‘Wavelets’ entry for *Encyclopedia of Environmetrics*

by

D. B. Percival

Applied Physics Laboratory, Box 355640, University of Washington, Seattle, WA 98195–5640

MathSoft, Inc., 1700 Westlake Ave. N, Seattle, WA 98109

INTRODUCTION

Wavelets are a special class of functions (or sequences) that are widely used for analyzing time series, i.e., a sequence of observations recorded over time (an example of such a series is plotted at the bottom of Figure 3, which shows deseasonalized monthly average temperature anomalies in the Northern Hemisphere formulated once per month from January 1856 onward [30]). Just as Fourier analysis is based upon the notion of representing (or re-expressing) a time series as a linear combination of sinusoids, the idea underlying wavelet analysis is to represent the series as a linear combination of wavelets. In Fourier analysis, each sinusoid is associated with a particular frequency f, so we can deduce what frequencies are important in a particular time series by studying the magnitudes of the coefficients of the various sinusoids in the linear combination. In contrast, each wavelet is associated with two independent variables, namely, time t and scale τ, because each wavelet is essentially nonzero only inside a particular interval of times, namely, $[t-\tau, t+\tau]$. Within that interval, the wavelet spends roughly an equal amount of time above and below zero, so it appears to be a 'small wave' centered at time t and having a width of 2τ. We can thus learn how a time series varies on particular scales across time if we re-express it using wavelets. (Although we concentrate entirely on data sampled over time here, in fact wavelets are used extensively with data sampled over other independent variables, including (i) two dimensional grids, (ii) parametric curves within a two dimensional surface and (iii) three dimensional objects [22, 31].)

As the term is presently used, wavelets as a subject go back to a seminal paper by Goupillaud et al. [14] in 1984 in the geophysical literature, but the topic is really a mixture of fairly old concepts along with some exciting new ideas and algorithms. The vitality of the subject is in part due to the fact that the theory and application of wavelets have formed bridges among the often disparate worlds of mathematics, statistics, signal processing and various physical sciences (including all that touch upon environmetrics). Some idea of the current level of interest in wavelets can be gleaned from the following: as of March 2000, one database of articles in the physical and engineering sciences listed over 10,000 articles and books when queried with the keyword 'wavelet' – this was an increase of over 2000 articles from a similar query done in March 1999!

There are two main classes of wavelets. The first includes various forms of the continuous wavelet transform (CWT) and was the main focus of waveleticians in the 1980s. The second class comprises the orthonormal discrete wavelet transform (DWT) and related transforms, which came into prominence starting in the late 1980s (the DWT we concentrate on below is related to a CWT, but in a rather subtle manner). In what follows, we introduce wavelets by considering the CWT of a 'signal' $x(t)$, $-\infty < t < \infty$, which we take to be a real-valued square integrable function (i.e., $\int x^2(t)\,dt < \infty$). We then describe the orthonormal DWT of a time series X_t, $t = 0, \ldots, N-1$, which we can often regard as a finite number of samples from a signal; i.e., $X_t = x(t\,\Delta t)$, where Δt is the spacing in time between adjacent observed values. After a description of an efficient 'pyramid' algorithm for computing the DWT, we discuss two interesting descriptive statistics that are based on the analysis/synthesis capabilities of the DWT, namely, a decomposition of the sample variance of a time series and an additive decomposition known as a 'multiresolution analysis.' We briefly describe a variation of the DWT called the 'maximal overlap' DWT (MODWT), which is one version of a 'shift invariant' DWT. We then review some important applications for the DWT and MODWT in the statistical analysis of time series, for which we will then assume that X_t is a random variable constituting the tth element of a stochastic process.

Because this article is an overview of wavelets and their applications, we must gloss over many technical details, which the reader can find in a book by the current author and A. T. Walden [28]. There are several good books that address the statistical application of wavelets and would be useful for a reader who wants to delve more into the subject, including Bruce and Gao [4], Carmona et al. [5], Mallat [22], Ogden [26], Vidakovic [33] and Wornell [36]. Commerical software for wavelets includes S+Wavelets for S-Plus and the Wavelets Extension Pack for Mathcad (MathSoft); Wavelet Explorer for Mathematica (Wolfram Research); and WavBox (Computational Toolsmiths) and the Wavelet Toolbox (MathWorks), both for Matlab. Public domain software includes WaveThresh for S-Plus (University of Bristol) and WaveLab for Matlab (Stanford Unversity). In addition, anyone interested in wavelets can benefit from perusing the Wavelet Digest at `http://www.wavelet.org`, which regularly issues newsletters and maintains other useful information for the wavelet community.

THE CONTINUOUS WAVELET TRANSFORM

Formally, a real-valued function $\psi(t)$ is called a wavelet if it satisfies the following assumptions:

$$\text{(i)} \int_{-\infty}^{\infty} \psi(t)\,dt = 0 \quad \text{and (ii)} \int_{-\infty}^{\infty} \psi^2(t)\,dt = 1 \tag{1}$$

(this is a bare bones definition: in practice, we must also impose a technical – but relatively mild – assumption known as the admissibility condition if we want to be able to reconstruct $x(t)$ from its wavelet transform). Assumption (ii) says that, for every small $\epsilon > 0$, there is some $T > 0$ such that

$$\int_{-T}^{T} \psi^2(t)\,dt < 1 - \epsilon.$$

While $[-T, T]$ might be quite large, this interval is still vanishingly small when compared to the entire real axis. Assumption (ii) also means that $\psi(t)$ must be nonzero somewhere, while assumption (i) says that $\psi(t)$ balances itself above and below zero. From these two assumptions we can picture a function that is basically time limited and oscillates up and down at least once; i.e., $\psi(t)$ is a 'small wave' or 'wavelet' (in contrast, we would consider the sinusoids $\sin(2\pi ft)$ underlying Fourier analysis as 'big waves' because they never damp down toward zero as $|t|$ gets large).

Figure 1 goes about here.

Figure 1 shows plots of two wavelets. The left-hand plot is of the Haar wavelet, which is defined as

$$\psi^{(\mathrm{H})}(u) \equiv \begin{cases} -1/\sqrt{2}, & -1 < u \le 0; \\ 1/\sqrt{2}, & \quad 0 < u \le 1; \\ 0, & \text{otherwise} \end{cases}$$

(this is arguably the first wavelet since it appeared in a 1910 article by Haar [15]). The other wavelet $\psi^{(\mathrm{Mh})}(u)$ is proportional to the second derivation of the Gaussian (or normal) probability density function. This wavelet is called the 'Mexican hat' wavelet for obvious reasons.

We can use the Haar wavelet to tell us something about how localized averages of a signal $x(t)$ vary across time. To quantify this description, let

$$A(\tau, t) \equiv \frac{1}{\tau} \int_{t-\frac{\tau}{2}}^{t+\frac{\tau}{2}} x(u)\,du.$$

Elementary books on calculus refer to $A(\tau, t)$ as the average value of $x(t)$ over the interval $[t - \frac{\tau}{2}, t + \frac{\tau}{2}]$, which is centered at t and has a scale (or width) of τ. In the physical sciences, average values of signals are of wide-spread interest. Examples include one second averages of air temperature or of vertical velocity of air currents above a forest, hourly rainfall rates and monthly average temperatures in the Northern Hemisphere. What is often of more interest than the averages themselves, however, is how the averages evolve over time. One way to study this evolution is to look at the difference between adjacent averages. Accordingly, let us define

$$D(\tau,t) \equiv A(\tau,t+\tfrac{\tau}{2}) - A(\tau,t-\tfrac{\tau}{2}) = \frac{1}{\tau}\int_{t}^{t+\tau} x(u)\,du - \frac{1}{\tau}\int_{t-\tau}^{t} x(u)\,du.$$

For example, if we let $x(t)$ be the temperature of the Northern hemisphere at time t and if we let τ be 30.5 days (one month), then a plot of $D(\tau,t)$ versus t would tell us how the monthly averages before and after time t differ as a function of time. To connect this to the Haar wavelet, note that we can write

$$D(\tau,t) = \int_{-\infty}^{\infty} \tilde{\psi}_{\tau,t}(u)x(u)\,du, \text{ where } \tilde{\psi}_{\tau,t}(u) \equiv \begin{cases} -1/\tau, & t-\tau < u \le t; \\ 1/\tau, & t < u \le t+\tau; \\ 0, & \text{otherwise.} \end{cases}$$

If we specialize to the case $\tau = 1$ and $t = 0$ and then compare $\tilde{\psi}_{1,0}(u)$ to the Haar wavelet, we find that $\tilde{\psi}_{1,0}(u) = \sqrt{2}\psi^{(\mathrm{H})}(u)$. Thus, to within a constant of proportionality, the Haar wavelet tells us how unit scale averages differ before and after time zero.

We can easily adjust the Haar wavelet so that it can be used to tell us about changes in $x(t)$ at other scales and times. Accordingly let us consider

$$\psi^{(\mathrm{H})}_{\tau,t}(u) \equiv \tfrac{1}{\sqrt{\tau}}\psi^{(\mathrm{H})}\left(\tfrac{u-t}{\tau}\right) = \begin{cases} -\frac{1}{\sqrt{2\tau}}, & t-\tau < u \le t; \\ \frac{1}{\sqrt{2\tau}}, & t < u \le t+\tau; \\ 0, & \text{otherwise.} \end{cases}$$

Conceptually we form $\psi^{(\mathrm{H})}_{\tau,t}(u)$ from $\psi^{(\mathrm{H})}(u)$ by taking the latter and stretching it out so that its nonzero portion covers $[-\tau,\tau]$ and then relocating it so that it is centered at time t. It is easy to check that $\psi^{(\mathrm{H})}_{\tau,t}(u)$ obeys the defining properties for a wavelet in Equation (1) (in particular, we need $\sqrt{\tau}$ in the above to satisfy the second assumption).

