Rapid surrogate testing of wavelet coherences

L. W. Sheppard, P. C. Reid, D. C. Reuman

Appendix S1 The Morlet wavelet transform

We applied a complex Morlet wavelet transform [Addison, 2002] as an example of a continuous wavelet transform suitable for effcient coherence testing. The mother wavelet is

$$\psi(t) = (e^{i2\pi f_0 t} - e^{-(2\pi f_0)^2/2}) \exp(-t^2/2), \tag{1}$$

with $f_0 = 0.5$. Wavelets associated with a range of timescales were produced using a rescaling technique:

$$\psi_\sigma(t) = s^{-1/2}(e^{i2\pi f_0 t/s} - e^{-(2\pi f_0)^2/2}) \exp(-t^2/2s^2). \tag{2}$$

Following earlier convention [Cazelles et al., 2014] we identify each wavelet with a timescale, $\sigma = s/f_0$, and frequency, $f = f_0/s$. The actual peak in the Morlet wavelet power spectrum of a sinusoidal signal with frequency f' and period $\sigma' = 1/f'$ is at $s = ((2\pi f_0 + (2 + (2\pi f_0)^2)^{1/2})/4\pi)(\sigma')$, so $f \approx f'$ only [Meyers et al., 1993]. The centre frequency f_0 of the mother wavelet, which has width $s = 1$ and $\sigma = 2$, was taken to be 0.5 to give a high degree of temporal resolution, but making necessary the subtraction of a constant to keep the mean of the wavelet equal to zero [Addison, 2002]. The mother wavelet was scaled so that one oscillation equaled two years, i.e., $\sigma = 2$, because a two-year period is the highest-frequency fluctuation that can be identified in an annual time series. Wavelets with a range of periods were generated, starting with $\sigma = 2$ and multiplying each period by 1.05 to get the next, up to $\sigma = 44$ for 53 years of data. Convolving a time series $x_n(t)$ from location n with wavelets having different periods produces a set of complex components $w_{n,\sigma}(t) = \sum_{t'} x_n(t + t')\psi_\sigma(t')$. See section 2.2 in the main text for details on how this convolution was performed.

Appendix S2 Details of method

For a given scale σ we define a wavelet $\psi_\sigma(t)$ (such as equation 2), obtained by rescaling some time-localised oscillatory function (a mother wavelet such as in equation 1) so that its Fourier transform has a peak around $1/\sigma$. We write the continuous Fourier transform of the wavelet $\Psi_\sigma(f)$.

A component of the wavelet transform of some data $x(t)$, with (discrete) Fourier transform $X(t)$, is the convolution of the data time series with the appropriate wavelet in the time domain, or (more efficiently) a multiplication in the frequency domain. Here and throughout this appendix we assume Fourier transforms are performed without zero padding the data or the wavelet when generating surrogates and wavelet transforms; thus the wavelet transform is the one based on circular convolution, as defined in section 2.2 of the main text. In short, the wavelet component with timescale σ of the wavelet transform, $w_\sigma(t)$, of $x(t)$, is taken to be the inverse discrete Fourier transform of $W_\sigma(f) = X(f)\overline{\Psi_\sigma(f)}$.

We consider a second time series $y(t)$ with Fourier transform $Y(f)$ and wavelet transform $v_\sigma(t)$ with Fourier transform $V_\sigma(f)$. The wavelet coherence of these unpadded unscalloped transforms is directly related (see equation 4 in the main text) to the quantity

$$\pi_\sigma^w = \sum_t w_\sigma(t)\overline{v_\sigma(t)}, \tag{3}$$

which by Parseval's theorem is

$$\pi_\sigma^w = \sum_f W_\sigma(f)\overline{V_\sigma(f)} \tag{4}$$

$$= \sum_f X(f)\overline{\Psi_\sigma(f)Y(f)}\Psi_\sigma(f). \tag{5}$$

Thus we see that the wavelet coherence at scale σ depends on a weighted average of the un-windowed cross spectrum $X(f)\overline{Y(f)}$.

