esa ECOSPHERE

A multiscale approach to water quality variables in a river ecosystem

El-Amine Mimouni,[1,]† Jeffrey J. Ridal,[1] Joseph D. Skufca,[2] and Michael R. Twiss[3]

[1]River Institute, Cornwall, Ontario, Canada
[2]Department of Mathematics, Clarkson University, Potsdam, New York 13699 USA
[3]Department of Biology, Clarkson University, Potsdam, New York 13699 USA

Citation: Mimouni, E.-A., J. J. Ridal, J. D. Skufca, and M. R. Twiss. 2020. A multiscale approach to water quality variables in a river ecosystem. Ecosphere 11(2):e03014. 10.1002/ecs2.3014

Abstract. Monitoring select ecosystem variables across time and interpreting the collected data are essential components of ecosystem assessment supporting management. Increasingly affordable sensors and computational capacity have made very large dataset assembly more common. However, these datasets initiate analytical challenges by their size and theoretical challenges due to the scale of the processes they encompass. Multiscale assessment of high temporal resolution water quality sensor data (temperature, in vivo chlorophyll *a*, colored dissolved organic matter) collected year-round was conducted for the Upper St. Lawrence River. Using numerical methods that directly integrate the concept of scale, we show that consideration of scale-dependent processes can lead to increased predictive power and a clearer understanding of ecosystem function. These results suggest that multiscale methods are not only an alternative way of approaching long-term data assessment, but also a necessity in order to avoid spurious interpretation. Consequently, the concept of scale as described here can be consistently integrated into long-term data studies to assist in the interpretation of high-resolution data that help describe natural phenomena in aquatic systems.

Key words: limnology; modeling; rivers; scale; temporal scale; water quality.

Received 28 March 2019; accepted 14 May 2019. Corresponding Editor: Debra P. C. Peters.

† **E-mail:** elamine.mimouni89@gmail.com

Introduction

Large rivers are ecologically, economically, and socially important ecosystems. Assessment and management of priority resource issues such as fish migration, detection of tributary and point-source nutrient enrichment, fecal bacterial contamination, contaminant transport and fate, harmful cyanobacterial blooms, and climate change impacts are issues that require attention. Strong anthropogenic influence resulting from the rivers flowing past the urban centers, agricultural regions, and dam construction makes their studies a complex matter. However, both short-term periodicities and long-term trends and changes in rivers require attention. To adequately study both components, long-term ecosystem research (LTER) represents an important tool. An essential step to LTER is developing techniques that capture data and interpret observations on appropriate timescales so that processes that impact river environments can be detected and understood with the aim of informing ecosystem-based management actions (Parr et al. 2003).

Advances in water quality instrumentation and data management tools (data recording, storage, analysis, and exchange) make it possible to monitor aquatic ecosystems in a comprehensive and cost-effective manner across a wide range of timescales. The combination of recording observations at high frequency over a long (e.g., annual) time period relative to the periods over which biogeochemical changes in the ecological system can be observed means that such studies can study ecosystem processes that occur over a wide range of temporal scales. Thus, the issue of scale is a central point in the theoretical and analytical approaches used to assess long-term and highly temporally resolved datasets. However, the concept and its measure of scale remain difficult to define, as they have been attached to various properties of systems (Wiens 1989, Levin 1992, He et al. 1996, Thrush 1997, Dungan et al. 2002). Nonetheless, it remains a central concept of ecological literature, as one gains insights into ecosystem function by studying time-series analysis with appropriate methods.

The capacity to collect highly resolved temporal data with automated water quality sensor arrays can provide a rich database to investigate relationships among measured parameters and to generate insights into the factors that structure the aquatic environment in the river. Using methods adapted to such data, various insights have been obtained regarding watershed and ecosystem behavior (Kirchner et al. 2000, 2001, Arora et al. 2016, Schmidt and Sutton 2018). Within the Great Lakes–St. Lawrence River system, there are a limited number of sensor arrays capable of supporting LTER with very few deployed in the rivers that provide lake to lake drainage throughout the Great Lakes–St. Lawrence River system. The most commonly used buoy-based sensors are restricted to ice-free periods, typically May–December (Twiss and Stryszowska 2016). To develop and assess the ability automated sensor arrays to resolve large river ecosystem processes, an observational platform of commercially available water quality sensors was installed inside of a hydropower dam on the St. Lawrence River. The location of the sensor array allows the project to support water quality measurements year-round.

The main goal of our study was to apply a consistent pathway and methodology to analyze high temporal resolution data while considering the notion of scale. Furthermore, we show that a theoretical and numerical integration of the concept of temporal scale allows for the construction of better interpretative models. Here, water quality sensor data are first analyzed using a wavelet transform in order to detect temporal structures not only across scales but also across time. Second, temporal models using Moran's eigenvector maps (MEM) spatial filtering variables are constructed and used to assess the importance of temporal structures in the data. Third, the possibility of establishing linear models among water quality and the measured variables is explored by verifying their assumptions using direct multiscale ordination. Finally, the variables are related to each other using spatial eigenvectors as modulating variables in multiscale codependence analysis that describe the various scales considered. The aim of this effort is to obtain information that can help explain the variability inherent in the observations of natural systems so that mechanisms that cause system change can be clearly identified.

Methods

Study site and data acquisition

A multisensor array was installed in Unit 32 power turbine of the Moses-Saunders hydroelectric dam, along the New York shoreline of the St. Lawrence River (45°0.253′ N, 74°47.945′ W; Fig. 1). The Moses-Saunders hydropower dam is in the Upper St. Lawrence River, between the state of New York (USA) and the province of Ontario (Canada). The Upper St. Lawrence River, defined by the water upstream of the Moses-Saunders hydropower dam to Lake Ontario, is characterized by little tributary input (~3% of its average annual discharge of 6800 m^3/s), and distinct nearshore and main channel regions (Ball et al. 2018).

The multisensor array consisted of a Turner Designs C6 multisensor platform equipped with Cyclops-7 sondes. The array measured various water quality parameters every minute, including water temperature, colored dissolved organic matter (CDOM; Suwannee River fulvic acid equivalents), in vivo chlorophyll *a*, and in vivo phycocyanin. Water from the penstock was drawn via a 30 cm diameter pipe to cool the stator of the turbine-driven electric generator; this water was effectively mixed surface and bottom (~20 m depth) river water but is restricted to water that flows along the southern shoreline of the river owing to the location of the Unit 32 turbine, which is nearest that shore. The C6 is housed in a watertight flow-through cell and is

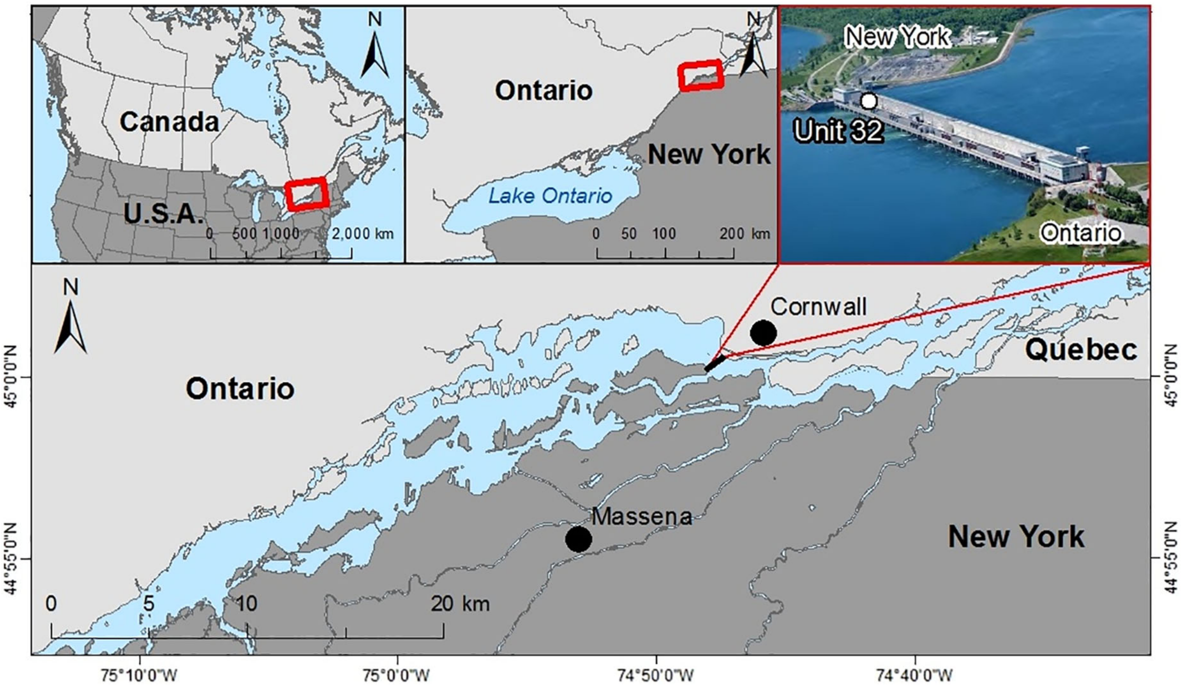

Fig. 1. Geographical location of the Unit 32 power turbine of the Moses-Saunders hydroelectric dam on the Upper St. Lawrence River.

connected to the cooling water pipe via a stainless-steel pipe (1 cm diameter) with a pressure reduction gate valve. The array was equipped with an anti-fouling brush, which performs one revolution every minute prior to recording water quality observations to prevent any fouling by debris or organisms on the optical sensor surfaces. The entire system was visited at 1- to 3-week intervals to download data, clean instruments, and recalibrate.

Data were averaged into three-hour blocks to avoid overly long computing times and memory roadblocks during analyses. We found that this was the best value that offered a trade-off between computational efficiency using a desktop personal computer and obtaining results within reasonable amounts of time using the available conventional computing power. The measured variables were individually transformed to reduce skewness as much as possible. Colored dissolved organic matter was $\log_e$ transformed, chlorophyll *a* was fourth root transformed, phycocyanin was square root transformed, and temperature was left untransformed.

Consideration of scale

Due to the possibility that the data contain several patterns of various types at different temporal scales, we chose to incorporate the notion of scale into our analyses at several points. Another term closely related to scale is structure (Borcard and Legendre 2002, Legendre and Legendre 2012). Essentially, an assessment of the various temporal scales over which a process may operate can be considered through the creation and integration of various temporal structures into analyses. To fully define these concepts in ecology would be beyond the scope of this paper, though we refer readers to references such as Fortin and Dale (2007) or Legendre and Legendre (2012). Nonetheless, we note that, to be useful in modeling, a temporal structure should be well defined in terms of location (i.e., when it occurs) and in scale (i.e., over how much time does it take place). Consideration of both aspects is necessary to fully analyze temporal datasets.

