Challenges in modeling detailed and complex environmental data sets: a case study modeling the excess partial pressure of fluvial CO2

Noname manuscript No.
(will be inserted by the editor)

Challenges in Modeling Detailed and Complex Environmental Data Sets: A Case Study Modeling the Excess Partial Pressure of Fluvial CO_2

Amira Elayouty · Marian Scott · Claire Miller · Susan Waldron · Maria Franco-Villoria

the date of receipt and acceptance should be inserted later

Abstract Advances in sensor technology enable monitoring programmes to record and store measurements at a high temporal resolution, enhancing the capacity to detect and understand short duration changes that would not have been apparent in the past with monthly, fortnightly or even daily sampling. However, there are challenges in the processing and analysis of these high-frequency data such as their complex behavior over the different timescales and the strong correlation structure that persists over a large number of lags. Here, we explore the complexities of modeling high-frequency data which arise from environmental applications. With increasing understanding of the importance of surface waters as a source of atmospheric CO_2 we consider a high resolution sensor-generated time series of the over-saturation of CO_2, $EpCO_2$, in a small order river system. We will discuss advanced statistical approaches used to analyze and model the data, which include visualization tools for exploratory analysis, wavelets and generalized additive models. These methods reveal the complex dynamics of $EpCO_2$ over different timescales, and the multi-variate relationships of $EpCO_2$ with hydrology and temporal auto-correlation structures, which are time and scale dependent.

Keywords High-Frequency Data · Wavelets · Generalized Additive Models · Excess Partial Pressure of Carbon Dioxide

1 Introduction

Understanding the drivers of crucial environmental challenges, such as climate change, water and air pollution, changes in water quantity, and loss of soil carbon is of great im-

A. Elayouty
School of Mathematics and Statistics, University of Glasgow, Scotland
E-mail: a.el-ayouti.1@reseasrch.gla.ac.uk

M. Scott and C.Miller
School of Mathematics and Statistics, University of Glasgow, Scotland

S. Waldron
School of Geographical and Earth Sciences, University of Glasgow, Scotland

M. Franco-Villoria
Department of Economics and Statistics, University of Torino, Italy

portance to society [8]. Therefore, environmental monitoring technologies are continually being developed to enhance the ability to understand environmental systems and detect changes occurring within these systems. In the past, monitoring programmes typically involved monthly, fortnightly, weekly, and occasionally daily sampling campaigns but rarely shorter time intervals ([9], [15]). However, many changes in stream water quality happen at sub-daily scales [12] and hence monitoring programmes of high temporal resolution are needed to observe and understand the significance of these rapid changes. Sensor technology is continuously developing and as a result, the ability to record and store measurements is ever-improving [25]. Accordingly, environmental monitoring programmes can use sensors that record hourly or sub-hourly (e.g., every minute) measurements [12]. Sensor data recorded at short time frames over a long time period are known as "High-Frequency Data (HFD)" [9]. Such HFD allow us to address new research questions which were previously inaccessible [9], but there are statistical modeling and analysis challenges. Many of the currently available statistical methods and software tools struggle to properly handle the complexity of such volumes of data [9] and hence new statistical methods are needed to analyze and model these large complex datasets.

Here, we investigate the complexities in the modeling and analysis of hydrological high-frequency data, such as persistent correlation between observations, complex dynamics and interactions over the different timescales. High resolution sensor-generated time series of partial pressure of carbon dioxide in a river is used as an illustrative dataset. The aqueous partial pressure of carbon dioxide is a measure of the capacity for CO_2 exchange between the water and the atmosphere [11]. The excess partial pressure of carbon dioxide ($EpCO_2$) in surface freshwater is a dynamic representation of the interacting biogeochemical and hydrological processes that produce, consume, and transport carbon dioxide [20]. If the river is over-saturated (an excess partial pressure, > 1), CO_2 can be effluxed, representing direct linkage of terrestrial and atmospheric carbon cycles [3]. As surface waters are capable of degassing large amounts of CO_2 to the atmosphere ([18], [4], [19], [10], [24]), they have been included recently in the assessment of the global carbon budget [3]. Therefore, understanding of the temporal variability in the capacity for degassing and the drivers of such variability is of value in refining uncertainity over such estimates.

Many studies have examined the temporal and spatial variations of $EpCO_2$ and the mechanisms controlling these variations in some high and low order rivers and large lotic systems ([3], [4], [5], [18], [19], [11], [10], [24]). All these studies show that high-order rivers are important sources of atmospheric CO_2, and that small streams contain high concentrations of dissolved CO_2 [3]. Large rivers are also over-saturated. For example, the Hudson River, which flows from north to south through eastern New York in the United States is over-saturated with CO_2 and $EpCO_2$ exhibits a diel cycle reaching its maximum in summer [18]. The evasion of CO_2 from rivers of the central Amazon basin constitutes an important carbon loss process and there is a pronounced seasonality in evasion linked to wet and dry seasons [19]. However, the estimates of the effluxed CO_2 are uncertain because of large temporal and spatial variability. $EpCO_2$ dynamics at six sites in the lower reaches of Xijiang River, southern China, were difficult to interpret due to sampling frequency - monthly - being insufficient [24]. Daily measurements have proved more useful in understanding spatio-temporal dynamics e.g. in the upper Yangtze River basin in China [11]. So high-frequency sampling in space and time is required due to the spatio-temporal heterogeneity in the catchment characteristics and anthropogenic activities.