We can now define the Haar continuous wavelet transform (CWT) of $x(t)$:

$$W^{(\mathrm{H})}(\tau,t) \equiv \int_{-\infty}^{\infty} \psi^{(\mathrm{H})}_{\tau,t}(u)x(u)\,du, \text{ where } 0 < \tau < \infty \text{ and } -\infty < t < \infty.$$

Note that $W^{(\mathrm{H})}(\tau, t) \propto D(\tau, t)$ so that the (t, τ)th value of this CWT can be interpreted as the differences between adjacent averages of scale τ located before and after time t. The CWT is fully equivalent to the signal $x(t)$ since we can recover $x(t)$ from its CWT:

$$x(t) = \frac{1}{C_\psi} \int_0^\infty \left[\int_{-\infty}^\infty W^{(\mathrm{H})}(\tau, u) \psi_{\tau,t}^{(\mathrm{H})}(u)\, du \right] \frac{d\tau}{\tau^2},$$

where C_ψ is constant depending just on $\psi^{(\mathrm{H})}(u)$. The above formula says that $x(t)$ can be rewritten as a linear combination of wavelets $\psi_{\tau,t}^{(\mathrm{H})}(u)$, all of which are centered around time t, and each one of which is associated with a particular scale between zero and infinity; i.e., the wavelet re-expression of $x(t)$ is both localized in time and scale-based. Note that, if $W^{(\mathrm{H})}(\tau, u)/\tau^2$ is large (small) in magnitude, then $\psi_{\tau,t}^{(\mathrm{H})}(u)$ is an important (insignificant) contributor to re-expressing $x(t)$ in terms of various wavelets. We also have the relationship

$$\int_{-\infty}^\infty x^2(t)\, dt = \frac{1}{C_\psi} \int_0^\infty \left[\int_{-\infty}^\infty [W^{(\mathrm{H})}(\tau, t)]^2\, dt \right] \frac{d\tau}{\tau^2}.$$

The left-hand side is called the ‘energy’ in the signal $x(t)$ (it is, however, not energy in the physical sense unless $x(t)$ has the proper units). We can thus interpret $[W^{(\mathrm{H})}(\tau, t)]^2/\tau^2$ as being proportional to an energy density function that decomposes the energy in $x(t)$ across different scales and times. Again, if $[W^{(\mathrm{H})}(\tau, t)]^2/\tau^2$ is large (small), we can say that there is an important (insignificant) contribution to the energy in $x(t)$ at scale τ and time t.

All of the above still holds if we replace the Haar wavelet with the Mexican hat wavelet (or any number of other wavelets defined in the literature). The physical interpretation of the Mexican hat CWT is quite similar to that of the Haar CWT and can be deduced by comparing the plots of $\psi^{(\mathrm{H})}(t)$ and $\psi^{\mathrm{Mh}}(t)$ in Figure 1: whereas the Haar wavelet compares simple averages before and after time zero, the Mexican hat wavelet compares a weighted average in a region centered about zero with weighted averages before and after that region. The interpretation of differences between weighted averages holds for many other commonly used wavelets (it is, however, problematic for the popular Morlet wavelet [14], which is complex-valued and is easier to interpret as yielding a highly localized Fourier transform).

THE DISCRETE WAVELET TRANSFORM

While the CWT has proven to be quite useful, there are a number of applications in which a discrete version of this transform (the DWT) is more appropriate for the following reasons. First, in contrast to the Fourier transform, the CWT does not give a succinct representation of a signal $x(t)$ because it changes a one dimensional signal into a two dimensional function. Full use of the CWT thus essentially leads us into image processing, which is considerably more involved computationally than just dealing with one dimensional signals. Second, the CWT for many time series of interest is highly redundant in both scale and time; i.e., there is little difference either between $W(\tau, t)$ and $W(\tau', t)$ when $|\tau - \tau'|$ is small compared to τ or between $W(\tau, t)$ and $W(\tau, t')$ when $|t - t'|$ is small compared to τ. Third, with the advent of modern digital computers, almost all signals are now collected digitally or subjected to a one-time 'analog to digital' conversion. The data with which scientists deal are thus discrete in nature, so it is necessary to discretize the CWT. How this discretization is accomplished leads to questions (e.g., how to handle boundary conditions) that are best addressed by directly dealing with a discrete transform. Fourth, as we argue below, the DWT that we describe here has considerable appeal in its own right because – in contrast to the CWT – it is an orthonormal transform that effectively decorrelates an important class of stochastic processes.

The DWT is defined in terms of a wavelet filter and an associated filter known as the scaling filter (these are sometimes called, respectively, the 'mother wavelet' and 'father wavelet' filters). Formally, a wavelet filter $h_{1,l}$ is a sequence that sums to zero, has unit energy and is orthogonal to its even shifts:

$$\text{(i)} \sum_{l=-\infty}^{\infty} h_{1,l} = 0 \text{ and (ii)} \sum_{l=-\infty}^{\infty} h_{1,l} h_{1,l+2n} = \begin{cases} 1, & \text{if } n = 0; \\ 0, & \text{otherwise} \end{cases} \qquad (2)$$

(while the assumptions on summation to zero and unit energy are analogous to those we made on $\psi(t)$ in Equation (1), the assumption about orthogonality to even shifts is new and is needed to obtain an orthonormal DWT). The most practical wavelet filters have a finite width L, by which we mean that $h_{1,l} = 0$ for $l < 0$ and $l \geq L$, while $h_{1,0} \neq 0$ and $h_{1,L-1} \neq 0$ (orthogonality to even shifts implies that L must be an even integer). The Haar wavelet filter is the simplest example of a wavelet filter:

$$h^{(\mathrm{H})}_{1,l} \equiv \begin{cases} 1/\sqrt{2}, & \text{if } l = 0; \\ -1/\sqrt{2}, & \text{if } l = 1; \\ 0, & \text{otherwise.} \end{cases}$$

Figure 2 goes about here.

This filter is plotted in the top row of Figure 2(a). Another wavelet filter is shown in the top row of Figure 2(c). This one is known as the Daubechies 'least asymmetric' wavelet of width $L = 8$, referred to henceforth as the LA(8) wavelet [7]. As the plot indicates, only three of the eight nonzero values in this filter are significantly different from zero.

The scaling filter $g_{1,l}$ that is associated with $h_{1,l}$ is constructed by reversing $h_{1,l}$, shifting it and then flipping the sign of every other variable: $g_{1,l} \equiv (-1)^{l+1} h_{1,L-1-l}$. For example, the Haar scaling filter is given by

$$g^{(\mathrm{H})}_{1,l} \equiv \begin{cases} -h^{(\mathrm{H})}_{1,1} = 1/\sqrt{2}, & \text{if } l = 0; \\ h^{(\mathrm{H})}_{1,0} = 1/\sqrt{2}, & \text{if } l = 1; \\ 0, & \text{otherwise.} \end{cases}$$

Plots of the Haar and LA(8) scaling filters are shown in the top rows of Figures 2(b) and (d).

Given $h_{1,l}$ and $g_{1,l}$, we can now define a DWT of unit level (the number of levels in a DWT corresponds to the number of scales that we are interested in). We do so first by using these filters to obtain

$$\widetilde{W}_{1,t} \equiv \frac{1}{\sqrt{2}} \sum_{l=0}^{L-1} h_{1,l} X_{t-l \bmod N} \text{ and } \widetilde{V}_{1,t} \equiv \frac{1}{\sqrt{2}} \sum_{l=0}^{L-1} g_{1,l} X_{t-l \bmod N}, \quad t = 0, \ldots, N-1, \tag{3}$$

where '$t - l \bmod N$' stands for '$t - l$ modulo N' and is defined as follows: if $0 \leq m \leq N - 1$, then $m \bmod N \equiv m$; if not, then $m \bmod N \equiv m + nN$, where nN is the unique integer multiple of N such that $0 \leq m + nN \leq N - 1$ (this 'circular' filtering effectively treats X_t as if it were periodic with a period of N, which – while admittedly problematic for many time series – is identical to how the discrete Fourier transform would treat X_t). Assuming for convenience that N is an even number, we then subsample the sequences $\widetilde{W}_{1,t}$ and $\widetilde{V}_{1,t}$ by taking every other value starting with $t = 1$ to obtain the wavelet and scaling coefficients of unit level (after a multiplication by $\sqrt{2}$):

$$W_{1,t} \equiv \sqrt{2}\widetilde{W}_{1,2t+1} \text{ and } V_{1,t} \equiv \sqrt{2}\widetilde{V}_{1,2t+1}, \quad t = 0, \ldots, N_1 - 1,$$

where $N_1 \equiv N/2$ (note that multiplication by $\sqrt{2}$ effectively cancels out the division by this factor in Equation (3) – we have defined $\widetilde{W}_{1,t}$ and $\widetilde{V}_{1,t}$ in this manner because they turn out to be the coefficients for a transform related to the DWT that we discuss at the end of this section). Together we have N coefficients in all, and these constitute the coefficients for a DWT of unit level. If we use the Haar wavelet, we have

$$W_{1,t} = \frac{X_{2t+1} - X_{2t}}{\sqrt{2}} \text{ and } V_{1,t} = \frac{X_{2t+1} + X_{2t}}{\sqrt{2}}. \tag{4}$$

Note that the wavelet (scaling) coefficients are proportional to differences (averages) of adjacent observations. If we use the LA(8) or other wavelets defined by Daubechies [7], we can make an analogous interpretation. Thus, while the Haar wavelet filter looks at differences between adjacent values when applied to a time series, the LA(8) filter yields essentially a contrast between X_t and values before and after X_t; likewise, whereas the Haar scaling filter produces two point averages, the LA(8) filter yields a weighted average whose effective width is two.