From time-referenced data, two individual components can be extracted: the environmental data itself and the timestamps. From the timestamps alone, a wide variety of temporal structures can be constructed and incorporated into analyses. To make our goals and methods clear as well as to allow readers to follow our reasoning, we have provided a flowchart along with the links between the types of temporal structures used and the question that each analysis seeks to answer (Fig. 2). The data processing flowchart is not exhaustive, as several additional links could be added. For example, variation partitioning (Borcard et al. 1992, Borcard and Legendre 1994) and multiscale codependence analysis (Guénard et al. 2010) are versatile methods that are not limited to their use with MEM and could be used with wavelets. Likewise, MEMs can also be used in direct multiscale ordination (DMSO; Borcard et al. 2016). Finally, there are also other methods that consider scale, which were not considered here, such as the method applied by Keitt and Urban (2005) or Legendre et al. (2009). Nonetheless, we note that the building blocks of these methods are still MEMs and wavelets, so they can be readily added to the flowchart.

We also note that such a data processing protocol is not absolute, as the question that DMSO answers given here differs from that provided by Guénard and Legendre (2017*a*). Likewise, Percival et al. (2004) give at least four questions that can be answered by wavelet analysis, which somewhat differ from the two used here. Additionally, the interpretation of some structures revealed by the different methods may vary (Wagner 2004, Borcard et al. 2016). The figure can be made much more complex than presented here. Nonetheless, it serves to summarize the types of questions that can be answered by these methods and why we chose these tools over others.

Wavelet transform analysis

Wavelets are mathematical functions that are capable of detecting patterns at different scales or frequencies for temporal and spatial data (Bradshaw and Spies 1992, Dale and Mah 1998,

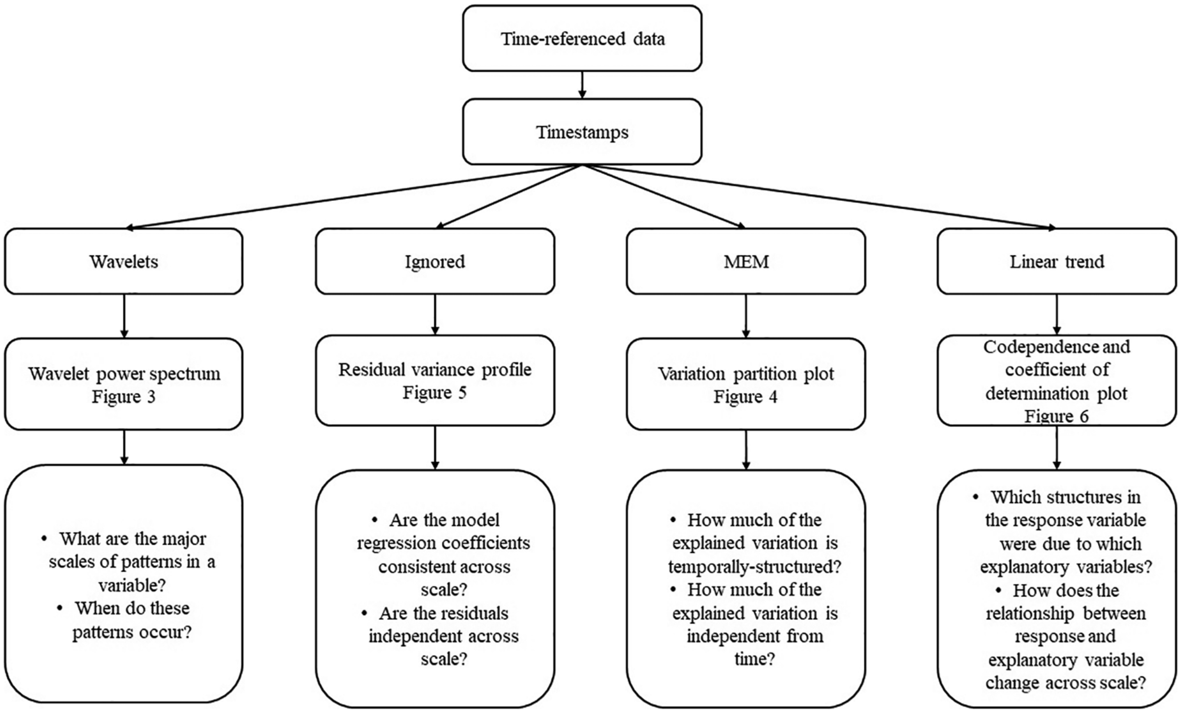
Time-referenced data
Timestamps
Wavelets
Ignored
MEM
Linear trend
Wavelet power spectrum
Figure 3
Residual variance profile
Figure 5
Variation partition plot
Figure 4
Codependence and
coefficient of
determination plot
Figure 6
• What are the major scales of patterns in a variable?
• When do these patterns occur?
• Are the model regression coefficients consistent across scale?
• Are the residuals independent across scale?
• How much of the explained variation is temporally-structured?
• How much of the explained variation is independent from time?
• Which structures in the response variable were due to which explanatory variables?
• How does the relationship between response and explanatory variable change across scale?

Fig. 2. Flowchart of the scale-incorporating methods. Blocks at the end of each path show the main questions that the method addresses. From timestamps, a variety of temporal structures can be obtained and integrated into numerical methods; ignored means that timestamps were not directly incorporated into the model, but rather serve only to decompose the variance across distance classes.

Percival et al. 2004, Cazelles et al. 2008, James et al. 2010). Their main feature is that they represent a local decomposition of a signal, as opposed to other methods, which assume that the signal retains its characteristics throughout the sampling extent. Consequently, information regarding the intensity of the studied process can be obtained not only with regard to scale, but location as well. A Morlet wavelet (Morlet et al. 1982*a*, *b*) was used as a template, whose mother wavelet function is

$$\psi(t) = \frac{1}{\sqrt[4]{\pi}}\left(e^{i\omega_0 t} - e^{-\frac{\omega_0^2}{2}}\right)e^{-\frac{t^2}{2}}, \qquad (1)$$

where ω_0, is the dimensionless frequency parameter, which determines the number of oscillations within the mother wavelet. We considered a value of $\omega_0 = 6$, which is an appropriate parameter value for feature extraction, as it considers a general peak, along with four smaller ripples on either side. Furthermore, we note that in this case, the second term inside the parentheses can be ignored, as it is only a correction factor, which accounts for the nonzero mean of the complex sinusoid (Addison et al. 2002).

Significant scales of variation for the data were tested using surrogate data testing using simulations. We considered random processes which can be modeled by power-law noise model (Marvin 1982). We refer to power-law noise as a process where the power spectral density is inversely proportional to a power of the frequency, so that

$$S(f) \propto \frac{1}{f^{\alpha}}, \qquad (2)$$

In this case, four different values of the exponent α were chosen, to represent different colors of noise. The first null model was a white noise model, where $\alpha = 0$. In this case, each new time-series consisted of n points drawn from a standard $\text{Normal}(\mu = 0, \sigma^2 = 1)$ distribution. This model is a basic null model (it is the default in the analyze.wavelet function), but it might be somewhat too liberal, as it considers that the data are completely independent and that there is inherently no temporal structure to be expected in the data. The second null model was a red noise model, where $\alpha = 2$. The third null model was a pink noise model, where $\alpha = 1$. The fourth and final model first estimated the α coefficient as the negative of the slope of a log-log regression between the spectrum of the data and the harmonic frequencies. Missing values were estimated by linear interpolation between available data points. To generate data according to a particular power-law noise model, the algorithm of Timmer and König (1995) was used.

Moran eigenvector maps analysis

In order to detect and assess temporal structures, MEM (Dray et al. 2006) were considered. They comprise a flexible family of eigen-based spatial filtering method (Griffith and Peres-Neto 2006) which allow the modeling of spatial processes at different scales. As has been raised, these variables are also well-suited to model temporal processes as well (De Cáceres et al. 2010, Legendre and Legendre 2012, Legendre and Gauthier 2014). Indeed, temporal studies can be considered like a unidimensional transect in spatial studies. Consequently, for the remainder of the paper, we will use the term temporal even though much of the literature considers their application to spatial problems.

Moran eigenvector maps variables were constructed by considering the eigendecomposition of the Hadamard product between a connectivity matrix and a weighing matrix (Dray et al. 2006, Legendre and Legendre 2012). If a significant linear trend was detected, the response variable was regressed against a straight line representing a linear temporal trend and the residuals comprised the detrended data to be used for the MEM analysis. We considered only variables modeling positive autocorrelation, whose corresponding eigenvalue satisfied the criterion $\lambda_k \geq -\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}/n(n-1)$.

Two different connectivity scheme matrices were considered, either based on the minimum-spanning tree (MST) or based on a distance criterion (DNN). In the case of time-series analysis, MST connections are equivalent to considering each observation connected only to the previous and following observations as the minimum-spanning tree is a straight line. However, despite labeling this connection scheme MST, it could actually be several types of connection schemes such as relative neighborhood graph (Toussaint 1980), Gabriel graph (Gabriel and Sokal 1969), or Delaunay triangulation (Delaunay 1934), as all these methods will give the same straight line for a temporal series. In the case of DNN, an upper threshold of 98.125 d was selected, as this represents the shortest distance that would keep all points connected. In addition, four different weighing schemes for the weighing matrix were considered in order to seek out the best possible model. These weights included no weights, which would give a binary model (BIN) or a weighing of the edges based on a linear ($f_1 = 1 - d_{ij}/\max\left(d_{ij}\right)$), concave-up ($f_2 = 1 - \left(d_{ij}/\max\left(d_{ij}\right)\right)^a$), or concave-down ($f_3 = 1/d_{ij}^{\beta}$) function of the distance between sampling days. Values of the exponent α were computed for all values between 2 and 10 (since $\alpha = 1$ would be the linear model) and those of the exponent β between 1 and 10. Principal coordinates of neighboring matrices (PCNM; Borcard and Legendre 2002) eigenfunctions were also considered.

A data-driven approach was taken (Getis and Aldstadt 2004, Dray et al. 2006), wherein several models were considered and the selected model was the one that best fit the data at hand. Two possible methods were considered to determine the best model for individual variables: Akaike's information criterion value (AIC; Akaike 1974) and forward selection of variables. Both of these methods are facilitated by the orthogonal nature of MEM, as they can simply be entered them into the model according to the value trace$\left(Y^T u_i u_i^T Y\right)$, which is the value by which the total sum of squares are decreased by the inclusion of MEM variable u_i in the model. We note that, in our case where there is only a single response variable, computing the trace is not necessary, as the product $Y^T u_i u_i^T Y$ will give a 1×1 matrix.

Second-order AIC values were considered (Sugiura 1978, Hurvich and Tsai 1989), which are defined as

$$\mathrm{AIC_c} = (-2\log(\mathcal{L}) + 2k) + \frac{2k(k+1)}{n-k-1}, \quad (3)$$

where n is the number of observations, k is the number of variables in the model, and $\mathcal{L}$ is the likelihood function. The correction is suggested mostly for cases in which the number of variables is much higher than the number of points. In MEM analysis, a large amount of temporal variables can be considered, which can make this consideration necessary. Furthermore, given that AIC is the limit of $\mathrm{AIC_c}$ as sample size tends to infinity, it should always be employed (Burnham and Anderson 2002). In our case, we considered a linear model, wherein the term $-2\log(\mathcal{L}) = n\log(RSS/n)$ (Burnham and Anderson 2002, Godínez-Domínguez and Freire 2003).

