlinear technique that makes possible to study rather weak interactions (Cazelles & Stone 2003). Indeed, the notion of phase synchronization implies only some

inter-dependence between phases, whereas the irregular amplitudes may remain uncorrelated (see Rohani *et al.* 2003 or Xia *et al.* 2004).

As a statistical analysis, wavelet approaches provide no information about the underlying epidemiological mechanisms. There is no single relation neither between cyclical features and biological or epidemiological mechanisms nor between relationships and mechanisms

since a given pattern of association between series may be generated by a wide variety of different mechanisms. However, like the other correlative approaches, wavelet analysis can provide useful clues about the nature of the underlying epidemiological processes (e.g. Grenfell *et al.* 2001 or Cazelles *et al.* 2005). Such clues pave the way for future modelling approaches in which explicit mechanisms can be incorporated. A great strength of wavelet analysis is its freedom from stationarity assumptions, providing a condensed, quantitative, temporally explicit summary of time-series. Wavelet analysis can be used as a first step for exploring the complexity of both the observed environmental and the epidemiological signals before the modelling stage. But this modelling approach must also take into account these non-stationary features. In this context, Bayesian approaches such as Kalman filtering (e.g. Cazelles & Chau 1997) or particle filters (e.g. Andrieu *et al.* 2001) seem very promising.

We hope that wavelet analysis will be more widely employed to analyse epidemiological time-series. The wavelet approach should be employed alongside other time-series tools to examine directly the relationship (association, dependence, synchrony) between studied time-series. We believe that wavelet analysis should yield significant future advances in our understanding of epidemiological process in our changing world.

REFERENCES

Anderson, R. M. & May, R. M. 1991 *Infectious diseases of humans.* Oxford, UK: Oxford University Press.

Andrieu, C., Doucet, A. & Fitzgerald, W. J. 2001 An introduction to Monte Carlo methods for Bayesian data analysis. In *Nonlinear dynamics and statistics* (ed. A. I. Mees), pp. 169–217. Boston, MA: Birkhäusser.

Aron, J. L. & Schwartz, I. R. 1984 Seasonality and period-doubling bifurcations in an epidemic model. *J. Theor. Biol.* **110**, 665–679.

Bell, M. L., Samet, J. M. & Dominici, F. 2004 Time-series studies of particulate matter. *Annu. Rev. Public Health* **25**, 247–280. (doi:10.1146/annurev.publhealth.25.102802.124329)

Bishop, Y. M. M. 1977 Statistical methods for hazards and health. *Environ. Health Perspect.* **20**, 149–157.

Bloomfield, P. 1976 *Fourier analysis of time series: an introduction.* New York, NY: Wiley.

Bouma, M. J. & Pascual, M. 2001 Seasonal and interannual cycles of endemic cholera in Bengal 1891–1940 in relation to climate and geography. *Hydrobiologia* **460**, 147–156. (doi:10.1023/A:1013165215074)

Broutin, H., Guégan, J.-F., Elguero, E., Simondon, F. & Cazelles, B. 2005 Large-scale comparative analysis of pertussis population dynamics: periodicity, synchrony and impact of vaccination. *Am. J. Epidemiol.* **161**, 1159–1167. (doi:10.1093/aje/kwi141)

Catalano, R. & Serxner, S. 1987 Time series designs of potential interest to epidemiologists. *Am. J. Epidemiol.* **26**, 724–731.

Cazelles, B. & Chau, N. P. 1997 Using the Kalman filter and dynamic models to assess the changing HIV/AIDS epidemic. *Math. Biosci.* **140**, 131–154. (doi:10.1016/S0025-5564(96)00155-1)

Cazelles, B. & Stone, L. 2003 Detection of imperfect population synchrony in an uncertain world. *J. Anim. Ecol.* **72**, 953–968. (doi:10.1046/j.1365-2656.2003.00763.x)

Cazelles, B., Chavez, M., McMichael, A. J. & Hales, S. 2005 Nonstationary influence of El Niño on the synchronous dengue epidemics in Thailand. *PLoS Med.* **2**, 313–318.

Chatfield, J. R. 1989 *The analysis of time series: an introduction.* London, UK: Chapman and Hall.