Therefore, sub-daily measurements across different seasonal periods should provide sufficient detail to understand fluctuations of free CO_2 concentrations at smaller timescales [5]. Here, three years of high-frequency sensor-generated data are used to investigate and reveal the temporal variations of the $EpCO_2$ and explain the mechanisms controlling these variations in a small-order river. This long-term, high-frequency dataset encompasses seasonality and differential time periods between hydrological events, but also allows many new features, including pulses and short duration events to be identified, which would not have been apparent with monthly or daily sampling.

To date, the temporal variations of $EpCO_2$ and the mechanisms controlling these variations have been considered using simple graphs, descriptive statistics, linear regression and multivariate statistics such as correlation analysis and analysis of variance ([18], [10], [11], [24], [5]). But, the dynamic responses of high-frequency data are complex and capturing them by conventional visualization and analysis techniques is difficult [16]. Recently, advanced statistical tools such as wavelets and generalized additive models have become useful tools for visualizing and analyzing localized time series variations and non-linear complex relationships, respectively; so their application to study the fluctuations of $EpCO_2$ is novel. The objectives of this paper are to: (i) visualize the temporal variations and decompose the variability of $EpCO_2$ at different timescales using wavelet analysis; and (ii) analyze and model the temporal variations of $EpCO_2$ and its relation with the water hydrology. The latter objective is achieved by fitting a set of hierarchal generalized additive models to describe the variations in $EpCO_2$ over a day, over a month, and finally over a full hydrological year. Using this temporal hierarchy, models which better explain the processes determining $EpCO_2$ are fitted, incorporating complex multi-variate interactions and lagged variables to account for the persistence of temporal correlations at the different temporal scales.

2 Materials and Methods

2.1 Study Site

The study site is in the Glen Dye catchment close to the terrestrial-aquatic interface of the River Dee in Aberdeenshire. Glen Dye is located in North-East Scotland at $56^o56'27$N and $2^o36'00$W. It is a headwater sub-catchment of the River Dee, a high-order river draining into the North Sea. The sensors were deployed at the Scottish Environment Protection Agency (SEPA) Charr gauging flume on the Water of Dye, a $41.7km^2$ catchment. Glen Dye is mainly upland in character, with altitude ranging between 100m and 776m. The climate is cold, with mean annual precipitation of 1130mm, of which $<$10% is snow. There is inter-annual variation in temperature with the winter months being December - February and the summer months being June - August. The underlying geology of the catchment is granite, with a small schist outcrop. The interfluves above 450m are covered by extensive peats ($<$ 5m deep) and peaty podzols ($<$ 1m). In some places peat is eroded to the mineral interface. Incised catchment slopes have the most freely-draining humus iron podzols ($<$ 1m deep); the main river valley bottoms generally have freely draining alluvial deposits. For a detailed description of the study site and its geology and climate characteristics, see [20].

2.2 Sampling Strategy and calculation of $EpCO_2$

Samples for measurement of Dissolved Inorganic Carbon (DIC) concentration were collected approximately every five hours over a 24-hour period and twelve times during June 2003 - August 2004. The sampling spanned a wide range of flow conditions. DIC (mmol L^{-1} C) is quantified by direct measurement using a headspace analysis approach [21], to internal precision better than $\pm$ 0.03 mmol L^{-1}. This was regressed onto discharge, which is measured semi-continuously, to generate a relationship from which DIC is predicted, thus creating a continuous DIC profile [20]. The generated relationship between discharge and DIC was indistinguishable from the same relationship constructed 10 years earlier, allowing confidence that this relationship is temporally stable over the constructed three years profile. Troll 9000EXP data loggers (In-Situ, Inc.) were used to generate 15 minutes frequency time series of temperature, pH and atmospheric pressure from October 2003 to September 2006. These parameters allowed the excess partial pressure of carbon dioxide ($EpCO_2$) to be indirectly calculated from the continuous DIC profile [21]. Estimates of the capacity for CO_2 efflux, are described as the "Excess partial pressure of CO_2", $EpCO_2$, a ratio of oversaturation (for more details, see [14] and [5]). The river system is over-saturated with CO_2 with respect to the atmosphere when $EpCO_2$ exceeds 1. The Troll loggers also generated 15 minute frequency time series of specific conductivity (SC). SC in streams and rivers is influenced by the river geology, in addition to the water flow and temperature [6]. It is usually higher in low flow periods when the groundwater contribution is proportionally highest [6].

2.3 Statistical Analysis and Methodology

The methodology applied here (i) visualizes and explores the variations of the $EpCO_2$ in the Glen Dye small-order river before, (ii) modeling and analyzing the temporal variations and the mechanisms controlling these variations. Approach (i) mainly uses graphical visualization methods and wavelet analysis to identify the temporal fluctuations of $EpCO_2$ and produce primary insights about its relationship with the water hydrology. Approach (ii) analyzes and explains the temporal variations of $EpCO_2$ and its relationship to catchment flow (as understood from SC) using generalized additive models. Describing and modeling the various patterns, fluctuations and interactions of $EpCO_2$ are the first step in identifying controls on the concentration.

2.3.1 Wavelets

Wavelet analysis is a useful tool for analyzing non-stationary and/or high frequency time series. Wavelets have the advantage of analyzing a time series by combining both, time and frequency domains. The time series is decomposed into a set of signals which relate to variations or changes at different scales. The result is a time-scale decomposition of the original signal that helps identify the cyclical components over different frequencies, as well as the long-term trend ([13], [17]).