Forward selection of variables using a double-stopping criterion (Blanchet et al. 2008) was also considered. In this case, variables are entered into the model as in classical forward selection, but the process also stops if the subset of variables has an adjusted R^2 value that is higher than the entire set of candidate variables. We note that the forward.sel function actually implements five stopping criteria. The most relevant of these are the previously mentioned two and the R^2 more criterion, wherein forward selection stops if the candidate variable explains less than a certain value of the variation. We used a value of 0.001, meaning that forward selection stopped if the candidate variable explained <0.1% of the variation in the response variable.

Variation partitioning (Borcard et al. 1992, Borcard and Legendre 1994) was used to compare the fractions of variations explained by the environmental model and the temporal model. Coefficients of multiple determination values were adjusted following Ezekiel's formula (Ezekiel 1930, Peres-Neto et al. 2006). Even though the linear trend is not an eigenfunction submodel, the same rules as those of variation partitioning involving orthogonal eigenfunction submodels were applied (Legendre et al. 2012). The method consists of transferring shared fractions of explained variation to the spatial model of higher scale. In our case, shared fractions of variation of the temporal model with the trend were transferred over to the trend, as it is taken to be a process occurring at a larger scale than considered by the study.

Direct multiscale ordination analysis

Conventional multivariate linear models such as those from redundancy analysis (RDA; Rao 1964) or canonical correspondence analysis (CCA; ter Braak 1986) consider that the effects of the constraining variables are the same at all scales and that the residuals of such a model are independent and identically distributed. Both assumptions can affect results, leading to various effects from tests being too liberal to regression coefficients being inconsistent across scales. To check whether these assumptions were met, we used Wagner's method (2003, 2004) which decomposes the covariance matrix of a multivariate dataset (Σ) into a series of distance-dependent covariance matrices $\Sigma(d)$, whose elements are computed as

$$\sigma_{ij}(d) = \frac{1}{2n_{d_k}} \sum_{a,b|d_{ab} \approx d_k} (x_{ai} - x_{bi})(x_{aj} - x_{bj}), \quad (4)$$

or a pair of observations a and b that are separated by distance d_{ab}. In this case, the covariance matrix used in principal components is a weighed sum (with weights equal to the proportions of number of pairs in each distance class); this partitioning between distance classes, combined with the outputs of constrained multivariate analyses, allows for the decomposition of both matrices of fitted values and residuals into scale-dependent classes. Accordingly, this decomposition forms the basis of DMSO (Wagner 2004) that allows testing of the previously mentioned assumptions. Scale dependence of the regression coefficients of a linear model predicting CDOM and chlorophyll *a* were assessed by verifying if the sum of scale-dependent variogram values for the fitted values and the residual values of the model fell outside of an envelope around the variogram for the response variable. The envelope for the response variable variogram was computed as

$$\gamma_Y(h) \pm z_\alpha \sqrt{\frac{\mathrm{Var}(\gamma_Y(h))}{n_h}}, \quad (5)$$

where $\gamma_Y(h)$ is the total semi-variance for class h and z_α is a critical value for a standard $\mathrm{Normal}(\mu = 0, \sigma^2 = 1)$ distribution. Significant temporal autocorrelation for the residuals was tested by carrying out a Mantel test with 1000 permutations (Mantel 1967, Legendre and Legendre 2012). To account for multiple testing, a Bonferroni correction was applied by dividing the significance level by the number of performed tests.

Multiscale codependence analysis

The relationship between variables can be decomposed and assessed across scales by considering covariables that represent the different scales at which this relationship may express itself. This is the central idea of multiscale codependence analysis (MCA; Guénard et al. 2010). In this analysis, codependence coefficient for two variables x and y at the temporal scale described by vector u_i is defined as

$$C_{y,x|u_i} = \frac{u_i^T y}{\sqrt{y^T y}} \frac{u_i^T x}{\sqrt{x^T x}}, \quad (6)$$

where u_i is drawn from a set of orthonormal spatial eigenvectors U. This is the product of two Pearson product–moment correlations of the two studied variables, each with u_i, a temporal variable which models the process at a certain scale. The scale-dependent temporal variables were the unit-normed eigenvectors of the best fitting temporal model, as determined by $\mathrm{AIC_c}$. In our case, response variables were always univariate as MCAs of CDOM or chlorophyll *a* against the other variables were considered. Contrary to the multivariate version of MCA (Guénard and Legendre 2017*a*), the sign of this codependence coefficient is meaningful and should be kept.

Tests of significance were carried out by permuting the observations of the variables under the null hypothesis of temporal independence between the two variables, as suggested by Guénard et al. (2010). In order to maintain adequate familywise error rate, a sequential Šidák correction was applied (Šidák 1967, Wright 1992). The entire set of spatial eigenvariables coding for positive autocorrelation for the best model were used as spatial predictors. We used an adaptive form of testing, where the number of permutations or simulations to be carried for each step was determined by the formula

$$n_{\mathrm{perm}} = \frac{1}{1 - (1 - \alpha)^{\frac{1}{pm-q}}}, \quad (7)$$

for a situation where p is the number of explanatory variables, m is the number of temporal variables, q is the number of temporal variables already in the model, and α is the desired significance level. In order to ensure that the null hypothesis could be adequately rejected, we considered 10 times the suggested number of permutations.

We investigated the effect of the null model and the surrogate data generating algorithm on the significance of the codependence coefficients by considering two additional methods of generating the surrogate dataset. The first procedure consisted of creating a new set of explanatory variables under a power-law noise model (Eq. 2). Surrogate variables were obtained using the generating algorithm of Timmer and König (1995). In our study, we assumed that the simulated explanatory variables followed a power-law noise model whose exponent α was determined as the negative of the slope of a log–log relationship between the spectrum and the harmonic frequencies of the variable. The second procedure consisted of bootstrapping the dataset, but pseudo replicates were obtained by resampling with replacement contiguous blocks of data, rather than single observations. This approach is essentially similar to block bootstrapping (Carlstein 1986, Künsh 1989, Efron and Tibshirani 1993). Block size was considered fixed within an iteration, but was also varied to assess its effect on the estimate. A new MCA was conducted using these variables as explanatory variables and the distribution of the test statistic τ was tabulated. Given the sometimes uneven sampling (i.e., presence of gaps), care was taken to generate temporal series with similar characteristics (i.e., gaps of the same size and at the same locations). The advantage of this methodology is that it does not assume complete randomness of observations, thus allowing for a more realistic null hypothesis. Using this method, the strongest codependence was tested for both models.

The reasoning behind this statement is that the three methods actually test three different null hypotheses.

1. Random permutation: The codependence between the two variables is similar to that which would be seen for a pair of nonstructured variables that have identical values, but are completely independent of the temporal axis;
2. Block bootstrap: The codependence between the two variables is similar to that which would be seen for a pair of structured variables that are drawn from a population of blocks similar to that of the observed variables, which show a similar autocorrelation structures below block size, but whose structure above block size has been destroyed; and
3. Power-law noise: The codependence between the two variables is similar to that which would be seen for a pair of structured variables following a stochastic random process, where each variable shows a similar autocorrelation structure to the measured variable (i.e., a similar spectrum).

We note that this is not a trivial issue, as it has been noted that some variants of randomization tests might be ill-advised in spatial statistics, where complete spatial randomness may prove to be an absurd null hypothesis (Fortin and Jacquez 2000). Additionally, it would be wrong to draw conclusions based on the rejection of a null hypothesis that is not well-grounded theoretically. Consequently, the matter represents not only a mathematical issue, but also an ecological one, as ecologists should decide which null hypothesis to test.

Within MCA, it is possible that the response variable shows significant codependence with more than one variable at a single scale. Such a situation was envisioned by the authors who suggested retaining the largest absolute statistic in order to preserve strictly exclusive sets of structuring variables (Guénard et al. 2010, Guénard and Legendre 2017*a*). However, the frequency of such situations is somewhat unknown. Furthermore, correlations between explanatory variables and random variation due to sampling could cause one variable to be selected on one occasion, but another on the next. To further explore this question, we introduce the notion of uncertainty coefficients. The uncertainty coefficient for explanatory variable *j* and structuring variable *i* is defined as

$$\mathrm{UC}(j,i) = 1 - \frac{\max\left(\left|C_{y,x|u_i}\right|\right) - \left|C_{y,x|u_i}\right|}{\max\left(\left|C_{y,x|u_i}\right|\right)}. \quad (8)$$

In other words, the uncertainty coefficient is one minus the standardized difference between the absolute codependence coefficient for variable *j* and the highest absolute codependence coefficient for structuring variable *i*. Only significant codependence coefficients were considered. This value will equal 1 if x_j is the variable that shows the maximum codependence coefficient (in absolute value) and will decrease toward 0 as the difference between the codependence coefficient and the maximum becomes larger. A variable is included in the uncertainty set if its uncertainty coefficient is equal to or higher than some predefined threshold *k*. Viewed alternatively, this is the same as stating that

$$\frac{\max\left(\left|C_{y,x|u_i}\right|\right) - \left|C_{y,x|u_i}\right|}{\max\left(\left|C_{y,x|u_i}\right|\right)} \leq k, \tag{9}$$

so that the codependence coefficient for the considered variable is at most k standardized difference units away from the highest codependence coefficient for structuring variable i. On one extreme, setting $k = 0$ leaves no room for uncertainty, as only the variable with the highest absolute codependence coefficient will be selected. On the other extreme, setting $k = 1$ leaves too much room for uncertainty, as all significant variables are selected (i.e., even those that would have very low absolute values of codependence coefficients). Between these two extremes lies an appropriate and adequate amount of uncertainty for researchers. There is obviously some arbitrariness associated with this threshold parameter k. However, we prefer to see it as a flexibility that studies should include. For our study, we considered a threshold value of $k = 0.10$ meaning that any variable whose uncertainty coefficient was at most 0.10 standardized units away from the highest standardized codependence coefficient could be included.

Computational tools

The majority of computations were done in R 3.4.1 (Ihaka and Gentleman 1996, R Core Team 2017). Various functions from the packages vegan (Oksanen et al. 2012), WaveletComp (Rösch and Schmidbauer 2014*b*) and codep (Guénard and Legendre 2017*b*) were used in order to carry out the analyses. In order to speed up computations, the majority of analyses were carried out by writing and integrating compiled C++ code into R using the Rcpp (Eddelbuettel and Francois 2011, Eddelbuettel 2014) package. The Armadillo (Sanderson and Curtin 2016) C++ linear algebra library was used through the RcppArmadillo (Eddelbuettel and Sanderson 2014) package. Parallel computing and use of multiple computer cores was enabled using the packages snow (Tierney et al. 2007, 2009) and snowfall (Knaus et al. 2009) were used.