We can define DWTs for levels higher than unity by conceptually 'stretching out' the wavelet and scaling filters so that they are effectively twice as wide as we go from one level up to the next (the actual width of the jth level filters is given by $L_j \equiv [2^j - 1][L - 1] + 1$). The Haar and LA(8) filters $h_{j,l}$ and $g_{j,l}$ for levels $j = 2, 3$ and 4 are shown in Figure 2 below the corresponding plots for unit level. When applied to X_t, the filter $g_{j,l}$ yields a (weighted) average over a scale of $\lambda_j \equiv 2^j$, whereas the filter $h_{j,l}$ produces a differences of averages over a scale of $\tau_j \equiv \lambda_j/2 = 2^{j-1}$ (these are standardized scales – the corresponding physical scales are $\lambda_j \, \Delta t$ and $\tau_j \, \Delta t$). The $h_{j,l}$ filters for the Haar case look like subsamples from the Haar wavelet function, whereas the LA(8) filters have a shape that is reminiscent of the Mexican hat wavelet function.

To obtain the jth level wavelet and scaling coefficients, we first apply the jth level filters to X_t to obtain

$$\widetilde{W}_{j,t} \equiv \frac{1}{2^{j/2}} \sum_{l=0}^{L_j-1} h_{j,l} X_{t-l \bmod N} \text{ and } \widetilde{V}_{j,t} \equiv \frac{1}{2^{j/2}} \sum_{l=0}^{L_j-1} g_{j,l} X_{t-l \bmod N}, \quad t = 0, \ldots, N-1. \tag{5}$$

Assuming for convenience that N is divisible by 2^j, we then subsample and renormalize every 2^jth value, thus yielding the jth level wavelet and scaling coefficients:

$$W_{j,t} \equiv 2^{j/2} \widetilde{W}_{j,2^j(t+1)-1} \text{ and } V_{j,t} \equiv 2^{j/2} \widetilde{V}_{j,2^j(t+1)-1}, \quad t = 0, \ldots, N_j - 1,$$

where $N_j \equiv N/2^j$. Note that, as j increases, we subsample less often, and hence the number of coefficients decreases; i.e., at scale τ_j, we subsample every τ_jth value of $\widetilde{W}_{j,t}$ to form $W_{j,t}$, and there are $N/(2\tau_j)$ coefficients on level j. The Jth level DWT consists of the $J+1$ sequences $\{W_{j,t} : t = 0, \ldots, N_j - 1\}$, $j = 1, \ldots, J$, and $\{V_{J,t} : t = 0, \ldots, N_J - 1\}$.

Figure 3 goes about here.

Figure 3 shows a level $J = 7$ LA(8) DWT for the Northern Hemisphere temperature series X_t (using reflection boundary conditions – see Section 4.11 of [28] for details). The series itself is shown in the left-hand plot on the bottom row and consists of $N = 1664$ values in all (the sampling time between data values is $\Delta t = 1/12$th of a year). Above X_t we have plotted the $W_{j,t}$ and $V_{7,t}$ series that make up the DWT. The wavelet coefficients are plotted as deviations from zero. Each coefficient $W_{j,t}$ is plotted against the midpoint of the time interval spanning the X_t values that largely determine $W_{j,t}$. This time interval has width $2\tau_j\,\Delta t$ (since there are $N_j = N/(2\tau_j)$ points at level j, the product of the width and N_j is always $N\,\Delta t$, i.e., the total span of the time series). Due to the shape of the LA(8) wavelet, a large value for $W_{j,t}$ indicates a large difference between (i) a weighted average over an interval of width $\tau_j\,\Delta t$ centered at the time associated with $W_{j,t}$ and (ii) the sum of weighted averages before and after that interval (each of these averages spans an interval of approximate width $\tau_j\,\Delta t/2$). Note that, while the variability in the larger scale $W_{j,t}$ is homogeneous over time, there is an evident decrease in variability with increasing time at smaller scales (this is most likely due to an improvement in quality of the measurements over time). In a similar manner the scaling coefficients are shown in the top row (but now as a line plot) and represent weighted averages over a scale of $\lambda_7\,\Delta t \doteq 10.67$ years. Note that the overall upward pattern in X_t is captured in the scaling coefficients.

In practice, we actually do not need to explicitly deal with any of the filters except for $h_{1,l}$ and $g_{1,l}$. This is because we can use an elegant 'pyramid' algorithm to compute the DWT of level J. To do so, we define a set of zeroth level scaling coefficients as $V_{0,t} = X_t$. Given the scaling coefficients $V_{j-1,t}$ for level $j-1$, we can use $h_{1,l}$ and $g_{1,l}$ to obtain the level j wavelet and scaling coefficients:

$$W_{j,t} = \sum_{l=0}^{L-1} h_{1,l} V_{j-1,2t+1-l \bmod N_{j-1}} \text{ and } V_{j,t} = \sum_{l=0}^{L-1} g_{1,l} V_{j-1,2t+1-l \bmod N_{j-1}} \text{ for } t = 0, \ldots, N_j - 1.$$

Intuitively, the pyramid algorithm makes sense if we recall how $h_{1,l}$ and $g_{1,l}$ modify $V_{0,t} = X_t$. If we regard X_t as having a scale of unity, then use of $h_{1,l}$ yields wavelet coefficients that can be interpreted in terms of changes at unit scale, while $g_{1,l}$ yields scaling coefficients related to averages at a scale twice as large (cf. Equation (4)). Since $V_{j-1,t}$ has scale $\lambda_{j-1} = \tau_j$, use of $h_{1,l}$ now yields wavelet coefficients whose scale is τ_j, while $g_{1,l}$ yields scaling coefficients with scale $2\tau_j = \lambda_j$. If we consider a level J DWT when $N = 2^J$, then there is but a single scaling coefficient $V_{J,0}$, in which case $V_{J,0}$ is proportional to the sample mean of the time series and is thus associated with a scale of $\lambda_J = N$ (i.e., the entire width of the time series). The pyramid algorithm allows us to compute the DWT using $O(N)$ arithmetic operations (this is fewer than the celebrated fast Fourier transform algorithm, which requires $O(N \log_2(N))$ operations).

We can also express the Jth level DWT in terms of an orthonormal transform of the vector $\mathbf{X} \equiv [X_0, X_1, \ldots, X_{N-1}]'$. Let $\mathbf{W}_j \equiv [W_{j,0}, W_{j,1}, \ldots, W_{j,N_j-1}]'$ and $\mathbf{V}_J \equiv [V_{J,0}, V_{J,1}, \ldots, V_{J,N_J-1}]'$. Then we have the analysis equation $\mathbf{W} = \mathcal{W}\mathbf{X}$, where $\mathbf{W}$ contains the DWT coefficients, i.e.,

$$\mathbf{W} = \begin{bmatrix} \mathbf{W}_1 \\ \mathbf{W}_2 \\ \vdots \\ \mathbf{W}_J \\ \mathbf{V}_J \end{bmatrix}, \tag{6}$$

while $\mathcal{W}$ is an $N \times N$ matrix whose rows depend solely on the wavelet filter $h_{1,l}$. The properties of the wavelet filter (Equation (2)) imply that $\mathcal{W}$ is an orthonormal matrix; i.e., $\mathcal{W}'\mathcal{W} = I_N$, where I_N is the $N \times N$ identity matrix.

Orthonormality has two important consequences. First, an orthonormal transform preserves the 'energy' in $\mathbf{X}$ in the sense that $\|\mathbf{W}\|^2 = \|\mathbf{X}\|^2$, where $\|\mathbf{X}\|^2 \equiv \sum_{t=0}^{N-1} X_t^2$ is the squared norm of the vector $\mathbf{X}$. Using the partitioning of $\mathbf{W}$ given in Equation (6), we can write

$$\|\mathbf{X}\|^2 = \sum_{j=1}^{J} \|\mathbf{W}_j\|^2 + \|\mathbf{V}_J\|^2. \tag{7}$$

The above yields a scale-based decomposition of energy, in which $\|\mathbf{W}_j\|^2$ represents the contribution to the energy due to changes on scale τ_j. If we let $\overline{X} \equiv \frac{1}{N}\sum_{t=0}^{N-1} X_t$ be the sample mean of the time series, we can obtain a scale-based analysis of the sample variance (ANOVA):

$$\hat{\sigma}_X^2 \equiv \frac{1}{N}\sum_{t=0}^{N-1} (X_t - \overline{X})^2 = \frac{1}{N}\|\mathbf{X}\|^2 - \overline{X}^2 = \frac{1}{N}\sum_{j=1}^{J_0} \|\mathbf{W}_j\|^2 + \frac{1}{N}\|\mathbf{V}_{J_0}\|^2 - \overline{X}^2 \tag{8}$$

(multiplication of the last two terms by 2^{J_0} yields the sample variance for the scaling coefficients). Thus $\|\mathbf{W}_j\|^2/N$ is the contribution to the sample variance of X_t due to to changes on scale τ_j.