The Discrete Wavelet Transform (DWT) decomposes a time series into discrete scales. The DWT is an orthogonal transform of the equally spaced time series $\{X_t : t = 1,\ldots,N\}$. The vector $\mathbf{W}$ of the DWT coefficients $\{W_n : n = 1,\ldots,N\}$ is given by:

$$\mathbf{W} = \mathbf{RX} \tag{1}$$

where $\mathbf{X}$ is the time series vector of length N and $\mathbf{R}$ is an $N*N$ real valued matrix defining the DWT constructed using the chosen filter such that $\mathbf{R}^T\mathbf{R} = \mathbf{I}$ [17]. Hence, the original signal $\mathbf{X}$ can be reconstructed as follows:

$$\mathbf{X} = \mathbf{R}^T\mathbf{W} = \sum_{j=1}^{J} \mathbf{D}_j + \mathbf{S}_J \tag{2}$$

where J is the level of decomposition, $\mathbf{D}_j$ is known as the wavelet detail and is associated with changes in the time series at scale $\tau_j = 2^{j-1}$ and $\mathbf{S}_J$ is called the wavelet smooth and is related to variations over scales $\tau_{J+1} = 2^j$ and higher. The wavelet smooth S_J represents a smooth version of $\mathbf{X}$. This decomposition is known as multi-resolution analysis (MRA). The MRA of the DWT takes the original signal and distributes it into different signals over the different dyadic scales $\tau_j = 2^{j-1}$, $j = 1,\ldots,J$ without losing the original available information, then analyzes each component with a resolution matched to its scale [17].

The Maximum Overlap Discrete Wavelet Transform (MODWT) is a modified version of the DWT which similarly decomposes the time series into a set of wavelet details plus a smooth component. The MODWT has some advantages over the DWT: first, it does not have any restrictions on the series length (i.e. N is not necessarily a power of two); second, the MODWT coefficients and associated MRA are not affected by the choice of the starting point of the time series. The trade-off of these advantages is loss of orthogonality and higher computational cost [17]. The MRA is useful in identifying the major timescale contributor to the variability in the time series. The estimated wavelet variance at a particular scale τ_j, $\hat{v}_X(\tau_j)$, which determines the contribution of that timescale to the variability of the series is given by:

$$\hat{v}_X(\tau_j) = \frac{1}{M_j} \sum_{t=L_j-1}^{N-1} W_{j,t}^2 \tag{3}$$

where $L_j = (2^j - 1)(L-1) - 1$ such that L is the filter width, $M_j = N - L_j + 1$ is the number of coefficients not affected by the boundary conditions which guarantees the unbiasedness of the estimator and $W_{j,t}$ are the wavelet coefficients at scale j and time t [17].

Here, MODWT based on least asymmetric filter of width equal to 8, LA(8), is applied to the $EpCO_2$ and the associated hydrological variables series. Least asymmetric (LA) filters are a special class of the Daubechies filters. The phase function of LA filters is very close to that of a linear phase filter, thus making it easy to line up features in the filtered series with the original series [17]. LA filter is then appropriate since it is of interest to align events in time. A filter width equal to 8 provides a good smooth representation of the corresponding time series and is chosen after comparing a series of wavelet transforms obtained for a range of filter width values. The wavelet transform corresponding to smaller filter width values resulted in sharp peaks in the individual elements of the time series decomposition, and greater width values did not make any difference. The `wmtsa` `R` package developed for wavelet analysis by Constantine and Percival in 2013 is used to obtain the MRA of the studied time series via the MODWT of the corresponding series. The wavelet transform cannot be applied to time series with missing data. Hence, the missing values are first imputed using linear interpolation. The interpolation is done separately for each month and for each time within the month to better reproduce the variability of the series.

2.3.2 Generalized additive modeling

Generalized Additive models (GAMs) are tools for describing and visualizing non-linear and non-parametric effects of explanatory variables X_k on a response variable of interest Y, without specifying a particular form for the regression function (see, for example, [7], [1]). A GAM has the following structure:

$$Y_t = \beta_o + f_1(X_{1t}) + f_2(X_{2t}) + f_3(X_{3t}, X_{4t}) + ... + \varepsilon_t \tag{4}$$

where the observations $Y_t, t = 1, \ldots, n$ are assumed to be independent with means $E(Y_t) = \mu_t$ and f_j are smooth functions of covariates X_k whose shapes are unrestricted and need to be estimated. In this context, the error term ε_t denotes an independent normally distributed random variable with mean 0 and variance σ^2. Hence, the distribution of Y_t is assumed Gaussian and GAMs are reduced to additive models. Here, the univariate smooth functions are approximated by cubic regression splines, except for the periodic effects which are estimated using cyclic cubic regression splines, and the bivariate smooth functions are represented by tensor product splines. Tensor product splines are invariant to linear scaling of covariates and are good to smooth interactions of quantities measured in different units [22]. The smoothness of each curve f_j is controlled by a smoothing parameter. The basis dimension of each smooth function is set based on the Akaike Information Criteria (AIC) to identify a smooth interpretable relationship. Then, the appropriate smoothing parameter of each smooth curve is selected automatically using the restricted maximum likelihood criteria. Likelihood methods tend to be more robust for smoothing parameter selection [23]. Model variable selection is performed using AIC and approximate F-tests. The `mgcv` package [22] supplied with `R` for generalized additive modeling is used. GAMs are fitted using the fitting routine `bam` which is an alternative for the main routine `gam` for very large data sets. For detailed discussion on splines and GAMs, see [22].