A Fourier surrogate of $x(t)$ is obtained by Fourier transforming $x(t)$, randomising the phases (assuming real data, these random phase should also have the symmetry properties of the the phases of the Fourier transform of a real time series), and inverse transforming [Theiler et al., 1992]. Thus the surrogate $\tilde{x}(t)$ has a Fourier transform $\tilde{X}(f) = X(f)r(f)$, where $r(f) = e^{i\phi(f)}$ and $\phi(f)$ is the set of random phases defining the surrogate. The wavelet coherence of $\tilde{x}(t)$ and $y(t)$ is

$$\tilde{\pi}_\sigma^w = \sum_f X(f)r(f)\overline{\Psi_\sigma(f)Y(f)}\Psi_\sigma(f). \tag{6}$$

Thus the wavelet coherence depends on the unwindowed cross spectrum $X(f)\overline{Y(f)}$, $r(f)$, and the weighting $\overline{\Psi_\sigma(f)}\Psi_\sigma(f)$. The cross spectrum and weighting need only be calculated once and premultiplied. To obtain a surrogate coherence at a given scale σ it is only necessary to generate $r(f)$ values and perform a weighted sum. $r(f)$ can be generated as many times as necessary to accurately estimate the p-value with which unrelated surrogate data yields a coherence higher than that of the actual data $x(t)$ at a given scale.

An exact solution exists for the probability distribution of the magnitude of the sum of N complex numbers with known magnitudes and random phases, as given in chapter 3 of Jammalamadaka and SenGupta [2001], which is based on the solution provided by Kluyver [1906] to the generalized random walk problem of Pearson [1905]. However, in practice it is more straightforward to generate many random instances than to numerically evaluate the complex integrals in the solution of Jammalamadaka and SenGupta [2001].

It is well-known that wavelet coherence is biased towards higher values at longer timescales; our approach reveals immediately the reasons for this bias. The weighting function $\overline{\Psi_\sigma(f)}\Psi_\sigma(f)$ is sharply peaked around $f = 1/\sigma$, so only a few values make a large contribution to the sum. The width of $\Psi_\sigma(f)$ in the frequency domain is inversely proportional to σ. For example if $\psi_\sigma(f)$ is a Morlet wavelet then $\overline{\Psi_\sigma(f)}\Psi_\sigma(f)$ is a Gaussian with width equal to $1/\sigma$. Thus short-scale coherence values are determined by a sum over many Fourier components and long-scale coherence values are determined by a sum over few Fourier components. This lack of independent data can be addressed by using longer time series, for which more Fourier components fit inside a Gaussian weighting function of given width. If the data themselves contain strong spectral peaks then the coherence at these peaks is dominated by even fewer large components. Thus the wavelet coherence of two independent time series oscillating at the same mean frequency is strongly biased when only a limited number of cycles is available [Sheppard et al., 2012]. Equivalently, we can say that for these strongly periodic timseries relatively few independent values of phase difference are available to determine if there is evidence of a consistent phase difference being maintained.

Appendix S3 Efficient testing of spatial coherence

The linearity of the Fourier transform enables the spatial coherence of a set of spatial surrogates to be evaluated as quickly as the coherence of a single surrogate. Where data are available from N locations, indexed by n, the spatial coherence (equation 5 in the main text, [Sheppard et al., 2015]) is directly

related to

$$\pi_\sigma^s = \sum_n \sum_t w_{n,\sigma}(t)\overline{v_{n,\sigma}(t)}k_\sigma, \tag{7}$$

where

$$k_\sigma = \frac{1}{\sqrt{\sum_t \sum_n w_{n,\sigma}(t)\overline{w_{n,\sigma}(t)} \sum_t \sum_n v_{n,\sigma}(t)\overline{v_{n,\sigma}(t)}}} \tag{8}$$

is a normalisation keeping the coherence between 0 and 1. This can be treated as a prefactor in all calculations of coherence. The power normalisation can alternatively be neglected because the null hypothesis is that the Fourier power spectrum of each time series is as observed (but the relationship between transforms is absent), and the wavelet power spectrum is thus the same for the data and all surrogates neglecting edge effects due to scalloping.

At a given scale, this coherence is dominated by wavelet transform values from locations where the power of the transform is greatest. If this is not desired, a location and scale-dependent power normalisation

$$k_{n,\sigma} = \frac{1}{\sqrt{\sum_t w_{n,\sigma}(t)\overline{w_{n,\sigma}(t)} \sum_t v_{n,\sigma}(t)\overline{v_{n,\sigma}(t)}}} \tag{9}$$

can be used to define

$$\pi_\sigma^l = \sum_n \sum_t w_{n,\sigma}(t)\overline{v_{n,\sigma}(t)}k_{n,\sigma}. \tag{10}$$

This wavelet power normalisation restricts the contribution of the coherence at each location to between 0 and 1, where 1 occurs if $w_n(t) = a_n v_n(t)$ where a_n is a constant complex multiplier at each location n. Including power or other normalisation, the spatial coherence is thus

$$\pi_\sigma^l = \sum_f \sum_n X_n(f)\overline{\Psi_\sigma(f)Y_n(f)}\Psi_\sigma(f)k_{n,\sigma}. \tag{11}$$