Figure 4 goes about here.

As an example, Figure 4 shows $\|\mathbf{W}_j\|^2/N$ versus $\tau_j\,\Delta t$ for $j = 1, \ldots, 7$ (the o's) and $\|\mathbf{V}_7\|^2/N - \overline{X}^2$ versus $\lambda_7\,\Delta t$ (the single x) based upon the LA(8) DWT for the Northern Hemisphere temperature data shown in Figure 3. The dominant contributor to the sample variance is due to variations in averages at the scale $\lambda_7\,\Delta t \doteq 10.67$ years. For scales $\tau_j < \lambda_7$, the largest contributor is the smallest scale, i.e., changes in weighted averages on a scale of $\tau_1\,\Delta t = \Delta t = 1/12$th of a year. After τ_j, the contributions to the sample variance decrease in a roughly linear manner on a log/log plot (as discussed below, this is indicative of a process possessing 'long memory' in the sense that the correlation between observations that are separated by k units does not decrease rapidly as k increases).

A second consequence of orthonormality is that the inverse of the DWT matrix $\mathcal{W}$ is just its transpose $\mathcal{W}'$, so we can recover $\mathbf{X}$ from its DWT coefficients via the synthesis equation $\mathbf{X} = \mathcal{W}'\mathbf{W}$. Thus $\mathbf{X}$ and $\mathbf{W}$ are fully equivalent and can be regarded as two representations for the same mathematical entity. We can put this synthesis equation to good use by partitioning $\mathcal{W}$ commensurate with the partitioning of $\mathbf{W}$ into the $\mathbf{W}_j$ and $\mathbf{V}_J$ vectors:

$$\mathcal{W} = \begin{bmatrix} \mathcal{W}_1 \\ \mathcal{W}_2 \\ \vdots \\ \mathcal{W}_J \\ \mathcal{V}_J \end{bmatrix},$$

where $\mathcal{W}_j$ is an $N_j \times N$ matrix whose rows are constructed from the filter $h_{j,l}$, while $\mathcal{V}_J$ is an $N_J \times N$ matrix constructed from the filter $g_{J,l}$. We can then write

$$\mathbf{X} = \mathcal{W}'\mathbf{W} = \sum_{j=1}^{J} \mathcal{W}_j'\mathbf{W}_j + \mathcal{V}_J'\mathbf{V}_J \equiv \sum_{j=1}^{J} \mathcal{D}_j + \mathcal{S}_J, \tag{9}$$

where $\mathcal{D}_j \equiv \mathcal{W}_j'\mathbf{W}_j$ is an N dimensional vector called the jth level detail, while $\mathcal{S}_J \equiv \mathcal{V}_J'\mathbf{V}_J$ is called the Jth level smooth. Because the rows of $\mathcal{W}_j$ are based on the filter used to create $\mathbf{W}_j$, we can associate the vector $\mathcal{D}_j$ with changes in $\mathbf{X}$ on the scale τ_j; likewise, $\mathcal{S}_J$ is related to averages on a scale λ_J. The above additive decomposition of $\mathbf{X}$ into details and a smooth is known as a multiresolution analysis (MRA). It is interesting to note that $\|\mathcal{D}_j\|^2 = \|\mathbf{W}_j\|^2$ and $\|\mathcal{S}_J\|^2 = \|\mathbf{V}_J\|^2$ so that we can re-express the ANOVA of Equation (8) in terms of the components of the MRA.

Figure 5 goes about here.

As an example, Figure 5 shows the MRA for the Northern Hemisphere temperature data $\mathbf{X}$ corresponding to the DWT depicted in Figure 3. The bottom plot shows $\mathbf{X}$, above which are depicted (from bottom to top) the details $\mathcal{D}_j$, $j = 1, \ldots, 7$, and the smooth $\mathcal{S}_7$. While $\mathcal{S}_7$ tracks the large scale variations in the series, the details give us an indication of how the series varies over time at various scales. For example, the details for $j = 1, 2$ and 3 show higher variability at the early part of the series (roughly up to the 1880s), after which they appear to be fairly homogeneous. The details for $j = 4, \ldots, 7$ are homogeneous throughout, although the eye is drawn to some anomalous features (e.g., the bulge in $\mathcal{D}_7$ in the 1890s).

We can roughly regard the DWT as a scheme for sampling from some CWT, say, $W(\tau, t)$. In particular we can relate $W_{j,t}$ to $W(\tau_j\,\Delta t, 2t\tau_j\,\Delta t)$. In this scheme we make use of just the dyadic scales τ_j, and we sample across time from $W(\tau_j\,\Delta t, t)$ less often as τ_j increases. Coupled with the scaling coefficients $V_{J,t}$ (which can be considered to be a summary of the CWT at all scales $\tau \geq \tau_J\,\Delta t$), this yields a DWT with the same number of values as the original time series, which is a drastic reduction from the potential number of values in the CWT; nonetheless, we have not lost any information in the sense that we can recover $\mathbf{X}$ from $\mathbf{W}$ and vice versa.

While limiting ourselves to the dyadic scales is often quite acceptable, the subsampling within each scale can lead to certain undesirable 'alignment' effects (i.e., the exact time at which we start recording a time series can materially influence its DWT at all time points thereafter). One manifestation of this effect is the fact that a circular shift in a time series can yield a substantially different MRA. Thus, if $\mathcal{T}$ is the $N \times N$ matrix such that $\mathcal{T}\mathbf{X} = [X_{N-1}, X_0, X_1, \ldots, X_{N-2}]'$ and if we let $\mathcal{D}_j^{(0)}$ and $\mathcal{D}_j^{(1)}$ be the details in the MRAs for, respectively, $\mathbf{X}$ and $\mathcal{T}\mathbf{X}$, then in general $\mathcal{D}_j^{(0)} \neq \mathcal{T}^{-1}\mathcal{D}_j^{(1)}$ (in fact $\mathcal{D}_j^{(0)}$ and $\mathcal{T}^{-1}\mathcal{D}_j^{(1)}$ can differ at every element). These alignment effects can be alleviated if we eliminate the subsampling in the DWT, leading to a definition for a 'nondecimated' or 'shift invariant' DWT known as the maximal overlap DWT (MODWT). The

MODWT of level J consists of $J+1$ vectors, each of length N, namely, $\widetilde{\mathbf{W}}_j \equiv [\widetilde{W}_{j,0}, \ldots, \widetilde{W}_{j,N-1}]'$, $j = 1, \ldots, J$, and $\widetilde{\mathbf{V}}_J \equiv [\widetilde{V}_{J,0}, \ldots, \widetilde{V}_{J,N-1}]'$, where $\widetilde{W}_{j,t}$ and $\widetilde{V}_{J,t}$ are defined in Equation (5). While the MODWT is certainly not an orthonormal transform, we can create ANOVAs and MRAs from it; i.e., in analogy to Equations (7) and (9), we have

$$\|\mathbf{X}\|^2 = \sum_{j=1}^{J} \|\widetilde{\mathbf{W}}_j\|^2 + \|\widetilde{\mathbf{V}}_J\|^2 \text{ and } \mathbf{X} = \sum_{j=1}^{J} \widetilde{\mathcal{W}}_j' \widetilde{\mathbf{W}}_j + \widetilde{\mathcal{V}}_J' \widetilde{\mathbf{V}}_J \equiv \sum_{j=1}^{J} \widetilde{\mathcal{D}}_j + \widetilde{\mathcal{S}}_J, \tag{10}$$

where $\widetilde{\mathcal{W}}_j$ is the $N \times N$ matrix such that $\widetilde{\mathbf{W}}_j = \widetilde{\mathcal{W}}_j \mathbf{X}$. In contrast to the DWT, if $\widetilde{\mathbf{W}}_j^{(0)}$ and $\widetilde{\mathcal{D}}_j^{(0)}$ are the jth level MODWT coefficients and detail for $\mathbf{X}$ and if $\widetilde{\mathbf{W}}_j^{(1)}$ and $\widetilde{\mathcal{D}}_j^{(1)}$ are the corresponding quantities for $\mathcal{T}\mathbf{X}$, we have the appealing properties $\mathcal{T}^{-1}\widetilde{\mathbf{W}}_j^{(1)} = \widetilde{\mathbf{W}}_j^{(0)}$ and $\mathcal{T}^{-1}\widetilde{\mathcal{D}}_j^{(1)} = \widetilde{\mathcal{D}}_j^{(0)}$. The MODWT has the advantage of a natural definition for all sample sizes N whereas the Jth level DWT is only naturally defined for an N that is a multiple of 2^J (there are various schemes for getting around this limitation). There is also an efficient pyramid algorithm for the MODWT with a computational burden of $O(N \log_2(N))$ arithmetic operations (while this is admittedly greater than the $O(N)$ burden for the DWT, it is in fact the same as that of the fast Fourier transform algorithm).

USES FOR THE DISCRETE WAVELET TRANSFORM

While the ANOVAs and MRAs based upon Equations (7), (9) and (10) can lead to useful qualitative descriptions for a time series, there is much more that we can do when we combine the DWT or MODWT with statistical models (to do so, we now regard $\mathbf{X}$ as a vector of RVs). Here we consider briefly three important uses for the DWT and MODWT, all of which we can describe as departures from the basic ANOVA or MRA decompositions.