Chaves, L. F. & Pascual, P. 2006 Climate cycles and forecasts of Cutaneous Leishmaniasis, a non-stationary vector borne disease. *PLoS Med.* **3**, 1320–1328. (doi:10.1371/journal.pmed.0030295)

Colwell, R. R. 1996 Global climate and infectious disease: the cholera paradigm. *Science* **274**, 2025–2031. (doi:10.1126/science.274.5295.2025)

Colwell, R. R. & Huq, A. 2001 Marine ecosystems and cholera. *Hydrobiologia* **460**, 141–145. (doi:10.1023/A:1013111016642)

Constantin de Magny, G., Cazelles, B. & Guégan, J. F. 2006. Disease cholera threat to humans in Ghana, Western Africa, under the influence of both global and regional climatic variability. *EcoHealth* **3**, 223–231.

Constantin de Magny, G., Guégan, J. F., Petit, M. & Cazelles, B. Submitted. Regional-scale synchrony of cholera epidemics in western Africa with climate variability.

Cummings, D. A. T., Irizarry, R. A., Huang, N. E., Endy, T. P., Nisalak, A., Ungchusak, K. & Burke, D. S. 2004 Travelling waves in the occurrence of dengue haemorrhagic fever in Thailand. *Nature* **427**, 344–347. (doi:10.1038/nature02225)

Daubechies, I. 1992 *Ten lectures on wavelets.* SIAM monographs. Philadelphia, PA:SIAM.

Dominici, F., McDermott, A., Zeger, S. L. & Samet, J. M. 2002 On the use of generalized additive models in time-series studies of air pollution and health. *Am. J. Epidemiol.* **156**, 193–203. (doi:10.1093/aje/kwf062)

Duncan, C. J., Duncan, S. R. & Scott, S. 1996 Whooping cough epidemics in London, 1701–1812: infection dynamics, seasonal forcing and the effects of malnutrition. *Proc. R. Soc. B* **263**, 445–450. (doi:10.1098/rspb.1996.0067)

Efron, B. & Tibshirani, R. J. 1993 *An introduction to the bootstrap.* London, UK: Chapman and Hall.

Engbert, R. & Drepper, F. R. 1994 Chance and chaos in population biology: models of recurrent epidemics and food chain dynamics. *Chaos, Solitons Fractals* **4**, 1147–1169. (doi:10.1016/0960-0779(94)90028-0)

Epstein, P. R. 1993 Algal blooms in the spread and persistence of cholera. *Biosystems* **31**, 209–221. (doi:10.1016/0303-2647(93)90050-M)

Gabor, D. 1946 Theory of communication. *J. Inst. Electr. Eng.* **93**, 429–457.

Greenwood, M. 1924 Is the statistical method of any value in medical research? *Lancet* **2**, 153–158.

Grenfell, B. T., Bjørnstad, O. N. & Kappey, J. 2001 Travelling waves and spatial hierarchies in measles epidemics. *Nature* **414**, 716–723. (doi:10.1038/414716a)

Grinsted, A., Moore, J. C. & Jevrejeva, S. 2004 Application of the cross wavelet transform and wavelet coherence to geophysical time series. *Nonlin. Process. Geophys.* **11**, 561–566.

Helfenstein, U. 1986 Box-Jenkins modelling of some viral infectious diseases. *Stat. Med.* **5**, 37–47.

José, M. V. & Bishop, R. F. 2003 Scaling properties and symmetrical patterns in the epidemiology of rotavirus infection. *Phil. Trans. R. Soc. B* **358**, 1625–1641. (doi:10.1098/rstb.2003.1291)

Koelle, K. & Pascual, M. 2004 Disentangling extrinsic from intrinsic factors in disease dynamics: a nonlinear time series approach with an application to cholera. *Am. Nat.* **163**, 901–913. (doi:10.1086/420798)

Lau, K. M. & Weng, H. 1995 Climatic signal detection using wavelet transform: how to make a time series sing. *Bull. Am. Meteorol. Soc.* **76**, 2391–2402. (doi:10.1175/1520-0477(1995)076<2391:CSDUWT>2.0.CO;2)

Le Van Quyen, M., Foucher, J., Lachaux, J. P., Rodriguez, E., Lutz, A., Martinerie, J. & Varela, F. 2001 Comparison of Hilbert transform and wavelet methods for the analysis of neuronal synchrony. *J. Neurosci. Methods* **111**, 83–98. (doi:10.1016/S0165-0270(01)00372-7)

Liu, P. C. 1994 Wavelet spectrum analysis and ocean wind waves. In *Wavelets in geophysics* (eds E. Foufoula-Georgiou & P. Kumar), pp. 151–166. New York, NY: Academic Press.