When a spatial surrogate is generated the $X_n(f)$ at every location are subject to the same phase shift, and so we can write the spatial coherence of a surrogate data set as

$$\tilde{\pi}_\sigma^l = \sum_f r(f) \sum_n X_n(f)\overline{\Psi_\sigma(f)Y_n(f)}\Psi_\sigma(f)k_{n,\sigma}. \tag{12}$$

Once again each surrogate value is determined by a multiplication and a sum, with the product of all the terms but the $r(f)$ calculated one time, in advance. For a given σ, we can find the rank of the magnitude of π_σ^l relative to the distribution of $\tilde{\pi}_\sigma^l$ magnitudes.

MATLAB code to efficiently test wavelet coherence and spatial wavelet coherence is included in supplementary materials.

Appendix S4 Data

Figure S1 shows the areas of the North Sea and British seas which were used in this analysis. Organism counts for a range of species shown in table S1 were taken from the Continuous Plankton Recorder (CPR) data set maintained by the Sir Alistair Hardy Foundation for Ocean Science (SAHFOS). The CPR is a device which is towed across the ocean behind a ship, filtering plankton from seawater with a continuously unwinding silk ribbon. The ribbon is cut into 4 inch segments, each corresponding to a particular time and location in the ocean. Organisms are identified and counted (possibly with sub-sampling) microscopically (Colebrook and Robinson [1965]).

Data about physical oceanographic variables was downloaded from the International Comprehensive Ocean-Atmosphere Data Set (ICOADS) Release 2.5. Temperature and salinity measurements from the top ten metres of the water column were included, as was wind speed at the sea surface, and cloud cover, recorded as proportion of the sky area that was observed to be obscured by cloud, discretized into oktas, or eighths. See table S2.

All the samples available for a given variable in a given area in a given month of a given year were averaged, producing a monthly time series from 1958 to 2010 inclusive. Where data for a given month was unavailable the missing value was replaced with a median value for that month (drawn from all years in which data was available). Monthly values were converted to yearly values by averaging over all months. For the physical variables, seasonal time series were also generated by averaging over a subset of months within each year.

A Box-Cox algorithm was implemented to produce comparable time series with near-normal distributions. Unlike many ecological time series the values were not generally discretized into integer count or date values. For Box-Cox normalisation it is standard to add or subtract a constant from all values prior to transformation, setting the minimum to one. We set the minimum to be equal to the difference between the lowest two distinct values in each time series. For each ecological or environmental variable, the single optimal Box-Cox coefficient was found for all 26 time series of data. After the distribution was normalised each time series was reduced to zero mean and unit standard deviation. Box Cox coefficients for biological variables tended to be around 0, for physical variables around 1, indicating that the raw physical values were approximately normally distributed and biological counts approximately log normal.

Appendix S5 Multiple-frequency testing

We considered wavelet frequencies below 0.25 cycles per year as the low frequency band and frequencies above 0.25 cycles per year as the high frequency band. Using efficient methods to calculate coherences, we ranked the actual spatial coherence of the data against S (a large number of) surrogates for each frequency. We averaged this rank across each band. Then we ranked each surrogate spatial coherence against all the other surrogates (by an efficient sorting algorithm), and found the average rank across each band for the surrogate. The proportion of surrogate average ranks greater than the average rank of the actual data was the p-value for the data (ie. the probability of obtaining coherence at least as great as that observed in this band under the null hypothesis).