GAMs assume that the errors ε_t are mutually independent, while autocorrelation is one of the characteristics of hydrological time series. The GAMs only describe how the response variable is statistically related to the explanatory variables without accounting for the dependence of the response on its past values. Failure to account for autocorrelation appropriately may result in an underestimate of the standard errors for the estimated smooth curves, which makes the estimates inefficient and the inference about the estimates unreliable. One solution is a two-stage fitting procedure (TSP). The first stage involves fitting a GAM assuming independent distributed errors, and the second entails fitting an appropriate correlation structure to the residuals of the fitted GAM. An estimate for the correlation matrix of the residuals ε_t, $\hat{\mathbf{V}}$, can be obtained from the data based on the specified correlation structure. Each smooth component of the additive model has an estimate of the form $\hat{f}_j = \mathbf{S}_j\mathbf{y}$, where $\mathbf{S}_j$ is the smoothing matrix of component j and the standard errors are readily available as the square root of the diagonal entries $\mathbf{S}_j\hat{\mathbf{V}}\mathbf{S}_j^T\sigma^2$. The error variance σ^2 can be estimated from the $RSS = \mathbf{y}^T(\mathbf{I}-\mathbf{S})^T(\mathbf{I}-\mathbf{S})\mathbf{y}$ and the approximate degrees of freedom associated with error is given by $tr\{(\mathbf{I}-\mathbf{S})^T(\mathbf{I}-\mathbf{S})\mathbf{V}\}$. Then, the variability bands ($\pm$ 2 s.e.) of the estimated smooth curves can be adjusted based on the new standard errors [2]. An alternative to this two-stage procedure will be presented in the discussion.

3 Results

In this section, we first present the initial exploratory data analysis (EDA) and wavelets analysis results. Then, we show the results of the generalized additive models used to analyze the variations in $EpCO_2$ at the daily, monthly and yearly timescales.

3.1 Exploratory Data Analysis (EDA) and Wavelets

Fig. 1 displays the calculated $EpCO_2$ series and the recorded measurements of discharge, temperature, pH and SC for three hydrological years 2003-2006. The hydrological year runs from October to September and hereafter is abbreviated as HY. The $EpCO_2$ ranges from 0.26 to 10 and the average $EpCO_2$ is 2.57 ± 1.01 over the whole study period. Thus, our sample point on the Water of Charr is generally over-saturated with CO_2. $EpCO_2$ exhibits temporal variability. Similarly, water discharge is variable, with an average of 1.1 ± 4.5 m^3/s through the whole study period. Comparison of discharge between years shows that the HY2003/2004 had the wettest summer, HY2004/2005 had the driest winter and HY2005/2006 was the wettest overall (Table 1). The coldest months in the three hydrological years are December - February (Fig. 1) with an average water temperature of 2.9 ± 1.7^oC; the warmest months are June - August with an average temperature of 14 ± 2.8^oC.

Fig.2 illustrates the seasonal and diurnal responses in $EpCO_2$ in each of the HYs. The median $EpCO_2$ (represented by the black bar in the middle of each box) is generally higher in summer (June - August) than winter (December - February). $EpCO_2$ is more variable during the summer. The hourly boxplots show that the median $EpCO_2$ is smallest close to midday and largest just after midnight and that $EpCO_2$ is more variable during darkness.

	Discharge (m^3/s)		
HY	Winter	Summer	Overall
HY2003/2004	10083	8882	37063
HY2004/2005	6496	3880	35958
HY2005/2006	12385	3726	40044

Table 1 Total water discharge at the water of Charr sampling point across the winter (December - February) and summer (June - August) of each HY and the whole HY.

The EDA shows inter-annual and intra-annual variations in the $EpCO_2$ and the other hydrological variables, showing the time series to be non-stationary. Animated 3-D plots available in the supplementary material at (link) provide a better representation of the interactions between the $EpCO_2$ and the variables describing the water hydrology. The $EpCO_2$ is highly dynamic and the hydrodynamics might also contribute to the $EpCO_2$ variability, e.g. discharge events are sometimes associated with a particular feature in the $EpCO_2$ series. However, the response of $EpCO_2$ to flow events may differ depending on preceding events and differences in summer and winter biological productivity. In conclusion, it is not easy to draw interpretable conclusions for such complex high-frequency data using simple exploratory methods. One way to better visualize and describe the temporal variations of these HFD in the time-frequency domain is to use wavelets.