Mallat, S. 1998 *A wavelet tour of signal processing*. San Diego, CA: Academic Press.

Maraun, D. & Kurths, J. 2004 Cross wavelet analysis: significance testing and pitfalls. *Nonlin. Process. Geophys.* **11**, 505–514.

McCarthy, J. J., Canziani, O. F. Leary, N. A., Dokken, D. J. & White, K. S. 2001 Climate change 2001: impacts, adaptation and vulnerability. International Panel of Climate Change Report. Cambridge University Press, Cambridge. See http://www.grida.no/climate/ipcc_tar/wg2/index.htm.

McMichael, A. J., Woodruff, R. W. & Hales, S. 2006 Climate change and human health: present and future risks. *Lancet* **367**, 859–869. (doi:10.1016/S0140-6736(06)68079-3)

Morens, D. M., Folkers, G. K. & Fauci, A. S. 2004 The challenge of emerging and re-emerging infectious diseases. *Nature* **430**, 242–249. (doi:10.1038/nature02759)

Pascual, M., Rodó, X., Ellner, S. P., Colwell, R. & Bouma, M. J. 2000 Cholera dynamics and El Niño-southern oscillation. *Science* **289**, 1766–1769. (doi:10.1126/science.289.5485.1766)

Pascual, M., Bouma, M. J. & Dobson, A. P. 2002 Cholera and climate: revisiting the quantitative evidence. *Microbes Infect.* **4**, 237–245. (doi:10.1016/S1286-4579(01)01533-7)

Patz, J. A., Campbell-Lendrum, D., Holloway, T., Jonathan, A. & Foley, J. A. 2005 Impact of regional climate change on human health. *Nature* **438**, 541–543. (doi:10.1038/nature04188)

Pikovsky, A. S., Rosenblum, M. G. & Kurths, J. 2001 *Synchronization: a universal concept in nonlinear sciences.* Cambridge, UK: Cambridge University Press.

Rodó, X., Pascual, M., Fuchs, G. & Faruque, S. G. 2002 ENSO and cholera: a nonstationary link related to climate change? *Proc. Natl Acad. Sci. USA* **99**, 12 901–12 906. (doi:10.1073/pnas.182203999)

Rohani, P., Earn, D. J. D. & Grenfell, B. T. 1999 Opposite patterns of synchrony in sympatric disease metapopulations. *Science* **286**, 968–971. (doi:10.1126/science.286.5441.968)

Rohani, P., Green, C. J., Mantilla-Beniers, N. B. & Grenfell, B. T. 2003 Ecological interference between fatal diseases. *Nature* **422**, 885–888. (doi:10.1038/nature01542)

Ross, R. 1911 Some quantitative studies in epidemiology. *Nature* **87**, 466–467.

Salazar-Lindo, E., Pinell-Salles, P., Maruy, A. & Chea-Woo, E. 1997 El Niño and diarrhoea and dehydration in Lima, Peru. *Lancet* **350**, 1597–1598. (doi:10.1016/S0140-6736(05)64013-5)

Schwartz, J. 1993 Air pollution and daily mortality in Birmingham, Alabama. *Am. J. Epidemiol.* **137**, 1136–1147.

Tidd, C. W., Olsen, L. F. & Schaffer, W. M. 1993 The case for chaos in childhood epidemics: II Predicting historical epidemics from mathematical models. *Proc. R. Soc. B* **254**, 257–273. (doi:10.1098/rspb.1993.0155)

Torrence, C. & Compo, G. P. 1998 A practical guide to wavelet analysis. *Bull. Am. Meteorol. Soc.* **79**, 61–78. (doi:10.1175/1520-0477(1998)079<0061:APGTWA>2.0.CO;2)

Woolhouse, M. E. J., Haydon, D. T. & Antia, R. 2005 Emerging pathogens: the epidemiology and evolution of species jumps. *Trends Ecol. Evol.* **20**, 238–244. (doi:10.1016/j.tree.2005.02.009)

Xia, Y., Bjørnstad, O. N. & Grenfell, B. T. 2004 Measles metapopulation dynamics: a gravity model for epidemiological coupling and dynamics. *Am. Nat.* **164**, 267–281. (doi:10.1086/422341)