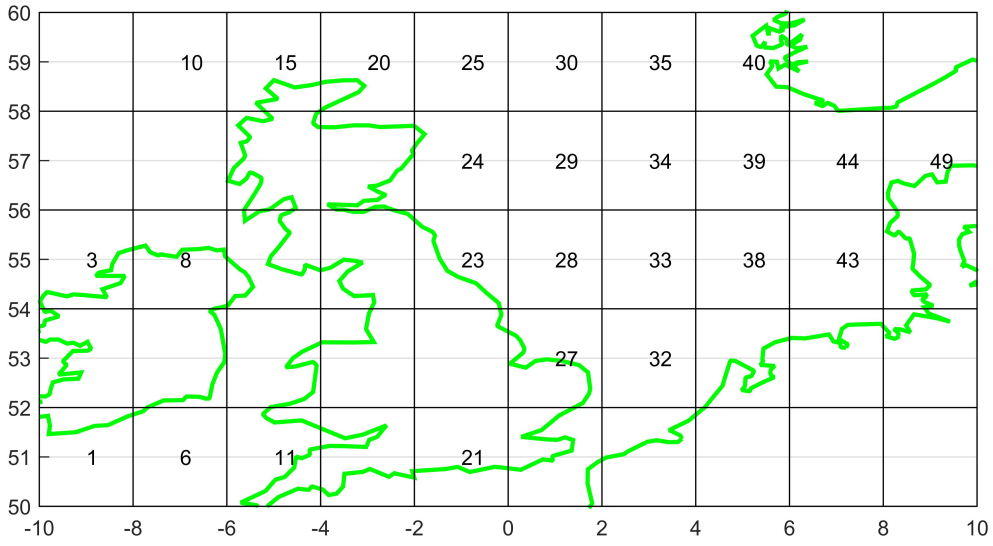

Figure S1: The 26 numbered areas of the North Sea and British seas which were used in this analysis. Some cells overlap land; oceanic parts of the cells were used.

Table of ecological variables

Phytoplankton: dinoflagellates
Thalassiosira spp.
Pseudo-nitzschia delicatissima complex
Pseudo-nitzschia seriata complex
Rhizosolenia styliformis
Proboscia alata
Phytoplankton: Ceratium
Ceratium fusus
Ceratium furca
Ceratium tripos
Ceratium macroceros
Zooplankton
Calanus stages I–IV
Para-pseudocalanus spp.
Acartia spp. (unidentified)
Oithona spp.
Pseudocalanus elongatus Adult
Temora longicornis
Centropages typicus
Calanus finmarchicus
Calanus helgolandicus
Metridia lucens
Echinoderm larvae
Decapoda larvae (Total)
Euphausiacea (Total)

Table S1: The names of the plankton investigated. Each time series was constructed by averaging monthly values over all twelve months, to produce one value per year per location.

Table of environmental variables

Env. variables	Months averaged
Yearly temperature	1 to 12
Spring temperature	3 to 5
Summer temperature	6 to 8
Autumn temperature	9 to 11
Growing season temperature	3 to 9
Yearly wind speed	1 to 12
Spring wind speed	3 to 5
Summer wind speed	6 to 8
Autumn wind speed	9 to 11
Growing season wind speed	3 to 9
Yearly salinity	1 to 12
Spring salinity	3 to 5
Summer salinity	6 to 8
Autumn salinity	9 to 11
Growing season salinity	3 to 9
Yearly cloud cover	1 to 12
Spring cloud cover	3 to 5
Summer cloud cover	6 to 8
Autumn cloud cover	9 to 11
Growing season cloud cover	3 to 9

Table S2: The names given to the environmental variables investigated. Each time series was constructed by averaging monthly values over the relevant months, to produce one value per year per location.

References

P. S. Addison. *The Illustrated Wavelet Transform Handbook: Introductory Theory and Applications in Science, Engineering, Medicine and Finance*. Taylor and Francis, New York, 2002.

B. Cazelles, K. Cazelles, and M. Chavez. Wavelet analysis in ecology and epidemiology: imapct of statistical tests. *Journal of the Royal Society Interface*, 11:20130585, 2014.

J. M. Colebrook and G. A. Robinson. Continuous Plankton Records: Seasonal cycles of phytoplankton and copepods in the North-Eastern Atlantic and the North Sea. *Bull. Mar. Ecol.*, 6:123–139, 1965.

S. R. Jammalamadaka and A. SenGupta. *Topics in circular statistics vol. 5*. World Scientific, Singapore, 2001.

J. C. Kluyver. A local probability theorem. *Ned. Akad. Wet. Proc., Ser. A*, 8:341–350, 1906.

S. D. Meyers, B .G. Kelly, and J. J. O'Brien. An introduction to wavelet analysis in oceanography and meteorology: with application to the dispersion of yanai waves. *Mon. Weather Rev.*, 121:2858–2866, 1993.

K. Pearson. The problem of the random walk. *Nature*, 72:294–342, 1905.

L. W. Sheppard, J. R. Bell, R. Harrington, and D. C. Reuman. Changes in large-scale climate alter spatial synchrony of aphid pests. *Nature Climate Change*, 2015. doi: 10.1038/nclimate2881.

L. W. Sheppard, A. Stefanovska, and P. V. E. McClintock. Testing for time-localized coherence in bivariate data. *Phys. Rev. E*, 85(4, 2), APR 9 2012. ISSN 1539-3755. doi: 10.1103/PhysRevE.85.046205.

J. Theiler, S. Eubank, A. Longtin, B Galdrikian, and J.D. Farmer. Testing for nonlinearity in time series: the method of surrogate data. *Physica D*, 58(1–4):77–94, 1992.