Results

Study site environmental conditions

The data analyzed begins at 10:55 hours on 18 June 2014 and ends at 07:12 hours on 25 April 2016. Sensor maintenance and cleaning produced occasional gaps of about two hours approximately every two weeks. Likewise, events such as power outages, water shut off, and dam maintenance produced additional gaps of variable size. Stating each of these gaps would be unnecessary as they were usually of short length (usually a couple of days). However, two gaps are of noticeable length and worth pointing out. The first of these was between 7 November 2014 and 13 February 2015, accounting for 98 d due to the New York Power Authority repairing the generating unit. The second of these was between 24 April 2015 and 10 June 2015, and accounted for 48 d; this was due to sensor instrumentation warranty repair.

Average environmental conditions were within the range of values expected (Table 1). However, individual variables showed different trends and patterns across time. Colored dissolved organic matter values were usually highest in spring and lowest in fall. Chlorophyll *a* was highest in summer and lowest in winter. However, we note that in 2016, two distinct peaks occurred, with one in summer and one in fall. These two peaks were much more pronounced than the peak seen in 2015. In vivo-based fluorescence of phycocyanin showed a pattern similar to in vivo chlorophyll *a* fluorescence and was highest in summer and fall and lowest in winter, but without showing two peaks in 2016. Temperature was highest in spring and summer, and lowest in fall and winter (≈0.1°C), as expected.

Table 1. Summary results of the measured environmental variables.

Variables	Mean	Median	Minimum	Maximum	Standard deviation
Colored dissolved organic matter (mg/L)	5.35	4.92	3.56	10.73	1.41
In vivo chlorophyll *a* (μg/L)	0.18	0.17	0.01	0.56	0.11
In vivo phycocyanin (mg/L)	0.01	0.01	0	0.07	0.01
Temperature (°C)	12.35	13.5	0.11	23.49	8.3

Wavelet transform analysis

Contour plots of the wavelet transform coefficients showed that all variables showed various levels of periodicities and at varying moments (Fig. 3). Colored dissolved organic matter, phycocyanin, and water temperature showed an extremely pronounced peak of average wavelet coefficient at a period of about 365 d, which would indicate strong yearly patterns in the data. For chlorophyll *a*, an additional distinct peak with a period of about 128 d was detected, indicating a quarterly pattern. However, inspection of the contour plot revealed that this pattern was not consistent throughout the study period and started only around the beginning of 2015. Chlorophyll *a* also showed a slight peak for daily periodicity which was strongest in summer and weakest in winter. The reason for this could be due to the weakening of daily variations in winter and the linear interpolation that was necessary to account for missing values that occurred primarily during winter months.

The four null models differed in the periods they detected as significant (see Fig. 3). Indeed, the white noise model was the most liberal and showed the widest range for periods, often indicating a large range of periods as being significant. Significant periods were noted as being between 16 and 1024 d. On the other hand, models that did not consider independence between points (i.e., red noise, pink noise, and estimated noise) were more reserved and only detected smaller patches of periods as significant. It is only for water temperature that the three latter models agree and identified only the yearly period as being significant. For CDOM, chlorophyll *a* and phycocyanin, null models with redder noise tended to show smaller periods as being significant, even though they showed lower average wavelet power than the larger periods. This could be since high (close to or higher than $\alpha = 2$) values of coefficients for the colored noise produce an extremely smooth sine wave. Therefore, the test might be more sensitive to small-scale variation in the data. However, we note that longer datasets might be necessary to test for higher-scale patterns in the data for some variables. Indeed, as can be seen for chlorophyll a, the yearly patterns are significant under the estimated noise model, but only within the cone of influence.

Moran eigenvector maps analysis

Temporal structures were identified for all three variables (CDOM, chlorophyll *a*, and phycocyanin), as the null model showed the high values of AIC_c that were very different than those of models that incorporated temporal variables (Table 2). For CDOM, the best temporal model based on AIC_c considered a DNN connectivity matrix with a concave-down weighing function and a parameter of $\beta = 1$. However, for chlorophyll a, the best temporal model considered a MST connectivity matrix with a linear weighing function.

After the null model, the worst-fitting models were usually those that considered a distance criterion and a binary weighing matrix. This result is understandable as, given the somewhat large cutoff value (98.125 d, which is the size of the largest mentioned gap), not weighing these distances considers giving them all the same importance or considering that the transfer of organisms or matter is very easy. Therefore, when considering connection schemes more connected than MST, then links between observations should be properly weighed. Even though there exists no objective scale of AIC_c interpretation, differences between AIC_c values were very large between classes of models. Indeed, based on suggested scales for ΔAIC (Burnham and Anderson 2002), the differences in AIC_c were indicative of no support for alternative models other than the best model.

We note that the forward selection procedure selected a lot fewer dbMEMs than the AIC_c method, around 20 times less. However, this difference is not due to the two criteria discussed by Blanchet et al. (2008). Rather, in these cases the algorithm terminated due to the R^2 more criterion. Had this criterion not been enforced, the forward selection procedure would have gone on for quite a while, adding significant variables, but which explain a very small amount of variation. Such a stopping criterion can easily be coded into an AIC selection algorithm for dbMEMs and doing so does make the two methods choose the same number of dbMEMs within each model. However, there was no clear relationship between such a prematurely terminated AIC_c and the conventional AIC_c.

When temporal models are not considered, environmental models explained close to half of

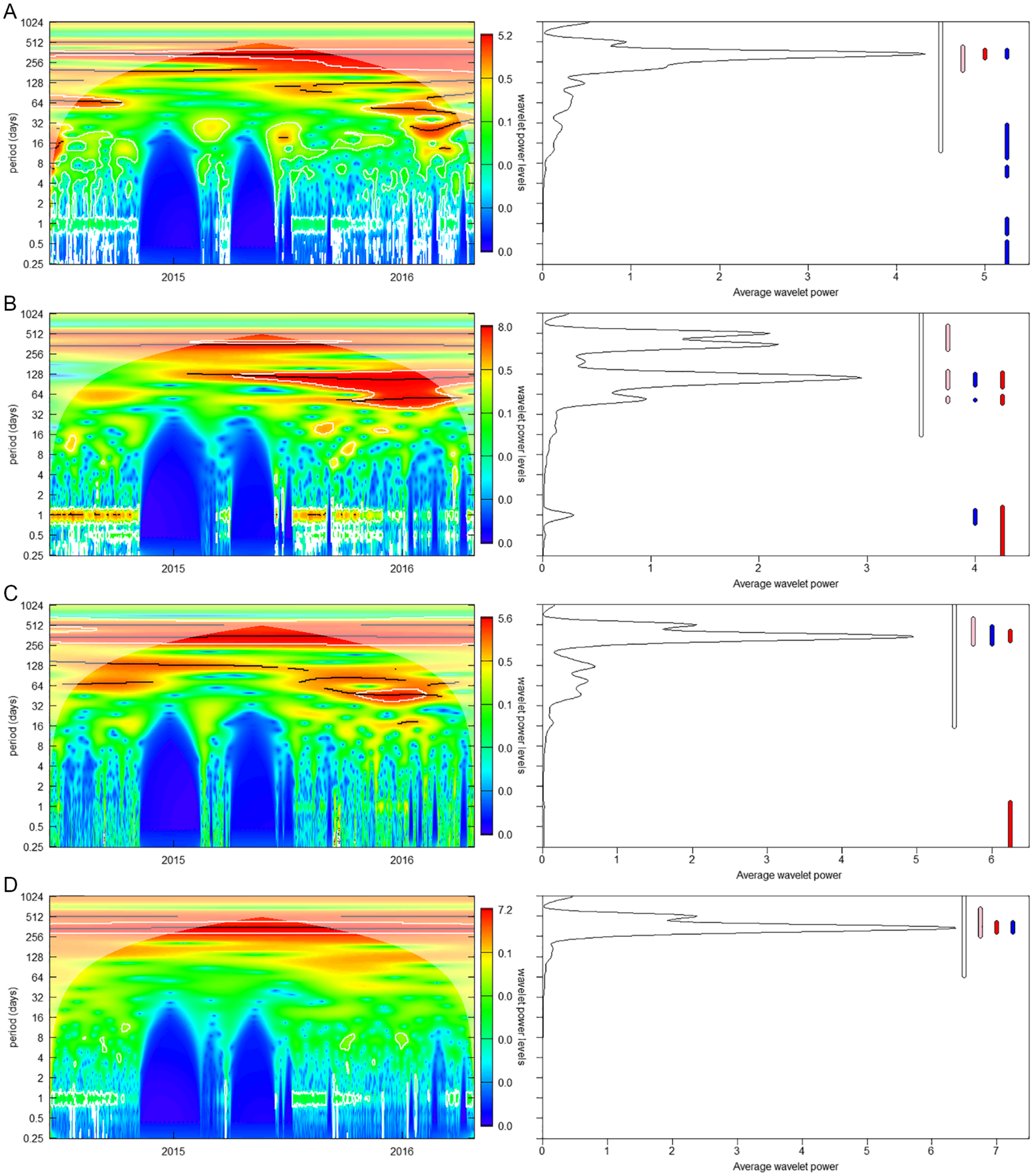

Fig. 3. Contour plots of the wavelet transform coefficients for colored dissolved organic matter (A), chlorophyll *a* (B), phycocyanin (C), and temperature (D). On the right of each contour plot, the wavelet power spectrum of the variable is shown; with white bars correspond to significant periods under the white noise model, red bars under the red noise model and pink bars under the pink noise model. The white-shaded area on the contour plots corresponds to the cone of influence or regions of the map where coefficients should not be interpreted due to margin effects.

Table 2. Results of the data-driven procedure specification of the spatial weighing matrix for the dbMEM analysis of the water quality data.

Connection	Weighing function	Number of variables	AIC$_c$
Colored dissolved organic matter			
NULL	NA	1	−9945.71
MST	BIN	1141	−33,007.84
	F1	1163	−33,262.85
	F2 (α = 10)	1122	−33,288.65
	F3 (β = 1)	1226	−30,177.62
DNN	BIN	42	−16,582.38
	F1	56	−22,894.45
	F2 (α = 2)	27	−17,470.62
	F3 (β = 1)†	826	−35,196.85
PCNM	PCNM	69	−16,922.91
Chlorophyll *a*			
NULL	NA	1	−14,157.2
MST	BIN	822	−31,648.4
	F1†	802	−31,963.86
	F2 (α = 2)	816	−31,928.9
	F3 (β = 1)	922	−31,051.79
DNN	BIN	37	−16,888.81
	F1	58	−24,396.61
	F2 (α = 2)	37	−16,628.78
	F3 (β = 1)	623	−31,372.66
PCNM	PCNM	68	−17,129.88

Notes: For models with variable parameter values (exponent α or β), only the best model values are reported. AIC$_c$, Akaike's information criterion, corrected for sample sizes; BIN, binary model; MST, minimum-spanning tree; PCNM, principal coordinates of neighboring matrices.