Wavelet Shrinkage

Suppose that we entertain a 'signal plus noise' model for our time series; i.e., we write $\mathbf{X} = \mathbf{S} + \mathbf{n}$, where $\mathbf{S}$ is an unknown signal of interest that we are prevented from observing directly due to the addition of unwanted noise $\mathbf{n}$. If we use $\mathcal{W}$ to form the DWT of both sides of this model, we can write $\mathbf{W} = \mathcal{W}\mathbf{X} = \mathbf{W}^{(\mathrm{S})} + \mathbf{W}^{(\mathrm{n})}$, where $\mathbf{W}^{(\mathrm{S})} \equiv \mathcal{W}\mathbf{S}$ and $\mathbf{W}^{(\mathrm{n})} \equiv \mathcal{W}\mathbf{n}$. Due to the time/scale nature of the DWT, we can argue that, for a wide variety of signals, the DWT $\mathbf{W}^{(\mathrm{S})}$ is a more compact representation of $\mathbf{S}$ than $\mathbf{S}$ itself; i.e., whereas a typical signal can have many large values spread out over time, its DWT will tend to have a few large coefficients and many small coefficients. On the other hand, noise typically is spread out homogeneously across time, so its DWT will be homogeneous across different scales and times. If $\|\mathbf{S}\|^2 \gg \|\mathbf{n}\|^2$ and hence $\|\mathbf{W}^{(\mathrm{S})}\|^2 \gg \|\mathbf{W}^{(\mathrm{n})}\|^2$ also, then $\mathbf{W}$ should consist of (i) a few large coefficients that are largely attributable to the signal and (ii) many small coefficients that are a combination of $\mathbf{W}^{(\mathrm{n})}$ and relatively unimportant coefficients in $\mathbf{W}^{(\mathrm{S})}$.

These considerations lead to the following scheme for estimating $\mathbf{S}$ based upon $\mathbf{X}$. We take a Jth level DWT of $\mathbf{X}$ and then consider modifying each wavelet vector $\mathbf{W}_j$ to form a new vector, say, $\mathbf{W}_j^{(\mathrm{t})}$. We do the modification by examining the magnitude of each wavelet coefficient. If a given coefficient is large (i.e., greater in magnitude than, say, $\delta^{(\mathrm{t})}$), we regard it as capturing an important part of the signal and put it into $\mathbf{W}_j^{(\mathrm{t})}$ unaltered; on the other hand, we consider each small coefficient as being due to the noise, so we put a zero into $\mathbf{W}_j^{(\mathrm{t})}$ in its place. When we are done, we estimate $\mathbf{S}$ via

$$\widehat{\mathbf{S}} \equiv \sum_{j=1}^{J} \mathcal{W}_j' \mathbf{W}_j^{(\mathrm{t})} + \mathcal{V}_J' \mathbf{V}_J;$$

i.e., we take the inverse DWT of the coefficient vectors $\mathbf{W}_j^{(\mathrm{t})}$ and $\mathbf{V}_J$. Note that, because the scaling coefficients are typically large scale weighted averages of the time series and because such averages usually reflect the large scale characteristics of a signal, we administratively assign all the scaling coefficients to the signal. Note also that we can regard the construction of the estimated signal as a modification of the basic MRA scheme of Equation (9) in which we create a denoised detail for each j.

If we try to implement this scheme in practice, we are immediately faced with a plethora of practical questions. We can address these questions in a statistically sensible manner if we can make some additional assumptions about the statistical properties of the noise (we also need to make assumptions about the signal, e.g., deterministic with certain smoothness properties or stochastic with a certain covariance structure). The various assumptions that investigators have been willing to entertain regarding the nature of the signal and the noise have led to a rather large literature on what is most commonly referred to as 'wavelet shrinkage' [8, 9, 10, 33] (this name stems from the fact that, if we want to estimate $\mathbf{S}$ optimally under certain commonly used assumptions, then we must abandon the simple 'keep/zero' strategy we just described in favor of one that shrinks wavelet coefficients toward zero). For example, if we are willing to assume that $\mathbf{n}$ is Gaussian white noise with variance σ_n^2, then $\mathbf{W}^{(\mathrm{n})}$ is also such, and asymptotic theory suggests comparing the magnitude of the wavelet coefficients to the so-called 'universal' threshold of $\delta^{(\mathrm{u})} \equiv \sqrt{[2\sigma_n^2 \log(N)]}$ (the theory also assumes $\mathbf{S}$ to be deterministic and sampled from certain quite general classes of functions). In practice, we can estimate σ_n using a robust estimate of variance based upon the median absolute deviation (MAD) of the smallest scale wavelet coefficients $\mathbf{W}_1$ about their median [16, 21] (a robust method is required to allow for the possibility of a few large signal coefficients influencing the coefficients in $\mathbf{W}_1$).

Figure 6 goes about here.

Figure 6 shows a very simple example of wavelet-based signal estimation. Here we assume that the Northern Hemisphere series (the dots) can be decomposed as a trend component (the signal) plus additive noise. We assume that the trend is inherently smooth, and we propose to estimate it based upon the DWT displayed in Figure 2. To ensure that the trend estimate is sufficiently smooth, we forcibly set all the wavelet coefficients on the three smallest scales to zero; i.e., we take the elements of $\mathbf{W}_j^{(\mathrm{t})}$, $j = 1, 2$ and 3, to be zero (this amounts to limiting the definition of the trend here to variations over scales of eight months or higher). We also assume that the noise variances for the elements of $\mathbf{W}_j$, $j = 4, \ldots, 7$, are all the same, and we estimate this variance based upon a MAD scale estimate. We then compute the universal threshold and use it to form the elements of $\mathbf{W}_j^{(\mathrm{t})}$, $j = 4, \ldots, 7$. The resulting estimate of trend is the solid curve in Figure 6. Note that this estimate is markedly different from traditional smoothers with a fixed bandwidth (e.g., a running average with a bandwidth of, say, one year): while the estimated trend is quite smooth from 1900 and on, it is less so prior to 1900 because the wavelet scheme has captured two intermediate scale variations (one near 1870, and the other in the 1890s). This simple example hints at the power of the wavelet-based approach in estimating signals with time-varying smoothness properties.

Wavelet Variance

If we take X_t to be a stochastic process, then the wavelet coefficients $W_{j,t}$ define a new stochastic process that reflects variations in X_t on scale τ_j. Assuming that it exists, the wavelet variance is by definition just the variance of $W_{j,t}$, namely, $\nu_X^2(\tau_j) \equiv \mathrm{var}\,\{W_{j,t}\}$ (theory and applications of the wavelet variance are discussed in, e.g., [3, 5, 11, 12, 13, 18, 19, 20, 27, 28,29, 32]). If X_t is a stationary process with variance σ_X^2, then we have the fundamental relationship

$$\sigma_X^2 = \sum_{j=1}^{\infty} \nu_X^2(\tau_j),$$

which is analogous to the ANOVAs that can be formed from Equations (8) and (10). This scaled-based decomposition of σ_X^2 is also analogous to the frequency-based decomposition given by the power spectral density function $S_X(f)$, which has an integral equal to σ_X^2 (indeed, $\nu_X^2(\tau_j)$ is related to $S_X(f)$ over the frequency interval $[1/\tau_{j+2}, 1/\tau_{j+1}]$). The advantages of the wavelet variance over $S_X(f)$ are that (i) it is scale-based and hence of interest to physical scientists who naturally regard phenomena as having different scales of variation (note that $\nu_X(\tau_j)$ has the same units as X_t itself, which makes it easier to interpret than $S_X(f)$); (ii) it offers a more succinct decomposition for many processes routinely encountered in the physical sciences (iii) it has a simpler estimation theory; and (iv) it extends readily to certain nonstationary processes, including those having stationary increments (i.e., $Y_t \equiv X_t - X_{t-1}$ is a stationary process even though X_t is not). Given a time series $\mathbf{X}$ that is either stationary or has stationary increments, we can readily estimate $\nu_X^2(\tau_j)$ based upon properly normalized averages of the squares of the DWT coefficients $\mathbf{W}_j$ or MODWT coefficients $\widetilde{\mathbf{W}}_j$.

If we revisit Figure 4 and if we assume that the Northern Hemisphere data $\mathbf{X}$ is a realization of a process with stationary increments, we can now regard the circles as DWT-based estimates of the wavelet variances $\nu_X^2(\tau_j)$ for $j = 1, \ldots, 7$. Asymptotic theory allows us to assess the quality of these estimates by forming approximate 95% confidence intervals for the unknown $\nu_X^2(\tau_j)$. Examples of such intervals are shown in the figure. (In practice we would prefer to use MODWT-based estimates because asymptotic theory says that these should be more efficient than the corresponding DWT-based estimates).

One consequence of the assumption that $\mathbf{X}$ has stationary increments is that each $\mathbf{W}_j$ should be a realization of a stationary process (here we are ignoring coefficients influenced by the periodicity assumption). If we look back at Figure 3, there is good reason to question the validity of the stationarity assumption for, say, $\mathbf{W}_1$ since these coefficients appear to decrease in variability with increasing time. We should thus really consider $\operatorname{var}\{W_{1,t}\}$ to be a function of time. If we can assume that var $\{W_{1,t}\}$ is slowly varying across time, we can readily adapt the wavelet variance estimator developed under the assumption of stationarity: we just need to assume that $\mathbf{X}$ can be regarded as having stationary increments within certain blocks of time, but we allow the nature of the increment process to vary from one block to the next. We then compute a separate estimate of $\nu_X^2(\tau_1)$ for each block based upon just those $W_{1,t}$ with times $2t\tau_j\,\Delta t$ occurring within that block.