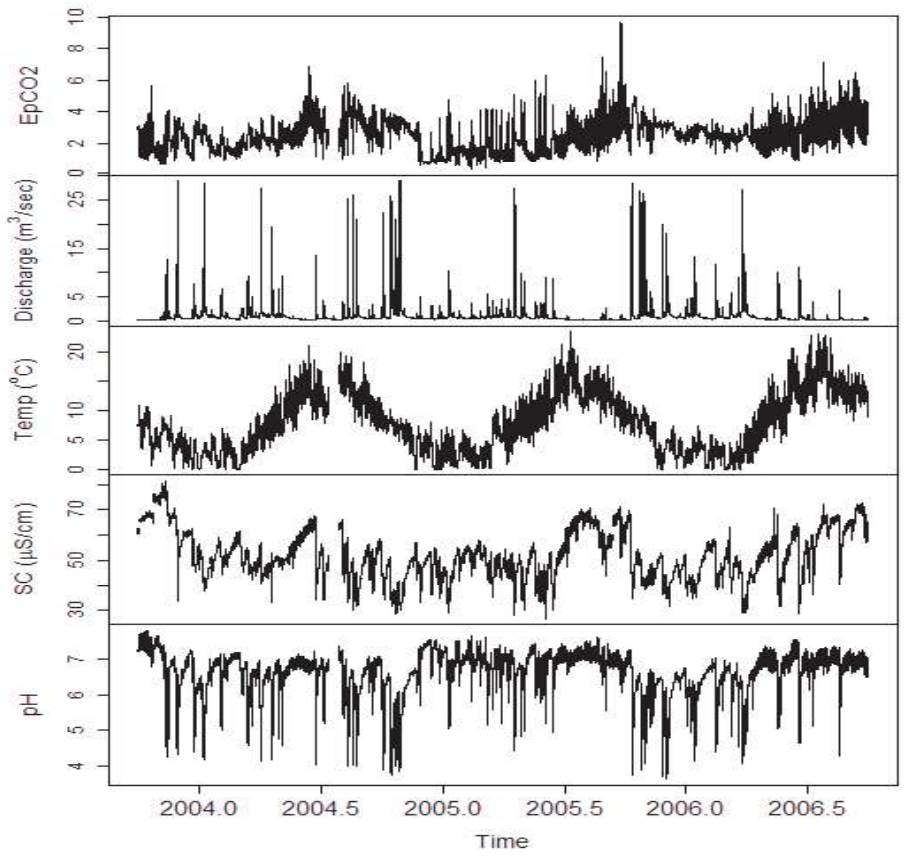

Fig. 1 Time plots of $EpCO_2$, Flow, Temperature, pH and SC series from October 2003 to September 2006. The tick marks on the x-axis corresponds to January 1^{st} and July 1^{st}.

There are some periods of missing data (Fig. 1). The $EpCO_2$ series in 2003/2004 has 1544 missing values in total, one in February 2004 and the rest in July 2004 (see the top panel of Fig. 1), which represent 4.4% of the total record. Each of the pH, temperature, and SC series of 2003/2004 and 2004/2005 has less than 5% missing values of the total record. All the missing values are imputed as mentioned earlier before starting the wavelet analysis. These imputed values are shown in grey in Fig. 3.

The MODWT with LA(8) filter decomposes the $EpCO_2$ series for each of the hydrological years into 12 wavelet details and one smooth component ($X = \sum_{j=1}^{12} D_j + S_{12}$), where 12 is the maximum number of scales. The wavelet details (D_j, $j = 1,\ldots,12$) reflect changes in the original series over scales of $15(2^{i-1})$ minutes and the smooth component (S_{12}) relates to variations at about 44 days and higher and represents the overall trend. The MRA of the $EpCO_2$ series (Fig. 3 for D_1, D_6 and D_{12}), represents changes in the original series on a scale of 15 minutes, 8 hours and $\sim$22 days, respectively. The MRA indicated that the detail components D_j, $j = 1,\ldots,4$ (only D_1 is shown here due to space limitations) are the least

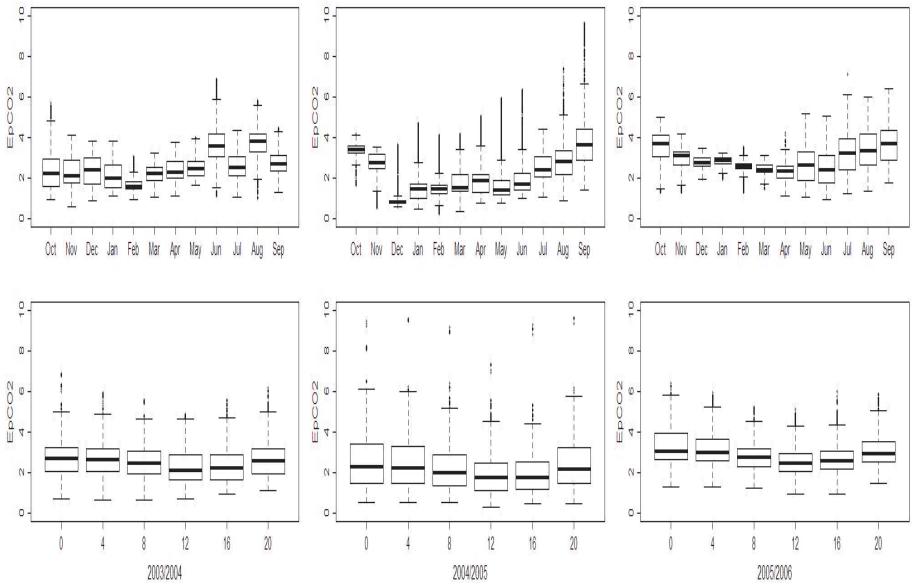

Fig. 2 $EpCO_2$ in each Month (top) and Hour (bottom) in the hydrological years 2003/2004 (right), 2004/2005 (middle) and 2005/2006 (left).

variable reflecting the small scales variability and can be related to weather or hydrological events. Therefore, these high-frequency scales capture the uncommon $EpCO_2$ levels which might be influenced by short-lived changes in the water hydrology such as intense periods of rainfall. D_6 is the main contributor to the sample variance of the $EpCO_2$ (see Fig. 4) and the associated temperature series reflecting the presence of an intra-daily cycle. This seems reasonable since changes over a scale of 8 hours correspond to the daylight cycle. However, the diel cycle is not constant throughout each hydrological year but larger fluctuations occur during summer. This MRA also shows that the $EpCO_2$ of the dry summer of HY2005/2006 exhibits this diel pattern clearly for longer compared to the wetter summers of HY2003/2004 and HY2004/2005. The wavelet detail D_{12} can be seen as an approximation for the monthly variations since it reflects changes over nearly 22 days.