† The best spatial weighing matrix for each variable.

the response variables, as values of R_a^2 were equal to 47.10% and 32.50% for the variables CDOM and chlorophyll *a*, respectively. For CDOM, a small, albeit significant linear trend was detected and incorporated into models. Variation partitioning showed that the variation explained by the environmental model was, for both variables, almost completely shared with the temporal model and the linear trend (Fig. 4). The pure environmental models had a negligible fraction of individual explained variation, as the purely environmental fractions were <1.00% for both cases. In contrast to this, despite sharing a large fraction of variation with the environmental models, the purely temporal fraction of variation was still substantial.

Direct multiscale ordination analysis

As stated earlier, both RDA models were significant and explained at least half of the variation in the response variables. However, upon carrying out DMSO analysis and computing the variance profiles of both the canonical axes and the residual axes (Fig. 5), several issues with all three models were raised. Firstly, these variograms did not resemble any of the classical variogram models (e.g., Gaussian, exponential, or spherical models), since they did not show a clear sill at which the semi-variance function levels off. Instead, the semi-variance function decreases after reaching a peak value; this pattern was especially pronounced for CDOM, but less so for chlorophyll *a*. Such a periodic pattern is most likely due to the inherent yearly periodicity of the system.

The second result was that, even after correction for multiple tests, there was still some temporal autocorrelation in the residuals; this implies that the regression residuals of the linear model are not independent, thus violating one of the assumptions of the linear model. Likewise, it also indicates that the temporal autocorrelation in the considered variables was not enough to fully explain the temporal autocorrelation in these variables. In addition, it should be noted that the distance classes with significant

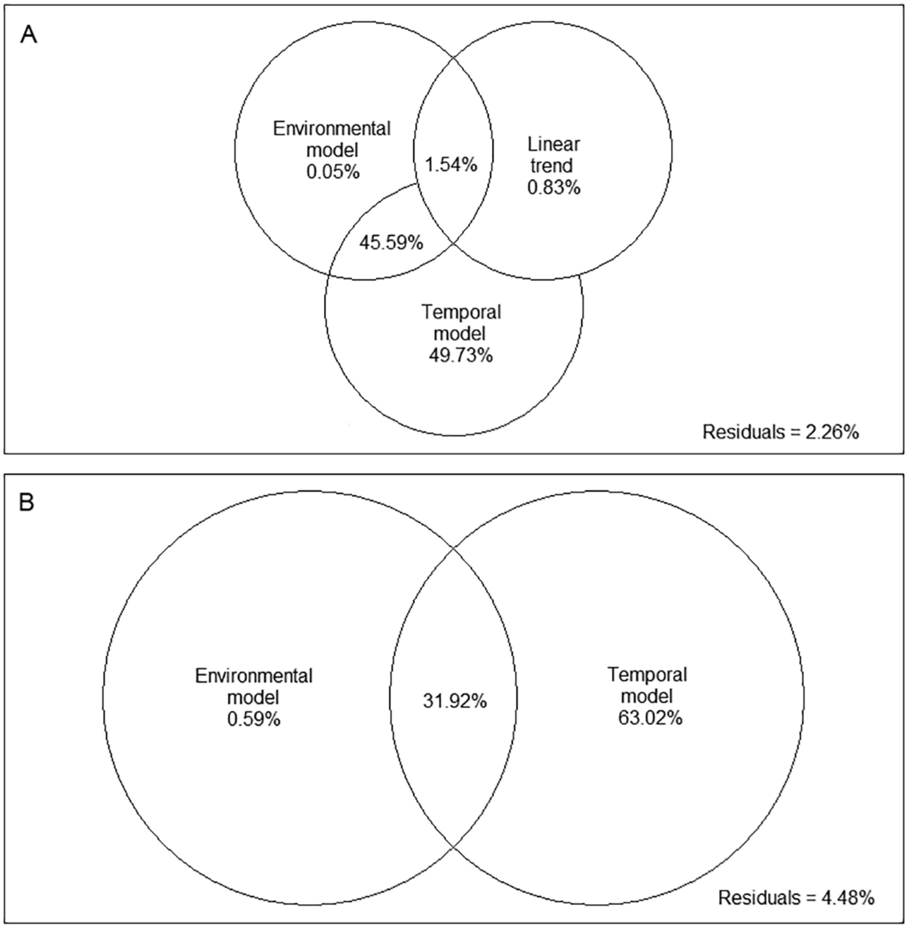

Fig. 4. Venn diagrams of hierarchical partitioning of colored dissolved organic matter (A) and chlorophyll *a* (B) between the environmental model, the temporal model based on Moran's eigenvector maps variables, and a linear trend, if necessary. When present, priority was given to the linear trend over the temporal model. Reported values are adjusted R^2 values.

temporal autocorrelation were not limited to the smallest classes. Finally, the residual variograms were usually not flat and tended to show a cyclic pattern like those of the response variable.

The third result was that the variogram of the explained variation often stepped out of the intervals derived from the variogram of total variance itself, which implies that the estimated coefficients were not consistent across scales. Consequently, computed linear models cannot be interpreted as global coefficients and that scale needs to be considered in these models.

Multiscale codependence analysis

Multiscale codependence analysis was significant and detected a different number of significant codependent structures for the environmental variables. Aside from the first few MEM variables, which model broadscale structures, codependence coefficients were somewhat low. To a certain extent, this was expected, as the coefficient is a product of two Pearson product–moment coefficients, which are bounded between [−1,1]. Likewise, other studies using MCA have also reported somewhat small values of codependence coefficients (Guénard et al. 2010). Nonetheless, it could also be that relationships between variables at scales other than the yearly scale are low.

The two newly described testing procedures (i.e., power-law noise simulation and block bootstrapping) found that the same broadscale

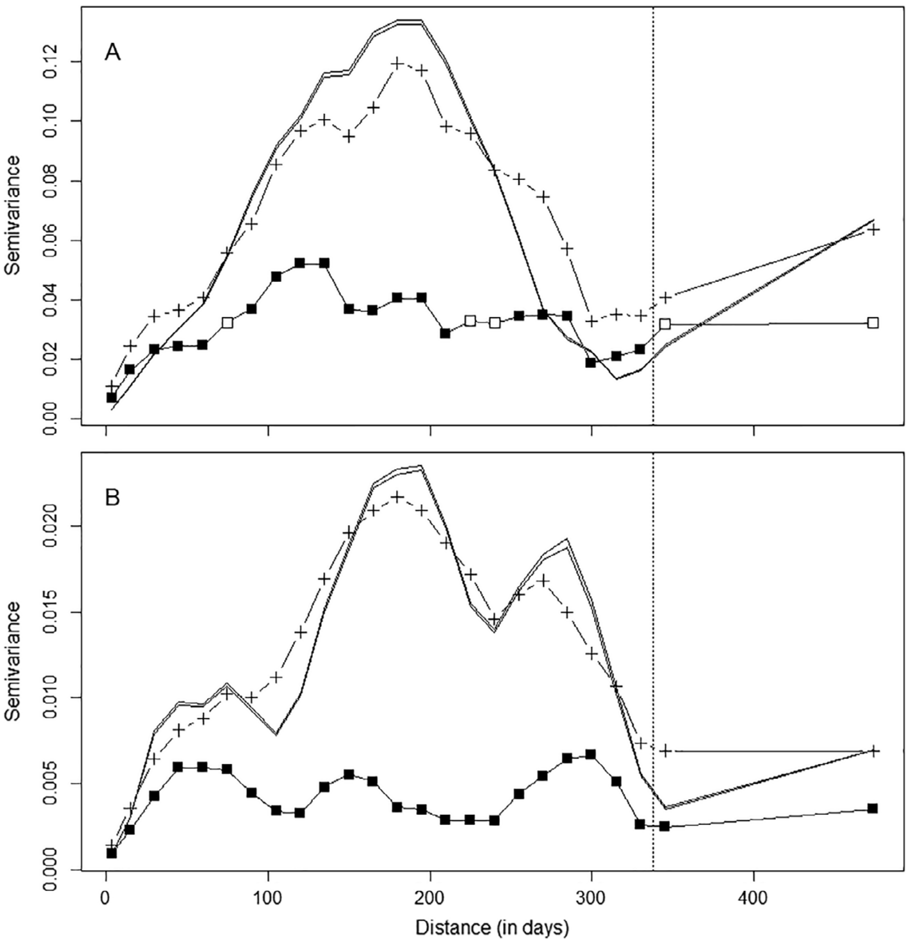

Fig. 5. Variance profiles for the residual (boxes) and the sum of the residual and explained (crosses) variances of the multiple regression models for colored dissolved organic matter (A) and chlorophyll *a* (B). Straight lines correspond to the point-wise envelope of the variogram of the total variance of the response variable. Significant temporal autocorrelations are indicated by black boxes.

codependence structures were significant. However, we found that both methods produced *P*-values that, even though they were still significant, were much higher than those by complete randomization. If the sequential Šidák correction were to be applied, then no significant codependence would be left. We also note that the results very similar to the complete randomization procedure were obtained under block bootstrapping when block size is equal to 1 and under power-law simulation when an exponent value of $\alpha = 0$ was considered (i.e., a white noise model). This was to be expected, as generating data under these models with these configurations creates variables without any temporal autocorrelation, which is close to the permutation procedure. The effect of increasing the exponent value of the power-law model was to make this distribution flatter, thus reducing the *P*-value associated with the test.

For CDOM, the strongest significant temporal structures were those that modeled large-scale, mostly yearly processes, such as MEM_2 and MEM_3 (Fig. 6). These MEM variables, respectively, accounted for 34.22% and 27.36% of the variation in CDOM. Had the most important structure been attributed to the variable with the highest τ -statistic, then it would have gone to phycocyanin. However, the uncertainty coefficient (see Eqs. 8 and 9) revealed that chlorophyll *a* was only 0.06 standardized units away from the value of phycocyanin. Therefore, it is not unreasonable to believe that these patterns are the result of either phycocyanin or chlorophyll *a* patterns. In addition, several MEMs that modeled smaller-scale structures were also noted as being significant. For chlorophyll *a*, selected MEMs were of higher order, but still represented large-scale patterns. The three most important structures were those associated with MEM_9, MEM_6, and MEM_{20}, which respectfully accounted for 44.03%, 10.00%, and 8.31% of the variation. Once again, MEM_{20} would have been attributed to CDOM patterns, but temperature was only 0.06 standardized units away from the value.