Figure 7 shows time-varying wavelet variance estimates (the circles) for scale τ_1 based upon the DWT for the Northern Hemisphere data shown in Figure 3. Here we assume stationarity over a block spanning a bit more than a decade, based upon which we can compute 95% confidence intervals for the unknown wavelet variance $\nu_X^2(\tau_1)$ in each block. These intervals verify what our eyes have picked out, namely, that there is indeed significantly higher variability early on in $W_{1,t}$ (this variability has been markedly stable since about 1950).

Analysis of Long Memory Processes

By definition we say that X_t is a long memory process if its spectral density function $S_X(f)$ is approximately equal to $C_S|f|^\alpha$ at low frequencies for some $\alpha < 0$ and $C_S > 0$ (see Beran [2] for a precise definition). This implies that $S_X(f) \to \infty$ as $f \to 0$, which means that realizations of X_t have prominent low frequency (large scale) fluctuations. When $-1 < \alpha < 0$, a long memory process is stationary, and its autocorrelation sequence (ACS) $\rho_\tau \equiv \text{corr}\,\{X_t, X_{t+\tau}\}$ is approximately equal to $C_\rho \tau^\beta$, where $\beta = -\alpha - 1$ and $C_\rho > 0$. While ρ_τ does decrease to zero with increasing τ, its rate of decay is much slower than that of more traditional models for time series (i.e., autoregressive processes, moving average processes, and combinations thereof). The slow rate of decay implies that X_t retains some 'memory' of its distant past. Investigators have found that long memory processes are effective models for a wide variety of time series, ranging from the microscopic (voltage fluctuations across cell membranes [17]) to the cosmic (time variability of X-ray from galaxies [25]).

The DWT is a remarkably effective tool for analyzing long memory processes [1, 11, 23, 24, 28, 34, 35, 36] (we require only a very mild condition on the width L of the wavelet filter to be able to handle α's of various sizes). To see the basic reason this is true, let us return to Figure 3 and examine the right-hand column of plots, which show, from bottom to top, the sample ACSs (plotted as deviations from zero) for the Northern Hemisphere data X_t and for its wavelet coefficients $W_{j,t}$. The sample ACS for X_t is slowly decaying and is reminiscent of ACSs for long memory processes.

By contrast, the sample ACSs for the wavelet coefficients are all close to zero. The two curves in each ACS plot delineate how much departure from zero we would expect to see 95% of the time if each $W_{j,t}$ were a realization of a white noise sequence. This figure is an illustration of the fact that, to a good approximation, the DWT decorrelates a long memory process. This property can be put to good use, e.g., in constructing approximate maximum likelihood estimators for the parameters of a long memory process (even if the time series is corrupted by the presence of a polynomial trend) or in bootstrapping certain common statistics for time series.

In addition, we can use the wavelet variance to obtain a preliminary indication of the presence of long memory in a time series. This result is based on the fact that $\nu_X^2(\tau_j)$ is approximately proportional to $\tau_j^{-\alpha-1}$ for a long memory process. Thus, if we plot estimate of $\nu_X^2(\tau_j)$ versus τ_j on a log/log scale, we should see a linear variation with a slope of $-\alpha - 1$ if X_t follows a long memory model.

As an example, let us look again at the wavelet variance estimates for the Northern Hemisphere data shown in Figure 4. Roughly speaking (and ignoring the time-varying nature of the wavelet coefficients at small scales), we see a linear decay (on the log/log scale) in the estimated wavelet variances with increasing scale. If we fit a least square line through these estimated values in log/log space, we obtain an estimated slope of -0.37. This translates into an estimate of $\alpha = -0.63$, which suggests that X_t might be well-modeled by a stationary long memory process [6].

REFERENCES

[1] Abry, P., Gonçalvès, P. and Flandrin, P. (1995). Wavelets, Spectrum Analysis and $1/f$ Processes. In *Wavelets and Statistics* (Lecture Notes in Statistics, Volume 103), A. Antoniadis and G. Oppenheim, eds. Springer–Verlag, New York, pp. 15–29.

[2] Beran, J. (1994). *Statistics for Long-Memory Processes.* Chapman & Hall, New York.

[3] Bradshaw, G. A. and Spies, T. A. (1992). Characterizing Canopy Gap Structure in Forests using Wavelet Analysis. *Journal of Ecology* **80**, 205–215.

[4] Bruce, A. G. and Gao, H.–Y. (1996). *Applied Wavelet Analysis with S-PLUS.* Springer, New York.

[5] Carmona, R., Hwang, W.–L. and Torrésani, B. (1998). *Practical Time-Frequency Analysis.* Academic Press, San Diego.

[6] Craigmile, P. F. (2000). Wavelet-Based Estimation for Trend Contaminated Long Memory Processes. Ph.D. dissertation, Department of Statistics, University of Washington, Seattle.

[7] Daubechies, I. (1988). *Ten Lectures on Wavelets.* SIAM, Philadelphia.

[8] Donoho, D. L. and Johnstone, I. M. (1994). Ideal Spatial Adaptation by Wavelet Shrinkage. *Biometrika* **81**, 425–455.

[9] Donoho, D. L. and Johnstone, I. M. (1995). Adapting to Unknown Smoothness via Wavelet Shrinkage. *Journal of the American Statistical Association* **90**, 1200–1224.

[10] Donoho, D. L., Johnstone, I. M., Kerkyacharian, G. and Picard, D. (1995). Wavelet Shrinkage: Asymptopia? *Journal of the Royal Statistical Society, Series B* **57**, 301–369.

[11] Flandrin, P. (1992). Wavelet Analysis and Synthesis of Fractional Brownian Motion. *IEEE Transactions on Information Theory* **38**, 910–917.

[12] Gamage, N. K. K. (1990). Detection of Coherent Structures in Shear Induced Turbulence using Wavelet Transform Methods. In *Ninth Symposium on Turbulence and Diffusion.* American Meteorological Society, Boston, pp. 389–392.

[13] Gao, W. and Li, B. L. (1993). Wavelet Analysis of Coherent Structures at the Atmosphere-Forest Interface. *Journal of Applied Meteorology* **32**, 1717–1725.

[14] Goupillaud, P., Grossmann, A. and Morlet, J. (1984). Cycle-Octave and Related Transforms in Seismic Signal Analysis. *Geoexploration* **23**, 85–102.

[15] Haar, A. (1910). Zur Theorie der Orthogonalen Funktionensysteme. *Mathematische Annalen* **69**, 331–371.

[16] Hampel, F. R. (1974). The Influence Curve and its Role in Robust Estimation. *Journal of the American Statistical Association* **69**, 383–393.

[17] Holden, A. V. (1976). *Models of the Stochastic Activity of Neurones* (Lecture Notes in Biomathematics, Volume 12). Springer–Verlag, Berlin.

[18] Hudgins, L. H., Friehe, C. A. and Mayer, M. E. (1993). Wavelet Transforms and Atmospheric Turbulence. *Physical Review Letters* **71**, 3279–3282.

[19] Kumar, P. and Foufoula–Georgiou, E. (1993). A Multicomponent Decomposition of Spatial Random Fields 1: Segregation of Large- and Small-Scale Features Using Wavelet Transforms. *Water Resources Research* **29**, 2515–2532.

[20] Kumar, P. and Foufoula–Georgiou, E. (1997). Wavelet Analysis for Geophysical Applications. *Reviews of Geophysics* **35**, 385–412.

[21] Launer, R. L. and Wilkinson, G. N. (1979). *Robustness in Statistics.* Academic Press, New York.

[22] Mallat, S. G. (1999). *A Wavelet Tour of Signal Processing* (Second Edition). Academic Press, San Diego.

[23] Masry, E. (1993). The Wavelet Transform of Stochastic Processes with Stationary Increments and its Application to Fractional Brownian Motion. *IEEE Transactions on Information Theory* **39**, 260–264.

[24] McCoy, E. J. and Walden, A. T. (1996). Wavelet Analysis and Synthesis of Stationary Long-Memory Processes. *Journal of Computational and Graphical Statistics* **5**, 26–56.

[25] McHardy, I. and Czerny, B. (1987). Fractal X-Ray Time Variability and Spectral Invariance of the Seyfert Galaxy NGC5506. *Nature* **325**, 696–698.

[26] Ogden, R. T. (1997). *Essential Wavelets for Statistical Applications and Data Analysis.* Birkhäuser, Boston.

[27] Percival, D. B. (1995). On Estimation of the Wavelet Variance. *Biometrika* **82**, 619–631.

[28] Percival, D. B. and Walden, A. T. (2000). *Wavelet Methods for Time Series Analysis.* Cambridge University Press, Cambridge, UK.

[29] Scargle, J. D. (1997). Wavelet Methods in Astronomical Time Series Analysis. In *Applications of Time Series Analysis in Astronomy and Meteorology*, T. Subba Rao, M. B. Priestley and O. Lessi, eds. Chapman & Hall, London, pp. 226–248.

[30] Smith, R. L. (1993). Long-range Dependence and Global Warming. In *Statistics for the Environment*, V. Barnett and K. F. Turkman, eds. Wiley, Chichester, West Sussex, England, pp. 141–161.

[31] Stollnitz, E. J., DeRose, T. D. and Salesin, D. H. (1996). *Wavelets for Computer Graphics: Theory and Applications.* Morgan Kaufmann, San Francisco.