Fig. 5 compares the 6^{th} wavelet detail series, which represents the changes over a scale of 8 hours, for the different hydrological variables with the $EpCO_2$ series. The timing, extent and number of occurrences of hydrological events differ from one hydrological year to another. The highest $EpCO_2$ variability is usually associated with little changes in discharge, consistent with internal fluvial carbon cycling, while hydrological events are associated with compressed $EpCO_2$ variability. The periods when pH and SC show most change occur with flow events. The variability of $EpCO_2$ evolves coherently with the variability in temperature, in itself a proxy for seasonality: $EpCO_2$ appears to be more variable during the summer when there are larger fluctuations between day and night temperatures. The differences in temperature and discharge influence the SC and pH, which in turn influence the $EpCO_2$ across the different years.

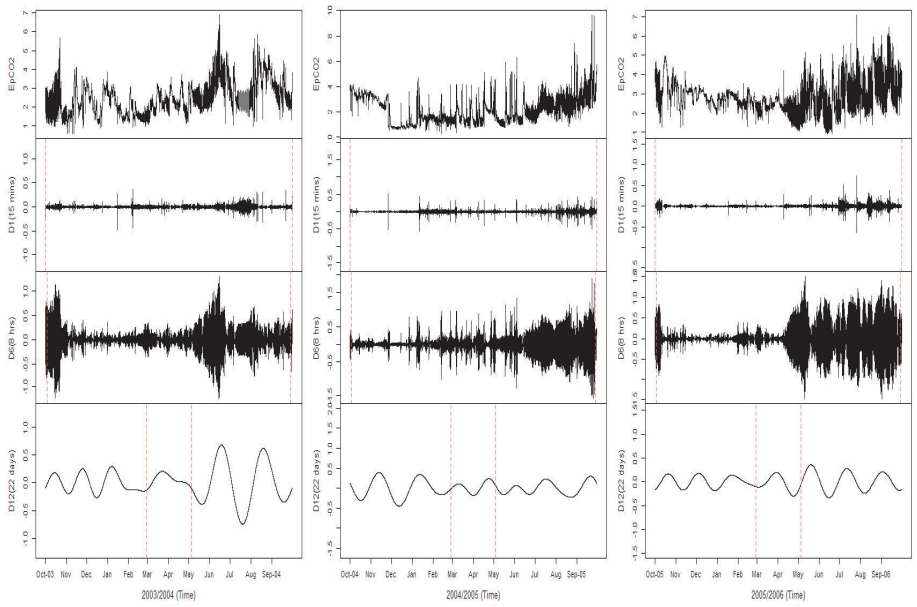

Fig. 3 Multi-resolution analysis of $EpCO_2$ series for the hydrological years 2003/2004 (right), 2004/2005 (middle) and 2005/2006 (left). The wavelet details D_1 (15 mins), D_6 (8 hrs) and D_{12} ($\sim$22 days) are on the same scale, different from the original series (top). The dashed vertical lines indicate the areas that might be affected by boundary coefficients.

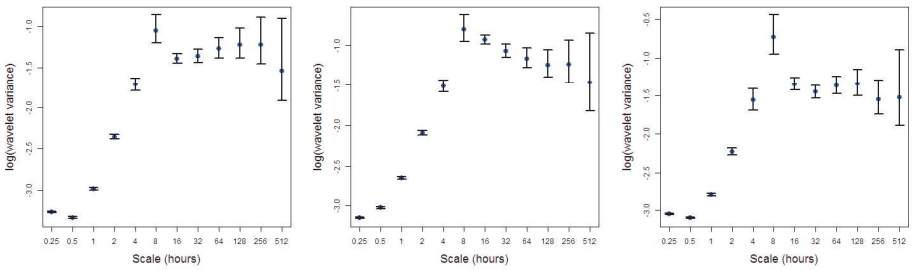

Fig. 4 Wavelet variance of the $EpCO_2$ series of the hydrological years 2003/2004 (right), 2004/2005 (middle) and 2005/2006 (left) for the scales $15(2^{j-1})$, $j = 1, \ldots, 12$.

The EDA and wavelets analysis highlighted the seasonal and diurnal fluctuations of $EpCO_2$ and the differences in these variations between the individual hydrological years. They also revealed that the hydrodynamics appear to contribute to part of the $EpCO_2$ variability although the nature of these relationships is very complex and difficult to explore and visualize through exploratory tools. It is not clear from the exploratory analysis whether or not the temporal patterns in $EpCO_2$ can be described entirely by hydrology. In addition, the EDA cannot highlight the persistent temporal correlation between the high-frequency measurements after we account for temporal dynamics. Therefore, a set of hierarchal GAMs are fitted at different temporal scales to better describe the variations in $EpCO_2$ at these timescales.