When relating CDOM and chlorophyll *a* to the set of MEMs related to each explanatory variables (Figs. 7, 8), differences between the explanatory variables could be seen. For chlorophyll *a*, CDOM was mostly responsible for smaller-scale patterns as the fitted values using the MEMs associated with CDOM had a coarser appearance and remained close to the mean,

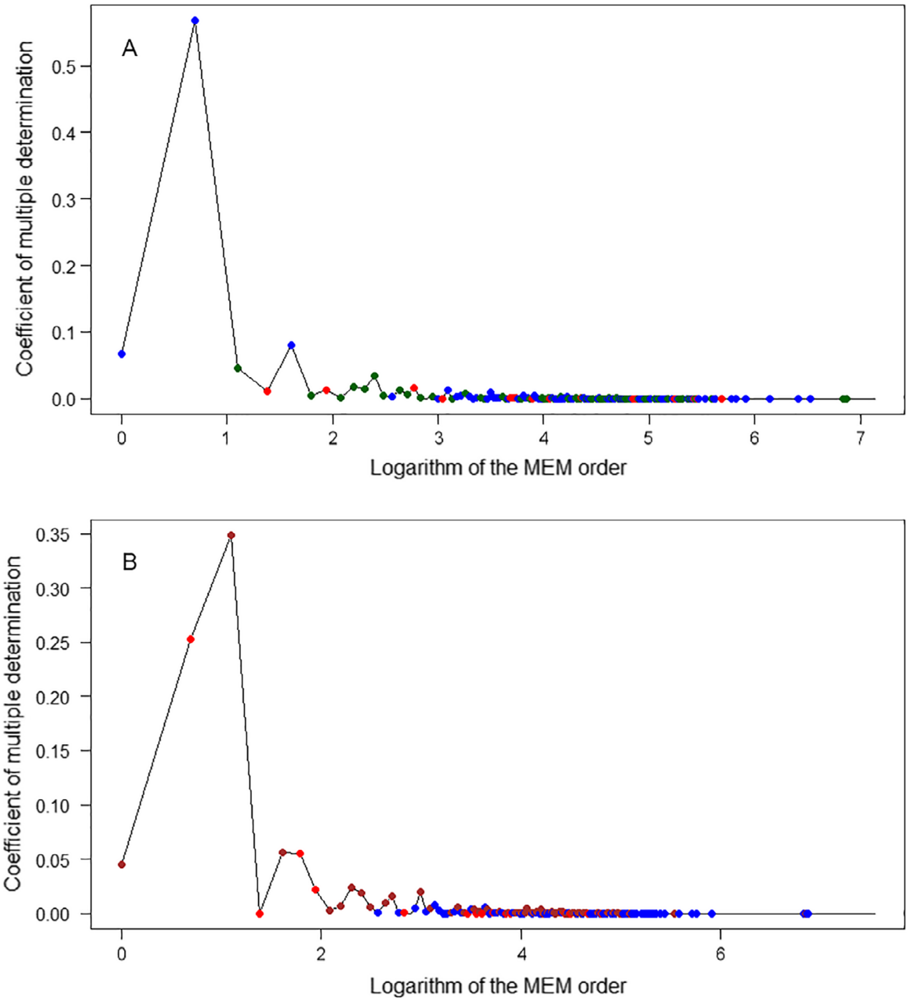

Fig. 6. Line plots showing the coefficients of determination of each Moran's eigenvector maps (MEM) with respect to the logarithm of the order of the MEM for a multiscale codependence analysis of colored dissolved organic matter (CDOM, A) and chlorophyll *a* (B). The values of nonsignificant codependence coefficients have been set to 0. Colored dots represent the variable to which the significant structure was attributed to and are brown for CDOM, green for chlorophyll *a*, blue for phycocyanin, and red for temperature.

accounting for a small portion of the variation. On the other hand, temperature was mostly responsible for large-scale patterns as the fitted values using the MEMs associated with temperature had a smoother appearance and often strayed far from the mean, accounting for a larger portion of the variation. Such a difference between the fitted values for different sets of MEMs related to explanatory variables was not as pronounced for CDOM. Indeed, aside from the fact that MEMs related to chlorophyll *a* did not explain a lot of the variation; MEMs associated with phycocyanin and temperature showed similar patterns.

Discussion

With the advent of high-performance sensors that autonomously record data, the capacity to accumulate environmental data has increased tremendously. In response, analytical tools that can provide direct and efficient appreciations of the dataset are invaluable. Wavelet analysis detected strong yearly patterns for almost variables except chlorophyll *a*. This was expected as these reflect seasonal changes in the environmental variables. However, in addition to this yearly pattern, another important periodic component with a period of about a quarter of a year was detected for chlorophyll *a*, starting around 2015. Such an observation seemingly contradicts classical limnological theory for temperate systems, which predicts that the main periodic component should be yearly. However, it is possible that this periodic component represents more of a localized event that occurred only within this timeframe, giving the appearance of a periodic signal, rather than hidden periodicity. Inspection of water samples taken from that point confirmed that this period was characterized by an important winter population of diatoms. Additionally, daily patterns were detected for chlorophyll *a*, even though these were more local and not globally significant.

The capacity to detect localized events implies that, as with its uses in epidemiological (Grenfell et al. 2001, Cazelles et al. 2013) and biological (James et al. 2010) cases, wavelet analysis could be used to identify ecological outbreaks of water quality variables such as chlorophyll *a*. Schmidt and Sutton (2018) examined the potential of wavelet coherence to show that certain limnological variables could be used to explain the dominant source of variability in observations. However, the mother wavelet should be chosen with care, as it represents the type of pattern over which the data are confronted and could have been changed (Bradshaw and Spies 1992, Mi et al. 2005). When testing for significant overall periods in the wavelet spectrum, null models which have some autocorrelation in their data (i.e., red noise, pink noise, and estimated noise) were found to be better suited at identifying peaks in the average wavelet power than those with completely independent data (i.e., white noise). A similar result was noted by James et al. (2010), who found that auto-correlated null models showed better distinction between significant patches of spruce budworm outbreaks across space. As time-series are inherently temporally auto-correlated objects, null models that do not disregard this aspect should be more appropriate as they incorporate the temporal non-independence of the data.

Chlorophyll *a* and CDOM fluorescence in the St. Lawrence River system had practically the entirety of the variation explained by the environmental model shared with the temporal model. Analogously to the interpretation of the shared fraction in variation in spatial analyses (Borcard et al. 1992, Borcard and Legendre 1994), this shared fraction can be referred to as temporally structured environment. Therefore, the conclusion is not that environment variation has no effect, but rather that this effect is extremely temporally structured. Moran's eigenvector maps variables can be used as covariables in linear models so as to give statistical tests with a correct type I error (Peres-Neto and Legendre 2010), but such an approach would be meaningless in this case, since after accounting for the temporal model, the environmental model accounts for virtually nothing. Therefore, understanding the patterns of this temporally structured environment should prove to be much more interesting than simply partialling out the effect of time.

Despite having appreciable R_a^2 values, environmental models that ignored the temporal aspects of the data (i.e., that did not include MEMs or the linear trend) did not respect several of the assumptions of linear models. Regression coefficients were not consistent across scales, implying that a global model would be inadequate.

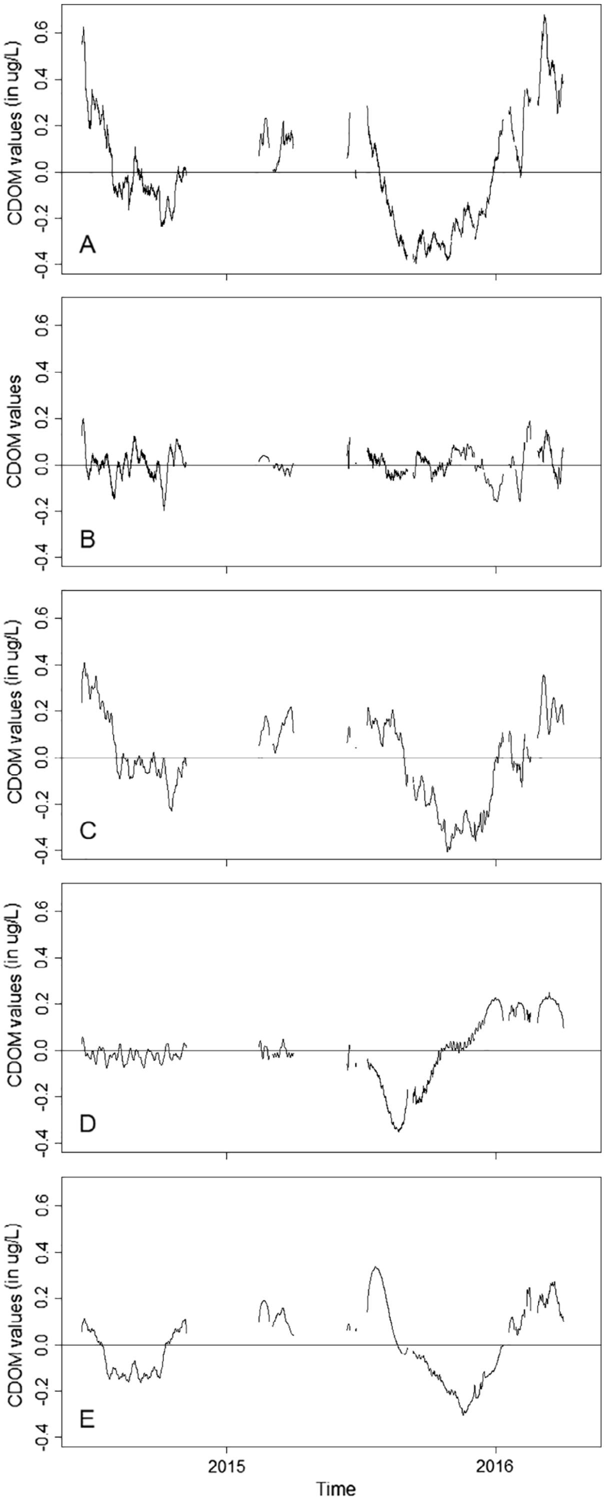

Fig. 7. Line plot of centered and detrended colored dissolved organic matter values on a loge scale (A) and fitted values according with respect to structuring variables associated with chlorophyll a (B), phycocyanin (C) and water temperature (D).