[32] Torrence, C. and Compo, G. P. (1998). A Practical Guide to Wavelet Analysis. *Bulletin of the American Meteorological Society* **79**, 61–78.

[33] Vidakovic, B. (1999). *Statistical Modeling by Wavelets.* John Wiley & Sons, New York.

[34] Wornell, G. W. (1990). A Karhunen–Loève-like Expansion for $1/f$ Processes via Wavelets. *IEEE Transactions on Information Theory* **36**, 859–861.

[35] Wornell, G. W. (1993). Wavelet-Based Representations for the $1/f$ Family of Fractal Processes. *Proceedings of the IEEE* **81**, 1428–1450.

[36] Wornell, G. W. (1996). *Signal Processing with Fractals: A Wavelet-Based Approach.* Prentice Hall, Upper Saddle River, New Jersey.

Legends for Figures

Figure 1: Two wavelet functions. The left- and right-hand plots show, respectively, the Haar and Mexican hat wavelet functions.

Figure 2: Two sets of wavelet and scaling filters. The top row in plot (a) shows the Haar wavelet filter, below which are shown the associated filters $h_{j,l}$ for scales τ_j, $j = 2, 3$ and 4. Plot (b) is like (a), but now for the Haar scaling filter. Plots (c) and (d) are like (a) and (b), but now for the Daubechies least asymmetric (LA) filter with eight nonzero coefficients (some of these coefficients are quite close to zero, which is why there appear to be fewer than eight coefficients in the top rows of (c) and (d)).

Figure 3: Discrete wavelet transform of level $J = 7$ for a time series X_t of deseasonalized monthly average Northern Hemisphere temperature anomalies. This time series was obtained from the Climatic Research Unit, University of East Anglia (`http://www.cru.uea.ac.uk/cru/cru.htm`) and deseasonalized using harmonic regression [6]. The series is shown in the bottom left-hand plot and consists of monthly values from January 1856 to August 1994. Above this series are shown the LA(8) wavelet coefficients $W_{j,t}$ for scales $\tau_j\,\Delta t = 2^{j-1}/12$ years, $j = 1, \ldots, 7$ (from bottom to top, plotted as deviations from zero). The top left-hand plot shows the corresponding scaling coefficients $V_{7,t}$, which are associated with scale $\lambda_7\,\Delta t \doteq 10.67$ years. The plots to the right of X_t and $W_{j,t}$ are of the sample autocorrelation sequences for the corresponding series (plotted as deviations from zero). In the case of each $W_{j,t}$, the upper and lower curves depict a 95% confidence interval for a given autocorrelation under the assumption that the wavelet coefficients are uncorrelated.

Figure 4: Wavelet-based analysis of the sample variance of the Northern Hemisphere temperature time series. The `o`'s depict the contribution to the sample variance due to variations on a scale of $\tau_j\,\Delta t = 2^{j-1}/12$ years, $j = 1, \ldots, 7$, while the single `x` is the contribution due to averages on a scale of $\lambda_7\,\Delta t \doteq 10.67$ years. The sum of the values depicted by the `o`'s and the `x` is equal to the sample variance. The vertical lines emanating each `o` depict 95% confidence intervals for a hypothesized true wavelet variance $\nu_X^2(\tau_j)$ (this analysis is based on the DWT shown in Figure 3).

Figure 5: Multiresolution analysis of the Northern Hemisphere temperature time series. The series itself is plotted on the bottom row, above which are displayed the details $\mathcal{D}_j$ and smooth $\mathcal{S}_7$ calculated from the DWT shown in Figure 3.

Figure 6: Wavelet-based denoising of the Northern Hemisphere temperature time series. The points in this plot show the time series itself, while the solid curve is a wavelet-based estimate of a hypothesized smooth trend in the data (this estimate is based on the DWT shown in Figure 3).

Figure 7: Evolution of wavelet variance across time at a scale of $\tau_1 \, \Delta t = \frac{1}{12}$th year for Northern Hemisphere temperature time series. Each wavelet variance estimate (the o's) is computed using the $W_{1,t}$ coefficients associated with times in nonoverlapping blocks spanning $10\frac{2}{3}$ years. The vertical lines emanating each o depict 95% confidence intervals for a hypothesized true wavelet variance $\nu_X^2(\tau_j)$ for a given block (this analysis is based on the DWT shown in Figure 3).