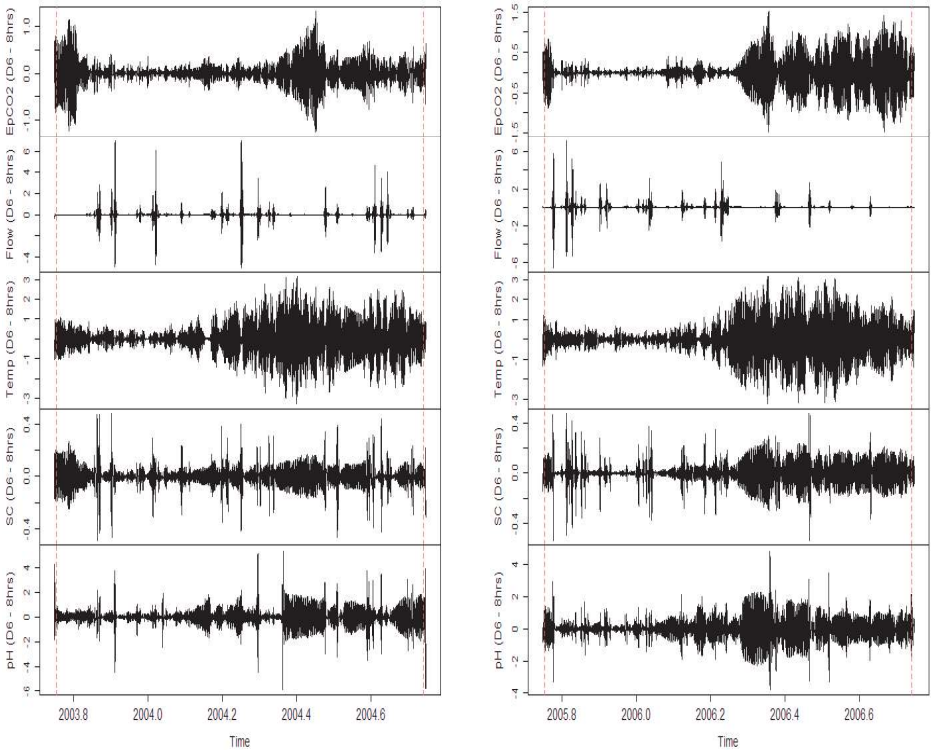

Fig. 5 6th wavelet detail (8 hrs scale) of MRA of $EpCO_2$ (top), flow, temperature, SC and pH (bottom) for the hydrological years 2003/2004 (left) and 2005/2006 (right). The dashed vertical lines indicate the areas that might be affected by boundary coefficients.

3.2 Generalized additive models

Initially, GAMs are developed for individual days followed by individual months and finally for each hydrological year. These GAMs are useful in explaining the variations in $EpCO_2$ and studying the relationship between $EpCO_2$ and the available physiochemical catchment variables, which are not used in deriving the $EpCO_2$ (i.e. SC), within the day, month and hydrological year. They also describe the differences in variations between the different days, months and hydrological years. This temporal hierarchy better shows the changes and the increased complexity of (i) the processes driving $EpCO_2$, (ii) the multi-variate interactions between $EpCO_2$, water hydrology and time components, and (iii) the temporal correlation structures from the daily to the yearly timescales.

3.2.1 Daily GAMs

It is evident from the previous EDA and MRA that the $EpCO_2$ exhibits a diel cycle with altering magnitude and pattern from one day to another. These alterations could be attributed to seasonal changes or other hydrological conditions. Let Y_t denote the $EpCO_2$ at time point t, and $\mathbf{X}_t = (X_t^{\text{Time within day}}, X_t^{\text{Hour of day}}, X_t^{\text{SC}})$ be the vector of explanatory variables, where $X_t^{\text{Time within day}}$ is a continuous variable representing the time within the day at

which the measurement is recorded, $X_t^{\text{Hour of day}}$ is an index of the hour within day and X_t^{SC} is the measured SC at time t. Then, the daily variations of $EpCO_2$ are described through the following GAM:

$$Y_t = f_1(X_t^{\text{Time within day}}) + f_2(X_t^{\text{SC}}) + f_3(X_t^{\text{Hour of day}}, X_t^{\text{SC}}) + \varepsilon_t \quad (5)$$

where the smooth functions f_j, $j = 1,2,3$, capture the daily cycle, the main effect of SC and the bivariate effect of hour within day and SC on $EpCO_2$, respectively; and ε_t accounts for the random effects not explained by the GAM. The functions f_1 and f_2 are represented using cubic regression splines and f_3 using tensor product spline, as described in section 2.3.2. The model is fitted to the data of some selected days. The model assumptions, including independence of the errors, are shown to be all valid. The estimated GAM explains about 99% deviance of the data of each selected day.

Fig. 6 shows only the results of 14/10/2005, 14/1/2006, 14/4/2006 and 14/7/2006. As can be seen, the fitted splines capture the response of $EpCO_2$ to time of day reflecting the changes in the biological activity according to the daylight cycle. This intra-daily cycle of $EpCO_2$ changes from one day to another, according to the seasonal and hydrological conditions and is significantly stronger in the summer days. It is also clear that the relationship between $EpCO_2$ and SC is significantly changing with hour and day, justifying the multivariate interactions between $EpCO_2$, hydrodynamics and time.

3.2.2 Monthly GAMs

The EDA has identified seasonal differences in the behavior and fluctuations of $EpCO_2$. These monthly/seasonal variations are explained via fitting the following GAM for each month of the hydrological year separately:

$$\begin{aligned} Y_t = {} & f_1(X_t^{\text{Time within month}}) + f_2(X_t^{\text{Hour of day}}, X_t^{\text{SC}}) + \\ & f_3(X_t^{\text{Hour of day}}, X_t^{\text{Day of month}}) + f_4(X_t^{\text{Day of month}}, X_t^{\text{SC}}) + \varepsilon_t \end{aligned} \quad (6)$$

where $X_t^{\text{Time within month}}$ is a continuous variable denoting the time within each month; f_1 determines the main behavior of $EpCO_2$ within each studied month; and f_2 describes the bivariate effect of hour within day and SC. Based on the daily GAMs results, the smooth functions f_3 and f_4 are added to the model to capture the changing effects of hour within day and SC from day to day, respectively. f_1 is approximated using cubic regression splines; and f_j, $j = 2,3,4$, are represented by tensor product splines.