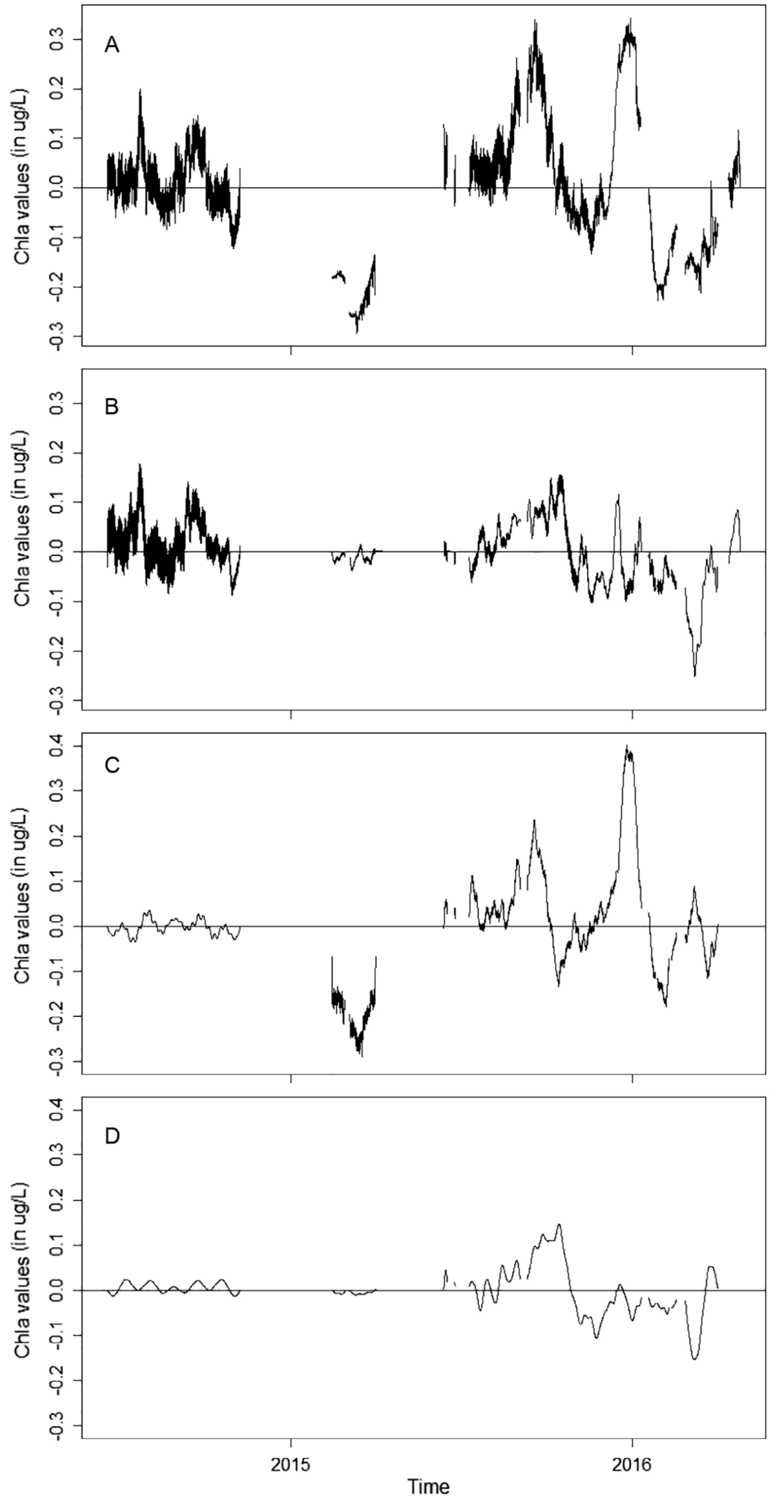

Fig. 8. Line plot of centered and detrended chlorophyll a values on a square-root scale (A) and fitted values according with respect to structuring variables associated with colored dissolved organic matter (B), phycocyanin (C) and water temperature (D).

Therefore, the use of scale-dependent methods becomes a necessity. The presence of temporal autocorrelation in the residuals at various distance classes could indicate that our models are missing a temporally auto-correlated variable, which would be necessary to produce temporally independent residuals. Possible candidates for this missing variable include weather variables or nutrient availability. We note that another alternative might be to use a generalized least squares model (Aitken 1934, Amemiya 1985) that allows for the specification of how the data are correlated. In this case, it might be worthwhile to experiment with autoregressive or autoregressive and moving average correlation structures and seek to integrate those into a generalized least squares framework. As with testing of fractions, MEM variables can be used as covariables to control for the spatial structure in the response-explanatory variable relationship in DMSO (Legendre and Legendre 2012, Borcard et al. 2016). However, this approach was not adopted since, as stated earlier, the resulting model would explain almost nothing.

It appears that chlorophyll *a* values in the St. Lawrence are the result of variables acting at several scales. Large-scale patterns, such as those over months and years, appear to be related to changes in water temperature. Smaller-scale patterns, such as those over weeks and days, appear to be related to CDOM concentrations. However, some events, such as the diatom bloom in late fall 2015, require further inspection, as at that point in time, water temperature was quite low (around 3°C) and it is likely that this bloom was driven by other factors such as nutrient availability.

It is possible that the response variables show significant codependence with more than one variable at a single scale. Such a situation was envisioned by the authors of MCA who posit retaining the largest absolute statistic in order to preserve strictly exclusive sets of structuring variables (Guénard et al. 2010, Guénard and Legendre 2017*a*). We agree with this procedure, but suggest that some form of uncertainty criterion, such as the newly described uncertainty coefficient should be considered in order to assess whether the selected variable is truly the only driving factor at said scale and can avoid potentially misleading conclusions. For example, the coefficient of determination for CDOM for dbMEM$_3$, which describes seasonal patterns, indicates that this is driven by patterns of phycocyanin, which showed the highest codependence coefficient. However, this could be the result of a positive association with either phycocyanin or chlorophyll *a*, as both variables are significantly associated with CDOM at this scale and show codependence coefficient values separated by ~0.06 standardized units. Consequently, if MCA is used as a tool to determine the identity of the driving variables at various scales, then the possibility of several variables affecting the response variable should be considered and assessed.

Changing the null testing method gave results that were somewhat different that the complete randomization method. Both alternative methods gave results showing that the first codependence was significant, just as the complete randomization method. However, the *P*-values reported were much higher, meaning that, had the sequential Šidák correction been applied, then no significant codependence would be left. The block bootstrap method showed intermediate results, due to the influence of block size. We find that increasing block size leads to an increase in the *P*-value. However, block size should be determined by the question at hand, as bootstrapping blocks below the scale of the structure would be too similar to bootstrapping individual values; and bootstrapping blocks above the scale of the structure would recreate too many of the same structures. Such methods require further study of their assumptions and sensitivity but might prove to be helpful in cases where complete randomization of data represents a poor null model, creating datasets which are too poorly structured to provide a reasonable background for time-series testing.

Despite the systematic and consistent recording potential of electronic sensors, some data gaps are unavoidable. Some of these gaps can be planned, such as sensor maintenance, whereas others, such as power outages or dam repairs, cannot. To some extent, these gaps need to be addressed for certain numerical methods to function adequately. In our case, missing values in the wavelet analysis and for the block bootstrap null model were obtained using linear interpolation. Other approaches may be considered; for example, Serdar (2011) filled gaps in water level sensor data with a neural network algorithm.

One of the strong points for methods that use MEMs is that they do not require equally spaced data which show no gaps. Consequently, MEMs provide a toolbox to test various types of structures in the data. However, the null model generating methods discussed earlier require some consideration of gaps for them to generate data suitable to their use.

We identify two future directions that should guide future work and research. First, much of the methods considered here are mostly used to predict rather than to forecast time-series (Legendre and Legendre 2012). This means that we gain an understanding of what caused previous values, but not necessarily adequate tools to predict future values. As much of the aim of environmental monitoring is to provide recommendations for future plans, this capacity should be essential. Second, the manner of testing for significance at various scales was approached in this paper by considering different null hypotheses. However, as with any procedure, it is only as valid as its assumptions. Therefore, more research should be directed toward how to make a test's assumptions as clear as possible so that they can be modified in order to represent that which the ecologist desires to verify.

A lot of effort has been put into the effects of climate change and its effect on the management of resources in the St. Lawrence River system (Mortsch et al. 2000, Millerd 2005). However, water quality in the St. Lawrence River system appears to be the result of various variables operating at several scales. The main source of variation appears to be seasonal variations associated with variations in water temperature and CDOM values. At the same time, patterns of a smaller scale are also significant and can appreciable in terms of magnitude, implying that there is a need to consider them as well. Therefore, forecast models and action plans should be changed to incorporate this scale dependence. In order to confront forecasts with actual data, long-term data collection programs in combination with multiscale analysis methods will be crucial to evaluate ecosystem changes.

Acknowledgments

This study was supported by funds from the Ontario Ministry of the Environment and Climate Change, the Five B Family Foundation, and the National Research Council Canada—Industrial Research Assistance Program for a research fellowship for Dr. El-Amine Mimouni. Funding was also provided by the New York State Water Research Institute, New York's Great Lakes Basin Small Grants Program (New York Sea Grant), and the Beacon Institute for Rivers and Estuaries (Clarkson University). We thank the New York Power Authority and Jonathan Mayette (NYPA) for providing access to the hydropower dam, Sean Doyle (NYPA) for assistance with permitting, and Faith C. Neff for assistance setting up the sensor array (funded by the National Science Foundation Research Experience for Undergraduate Program EEC-1359256). We thank Katherine Moir for providing the map of the study region.

Literature Cited

Addison, P., J. Watson, and T. Feng. 2002. Low-oscillation complex wavelets. Journal of Sound and Vibration 254:733–762.

Aguiar-Conraria, L. and M. Soares. 2011. The continuous wavelet transform: a primer. NIPE working paper series. NIPE, Universidade do Minho, Braga, Portugal.

Aitken, A. C. 1934. On least-squares and linear combinations of observations. Proceedings of the Royal Society of Edinburgh 55:42–48.

Akaike, H. 1974. A new look at the statistical model identification. IEEE Transactions on Automatic Control 19:716–723.

Amemiya, T. 1985. Advanced econometrics. Harvard University Press, Cambridge, Massachusetts, USA.

Arora, B., D. Dwivedi, S. S. Hubbard, C. I. Steefel, and K. H. Williams. 2016. Identifying geochemical hot moments and their controls on a contaminated river floodplain system using wavelet and entropy approaches. Environmental Modelling & Software 85:27–41.

Ball, E. E., D. E. Smith, E. J. Anderson, J. D. Skufca, and M. R. Twiss. 2018. Water velocity modeling can delineate nearshore and main channel plankton environments in a large river. Hydrobiologia 815:125–140.

Blanchet, F. G., P. Legendre, and D. Borcard. 2008. Forward selection of explanatory variables. Ecology 89:2623–2632.

Borcard, D., and P. Legendre. 1994. Environmental control and spatial structure in ecological communities: an example using oribatid mites (Acari, Oribatei). Environmental and Ecological Statistics 1:37–61.

Borcard, D., and P. Legendre. 2002. All-scale spatial analysis of ecological data by means of principal coordinates of neighbour matrices. Ecological Modelling 153:51–68.

Borcard, D., P. Legendre, and P. Drapeau. 1992. Partialling out the spatial component of ecological variation. Ecology 73:1045–1055.

Borcard, D., P. Legendre, and F. Gillet. 2016. Numerical ecology with R. Springer, Dordrecht, The Netherlands.

Bradshaw, G., and T. Spies. 1992. Characterizing canopy gap structure in forests using wavelet analysis. Journal of Ecology 80:205–215.

Burnham, K. P. and D. R. Anderson. 2002. Model selection and multimodel inference: A practical information-theoretic approach. Springer, Dordrecht, The Netherlands.

Carlstein, E. 1986. The use of subseries values for estimating the variance of a general statistic from a stationary sequence. Annals of Statistic 14: 1171–1179.

Cazelles, B., K. Cazelles, and M. Chavez. 2013. Wavelet analysis in ecology and epidemiology: impact of statistical tests. Journal of the Royal Society Interface 11:1–10.

Cazelles, B., M. Chavez, D. Berteaux, F. Ménard, J. O. Vik, S. Jenouvrier, and N. C. Stenseth. 2008. Wavelet analysis of ecological time series. Oecologia 156:287–304.

Dale, M., and M. Mah. 1998. The use of wavelets for spatial pattern analysis in ecology. Journal of Vegetation Science 9:805–814.