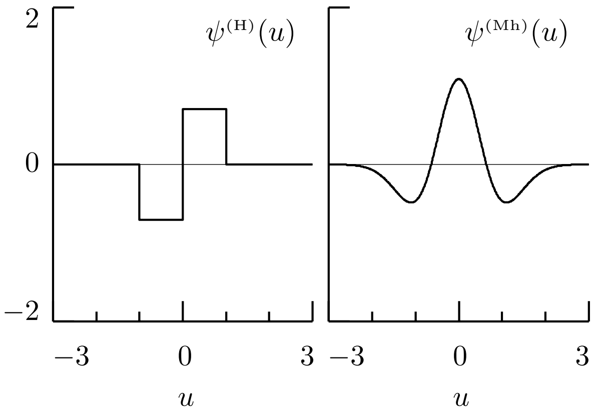


Figure 1

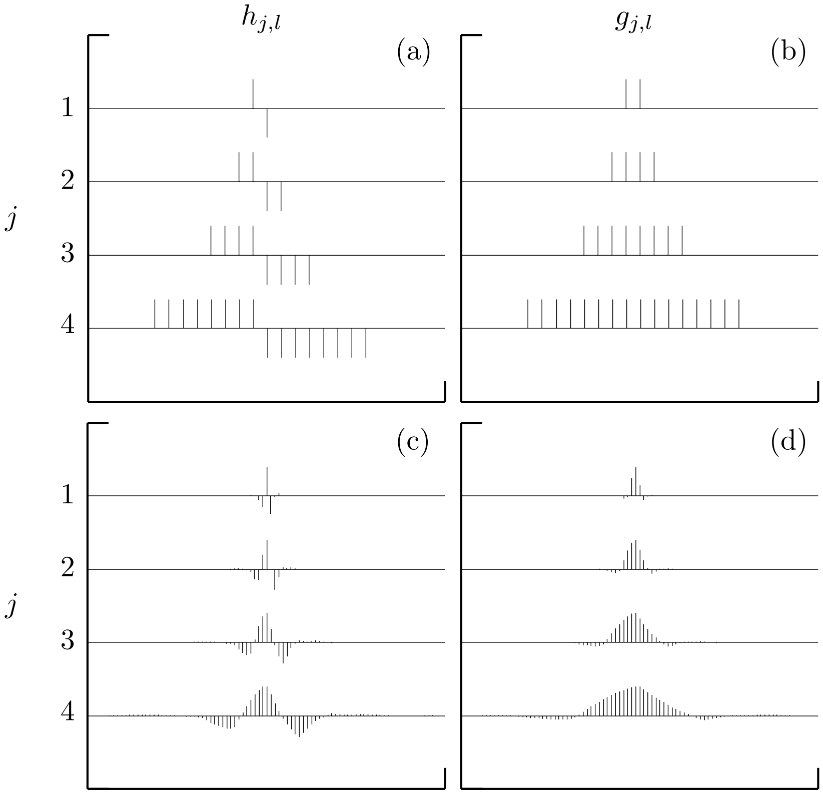


Figure 2

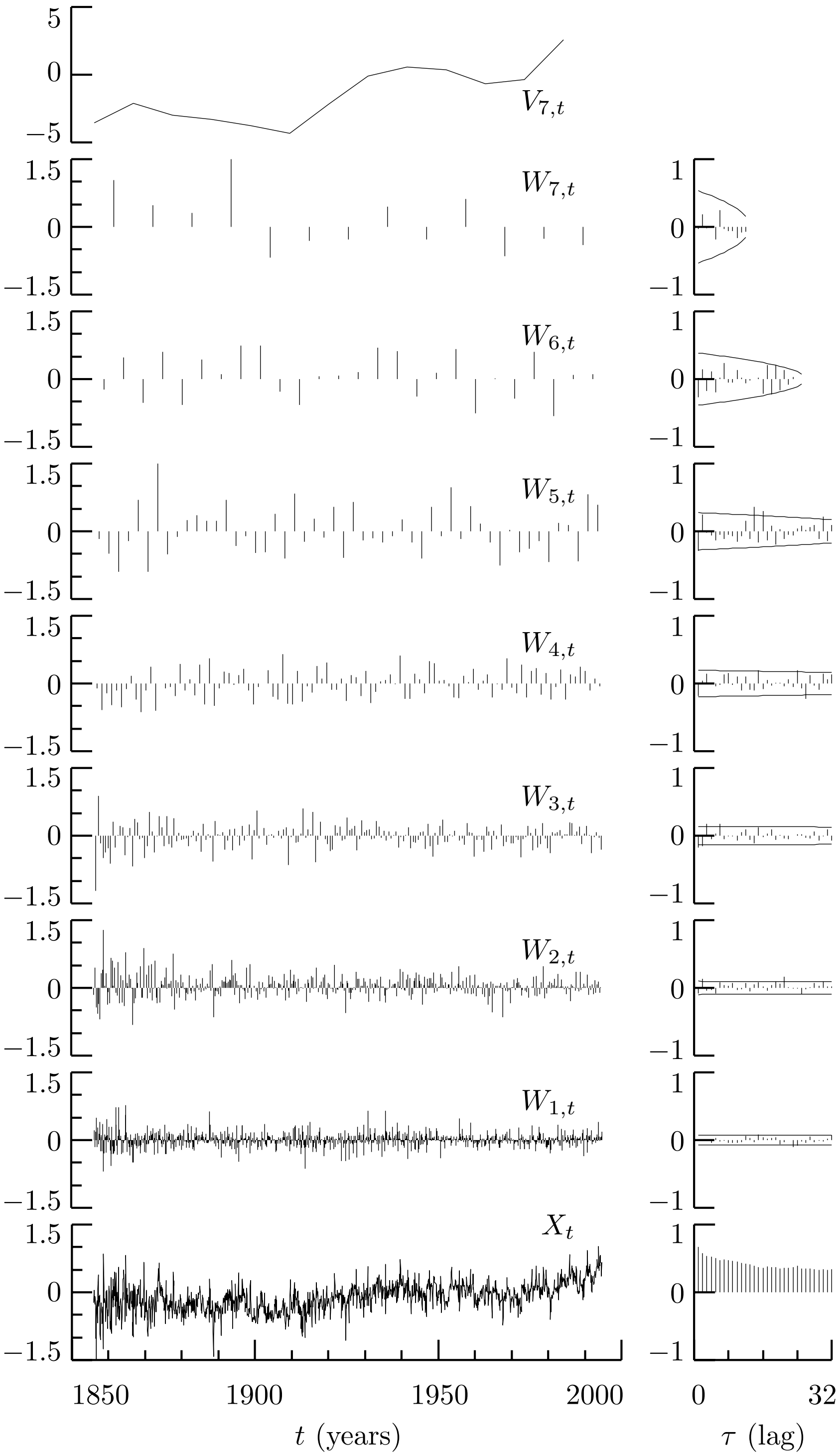

5
0
−5
$V_{7,t}$
1.5
0
−1.5
$W_{7,t}$
1
0
−1
$W_{6,t}$
$W_{5,t}$
$W_{4,t}$
$W_{3,t}$
$W_{2,t}$
$W_{1,t}$
X_t
1850
1900
1950
2000
t (years)
0
32
τ (lag)

Figure 3

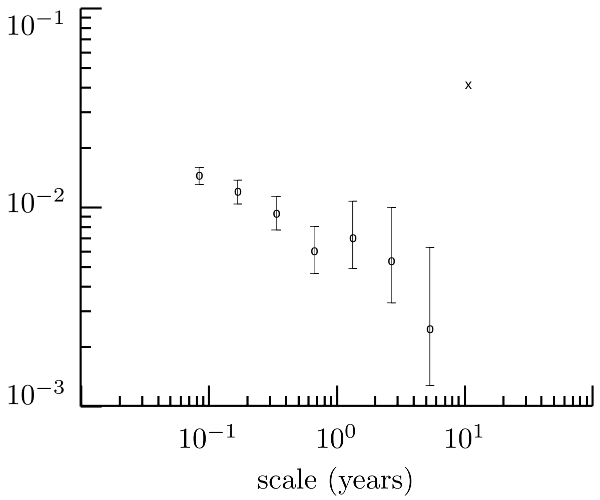
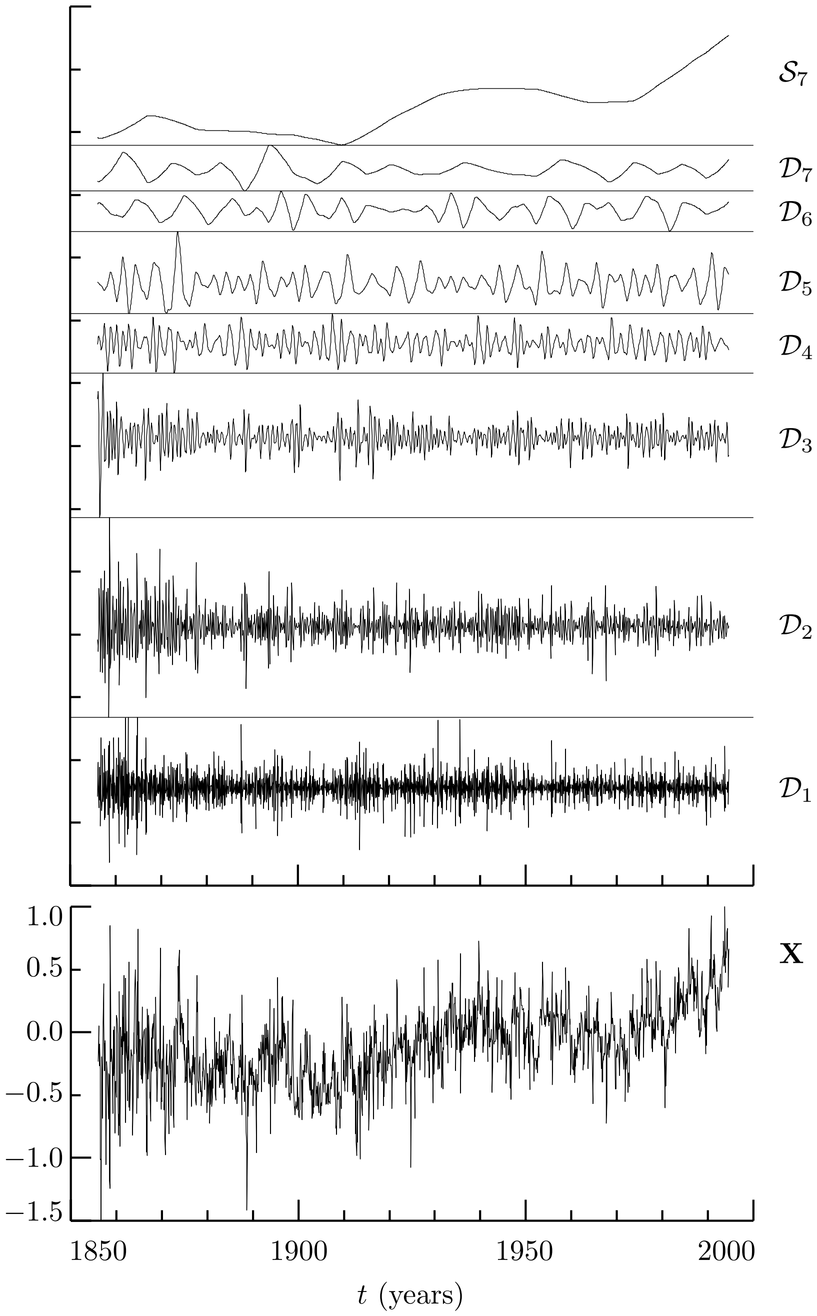


Figure 4



$\mathcal{S}_7$
$\mathcal{D}_7$
$\mathcal{D}_6$
$\mathcal{D}_5$
$\mathcal{D}_4$
$\mathcal{D}_3$
$\mathcal{D}_2$
$\mathcal{D}_1$
$\mathbf{X}$
1.0
0.5
0.0
−0.5
−1.0
−1.5
1850
1900
1950
2000
t (years)

Figure 5

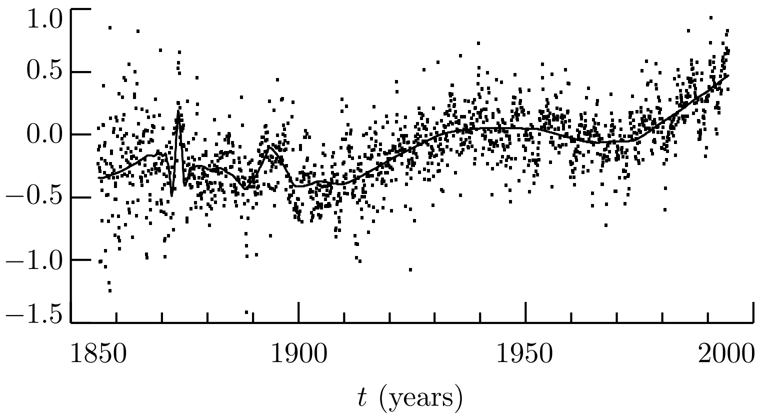


Figure 6

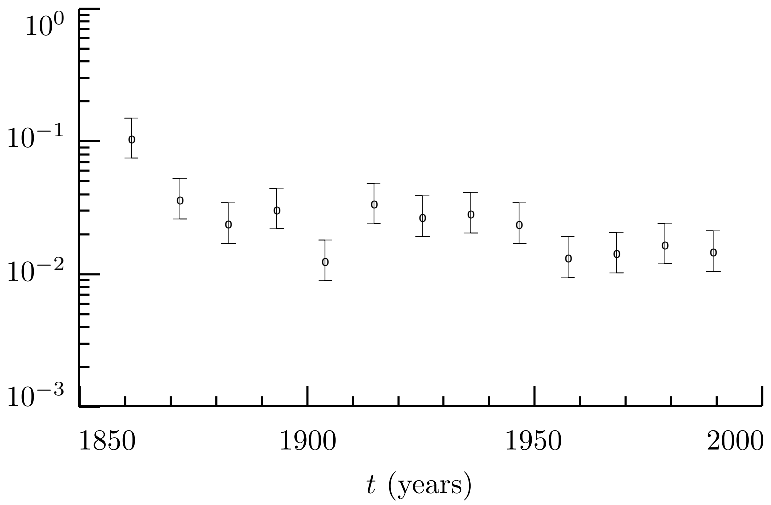


Figure 7