De Cáceres, M., P. Legendre, and D. Borcard. 2010. Community surveys through space and time: testing the space-time interaction in the absence of replication. Ecology 91:262–272.

Delaunay, B. 1934. Sur la sphère vide. Bulletin de l'Académie des Sciences de l'URSS - Classe des sciences mathématiques et naturelles 6:793–800.

Dray, S., P. Legendre, and P. R. Peres-Neto. 2006. Spatial modelling: a comprehensive framework for principal coordinate analysis of neighbour matrices (PCNM). Ecological Modelling 6:483–493.

Dungan, J. L., J. N. Perry, M. R. T. Dale, P. Legendre, M. Fortin, A. Jakomulska, M. Miriti, M. S. Rosenberg, and S. Fortin. 2002. A balanced view of scale in spatial statistical analysis. Ecography 25:626–640.

Eddelbuettel, D. 2014. Seamless R and C++ Integration with Rcpp. Springer, New York, New York, USA.

Eddelbuettel, D., and R. Francois. 2011. Rcpp: Seamless R and C++ Integration. Journal of Statistical Software 40:1–18.

Eddelbuettel, D., and C. Sanderson. 2014. RcppArmadillo: accelerating R with high-performance C++ linear algebra. Computational Statistics and Data Analysis 71:1054–1063.

Efron, B., and R. J. Tibshirani. 1993. An introduction to the bootstrap. Chapman & Hall, New York, New York, USA.

Ezekiel, M. 1930. Methods of correlational analysis. Wiley, New York, New York, USA.

Fortin, M.-J., and M. Dale. 2007. Spatial analysis A guide for ecologists. Cambridge University Press, Cambridge, UK.

Fortin, M.-J., and G. M. Jacquez. 2000. Randomization tests and spatially auto-correlated data. Bulletin of the Ecological Society of America 81:201–205.

Gabriel, K., and R. R. Sokal. 1969. A new statistical approach to geographic variation. Systematic Zoology 18:259–278.

Getis, A., and J. Aldstadt. 2004. Constructing the spatial weights matrix using a local statistic. Geographical Analysis 36:90–104.

Godínez-Domínguez, E., and J. Freire. 2003. Information-theoretic approach for selection of spatial and temporal models of community organization. Marine Ecology Progress Series 253:17–24.

Grenfell, B. T., O. N. Bjornstad, and J. Kappey. 2001. Travelling waves and spatial hierarchies in measles epidemics. Nature 414:716–723.

Griffith, D., and P. R. Peres-Neto. 2006. Spatial modeling in ecology the flexibility of eigenfunction spatial analyses. Ecology 87:2603–2613.

Guénard, G., and P. Legendre. 2017*a*. Bringing multivariate support to multiscale codependence analysis: assessing the drivers of community structure across spatial scales. Methods in Ecology and Evolution 9:292-304.

Guénard, G. and P. Legendre. 2017*b*. codep: multiscale Codependence Analysis. R package version 0.6-5. https://CRAN.R-project.org/package=codep

Guénard, G., P. Legendre, D. Boisclair, and M. Bilodeau. 2010. Multiscale codependence analysis: an integrated approach to analyze relationships across scales. Ecology 91:2952–2964.

He, F., P. Legendre, and J. V. LaFrankie. 1996. Spatial pattern of diversity in a tropical rain forest of Malaysia. Journal of Animal Ecology 23:57–74.

Hurvich, C., and C. Tsai. 1989. Regression and time series model selection in small samples. Biometrika 76:297–307.

Ihaka, R., and R. Gentleman. 1996. R: a language for data analysis and graphics. Journal of Computational and Graphical Statistics 5:299–314.

James, P. M. A., R. A. Fleming, and M.-J. Fortin. 2010. Identifying significant scale-specific spatial boundaries using wavelets and null models: spruce budworm defoliation in Ontario, Canada as a case study. Landscape Ecology 25:873–887.

Keitt, T., and D. Urban. 2005. Scale-specific inference using wavelets. Ecology 86:2497–2504.

Kirchner, J. W., X. Feng, and C. Neal. 2000. Fractal stream chemistry and its implications for contaminant transport in catchments. Nature 403:524–527.

Kirchner, J. W., X. Feng, and C. Neal. 2001. Catchment-scale advection and dispersion as a mechanism for fractal scaling in stream tracer concentrations. Journal of Hydrology 254:82-101.

Knaus, J., C. Porzelius, and H. Binder. 2009. Easier parallel computing in R with snowfall and sfCluster. R Journal 1:54–59.

Künsh, H. 1989. The jackknife and the bootstrap for general stationary observations. Annals of Statistics 17:1217–1241.

Legendre, P., D. Borcard, and D. W. Roberts. 2012. Variation partitioning involving orthogonal spatial eigenfunction submodels. Ecology 93:1234–1240.

Legendre, P., and O. Gauthier. 2014. Statistical methods for temporal and space-time analysis of community composition data. Proceedings of the Royal Society B: Biological Sciences 281:1-10.

Legendre, P., and L. Legendre. 2012. Numerical ecology. Elsevier, New York, New York, USA.

Legendre, P., X. Mi, H. Ren, K. Ma, M. Yu, I. F. Sun, and F. He. 2009. Partitioning beta diversity in a subtropical broad-leaved forest of China. Ecology 90:663–674.

Levin, S. A. 1992. The problem of pattern and scale in ecology. Ecology 73:1943–1967.

Liu, Y., X. San Lian, and R. H. Weisberg. 2007. Rectification of the bias in the wavelet power spectrum. Journal of Atmospheric and Oceanic Technology 24:2093–2102.

Mantel, N. 1967. The detection of disease clustering and a generalized regression approach. Cancer Research 27:209–220.

Marvin, S. 1982. 1/f noise. Proceedings of the IEEE 70:212–218.

Mi, X., H. Ren, Z. Ouyang, W. Wei, and K. Ma. 2005. The use of the Mexican Hat and the Morlet wavelets for detection of ecological patterns. Plant Ecology 179:1–19.

Millerd, F. 2005. The economic impact of climate change on Canadian commercial navigation on the Great Lakes. Canadian Water Resources Journal 30:269–280.

Morlet, J., G. Arenss, E. Fourgeau, and D. Giard. 1982*a*. Wave propagation and sampling theory - Part I: complex signal and scattering in multilayered media. Geophysics 47:203–221.

Morlet, J., G. Arenss, E. Fourgeau, and D. Giard. 1982*b*. Wave propagation and sampling theory - Part II: sampling theory and complex waves. Geophysics 47:222–236.

Mortsch, L., H. Hengeveld, M. Lister, L. Wenger, F. Quinn, M. Slivitzky, L. Mortsch, H. Hengeveld, M. Lister, and L. Wenger. 2000. Climate change impacts on the hydrology of the Great Lakes-St. Lawrence system. Canadian Water Resources Journal 25:153–179.

Oksanen, J., F. Blanchet, R. Kindt, P. Legendre, P. Minchin, R. O'Hara, G. Simpson, P. Sólymos, and M. Stevens. 2012. vegan: community Ecology Package. https://CRAN.R-project.org/package=vegan

Parr, T. W., A. R. J. Sier, R. W. Battarbee, A. Mackay, and J. Burgess. 2003. Detecting environmental change: science and society—perspectives on long-term research and monitoring in the 21st, century. Science of the Total Environment 310:1–8.

Percival, D. B., M. Wang, and J. E. Overland. 2004. An introduction to wavelet analysis with applications to vegetation time series. Community Ecology 5:19–30.

Peres-Neto, P. R., and P. Legendre. 2010. Estimating and controlling for spatial structure in the study of. Global Ecology and Biogeography 19:174–184.

Peres-Neto, P. R., P. Legendre, S. Dray, and D. Borcard. 2006. Variation partitioning of species data matrices: estimation and comparison of fractions. Ecology 87:2614–2625.

R Core Team. 2017. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Rao, C. R. 1964. The use and interpretation of principal component analysis in applied research. Sankhyā: The Indian Journal of Statistics, Series A 26:329–358.

Rösch, A. and H. Schmidbauer. 2014*a*. WaveletComp: a guided tour through the R-package. https://www.hs-stat.com/projects/WaveletComp/WaveletComp_guided_tour.pdf

Rösch, A., and H. Schmidbauer. 2014*b*. WaveletComp: computational Wavelet Analysis. R package version 1.0. https://CRAN.R-project.org/package=WaveletComp

Sanderson, C., and R. Curtin. 2016. Armadillo: a template-based C++ library for linear algebra. Journal of Open Source Software 1:1–26.

Schmidt B. E., and J. A. Sutton. 2018. High-resolution velocimetry from tracer particle fields using wavelet-based optical flow. American Institute of Aeronautics and Astronautics 2018:1767.

Serdar, E. 2011. Time-frequency analyses of tide-gauge sensor data. Sensors 11:3939–3961.

Šidák, Z. 1967. Rectangular confidence regions for the means of multivariate normal distributions. Journal of the American Statistical Association 62:626–633.

Sugiura, N. 1978. Further analysis of the data by Akaike's information criterion and the finite corrections.

Communications in Statistics - Theory and Methods 7:13–26.

ter Braak, C. 1986. Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis. Ecology 67:1167–1179.

Thrush, S. F., et al. 1997. Scaling-up from experiments to complex ecological systems: Where to next? Journal of Experimental Marine Biology 216:243–254.

Tierney, L., A. Rossini, and N. Li. 2007. Simple parallel statistical computing in R. Journal of Computational and Graphical Statistics 16:399–2007.

Tierney, L., A. Rossini, and N. Li. 2009. snow: A parallel computing framework for the R system. International Journal of Parallel Programming 37:78–90.

Timmer, J., and M. König. 1995. On generating power law noise. Astronomy and Astrophysics 300:707–710.

Torrence, C., and P. J. Webster. 1999. Interdecadal changes in the ENSO – Monsoon System. Journal of Climate 12:2679–2690.

Toussaint, G. T. 1980. The relative neighbourhood graph of a finite planar set. Pattern Recognition 12:261–268.

Twiss, M. R., and K. M. Stryszowska. 2016. State of emerging technologies for assessing aquatic condition in the Great Lakes-St. Lawrence River system. Journal of Great Lakes Research 42:1470–1477.

Várbíró, G., J. Padisák, Z. Nagy-László, A. Abonyi, I. Stanković, M. Gligora Udovič, V. B-Béres, and G. Borics. 2018. How length of light exposure shapes the development of riverine algal biomass in temperate rivers? Hydrobiologia 809:53–63.

Wagner, H. H. 2003. Spatial covariance in plant communities integrating ordination, geostatistics, and variance testing. Ecology 84:1045–1057.

Wagner, H. H. 2004. Direct multi-scale ordination with canonical correspondence analysis. Ecology 85:342–351.

Wiens, J. A. 1989. Spatial scaling in ecology. Functional Ecology 3:385–397.

Wright, S. 1992. Adjusted p-values for simultaneous inference. Biometrics 48:1005–1013.