Only the model results of January and June 2005 are presented due to space limitations. The estimated GAMs explain 72% and 91% deviance of the data in January and June, respectively. However, the ACF of the model residuals shows a slowly decaying correlation structure in January, and not only significant correlations at high lags, but a remaining periodic pattern every 24 hours that is not captured by the model in June. These dependence structures affect the efficiency of the estimates and make the inference procedure unreliable. In January, the estimated GAM residuals are modeled via an autoregressive process of order 1 (AR(1)), which has accounted for the remaining dependence. In June, a greater degree of structure was displayed in the residuals after fitting the GAM. Therefore, 2 and 8 hours

lagged dependent variables are added to the set of explanatory variables. The 2-hour lag denotes the average extent of short-term dependence, while the 8-hour lag represents the extent of long-term dependence. The 2-hour and 8-hour lagged $EpCO_2$ account for the long range dependence and the periodic dependence structure. The smooth functions of these two lagged dependent variables are approximated by cubic regression splines. Then, any remaining autocorrelation is accounted for via an AR(1) process fitted to the residuals of the adjusted model. The final residuals appear independent.

The monthly GAMs indicate that the $EpCO_2$ dynamics vary across the different months of the hydrological year. Fig. 7 illustrates the clear dissimilarities in the trend, variability of $EpCO_2$ and interactions with time and water hydrology between January and June. It is evident that the $EpCO_2$ is more variable in June. Fig. 7 typically shows the evidence of intra-daily cycle in June and its absence in January. The $EpCO_2$ and the magnitude of its diel cycle changes significantly from one day to another within June. The figure also indicates that the daylight cycle in June has a greater significant influence on the fluctuations of $EpCO_2$ than water hydrology during the absence of hydrological events (at high SC). Conversely, water hydrology dampens the diel cycle and dominates these fluctuations at periods of hydrological events. In January, $EpCO_2$ does not seem to exhibit an intra-daily cycle and variations are mostly attributed to hydrodynamics.

Generally, both time and hydrology contribute to the variations in $EpCO_2$. However, the contribution of the temporal and hydrological is time-dependent and changes from one season to another. Also, the temporal correlation remaining between the residuals after accounting for these variations changes seasonally and shows more complex structures in summer. It is evident that this temporal auto-correlation is more persistent when the model is extended to cover a longer time period.

3.2.3 Yearly GAMs

The monthly GAMs illustrated intra-annual variations in $EpCO_2$. However, the EDA indicated the non-stationarity of the full time series and the presence of inter-annual variations, as a result of the different climatological characteristics characterizing each HY. Therefore, the model is extended to describe the variations in $EpCO_2$ within each HY and highlight the differences between the three hydrological years. As the model covers a longer time period, the auto-correlation structure becomes more difficult to model. Hence, the yearly variations of $EpCO_2$ are described through the following TSP:

$$Y_t = f_1(X_t^{\text{Time within year}}) + f_2(X_t^{\text{Hour of day}}, X_t^{\text{Day of year}}) + f_3(X_t^{\text{Hour of day}}, X_t^{\text{SC}}) + f_4(X_t^{\text{Day of year}}, X_t^{\text{SC}}) + f_5(Y_{t-8}) + f_6(Y_{t-32}) + \varepsilon_t \quad (7)$$

$$\varepsilon_t = \phi\varepsilon_{t-1} + \xi_t \quad (8)$$

where $X_t^{\text{Time within year}}$ denotes a continuous variable representing the time within the year to reflect the yearly trend; and Y_{t-8} and Y_{t-32} denote the 2 and 8 hours lagged $EpCO_2$, respectively, which were successful in accounting for the periodic dependence structure in the monthly models. The smooth function f_1 captures the global trend of $EpCO_2$ along each hydrological year; f_2 describes the changing effect of the daily cycle from day to day; f_3 and f_4 explain the bivariate effect of SC with hour and day of year, respectively; and f_5

and f_6 capture the effect of the 2 hours and 8 hours lagged $EpCO_2$ on the current $EpCO_2$, respectively . As previous, f_1, f_5 and f_6 are represented by cubic regression splines and f_j, $j = 2,3,4$ by tensor product splines. The residuals ε_t in Equation 7 follow an AR(1) (see Equation 8), where ϕ is known as the autoregressive parameter and ξ_t is a white noise process with mean 0 and variance σ^2_ξ.

The fitted models explain about 95% of the variability in $EpCO_2$ in each hydrological year. It is evident that the $EpCO_2$ is higher when SC is lower (occurring during events when lower SC soil water is proportionally more important) and that the diel cycle is dominating the changes in $EpCO_2$ at high SC levels (Fig. 8) i.e. at low flow when in-stream biological processes are most dominant. By incorporating lagged dependent variables in the model, $EpCO_2$ for the three years exhibits the same patterns but with different magnitude. However, there is still some periodic structure left between the residuals of the yearly scale models after adding the lagged $EpCO_2$ to the GAM and modeling the residuals via AR process.

In brief, it is evident that the processes controlling the $EpCO_2$ are time and scale dependent. The multi-variate relationships between the $EpCO_2$, water hydrology and time components changes from one scale to another and become more complex when the model is extended to describe a longer time period within the hydrological year. In addition, the autocorrelation structure between the residuals remaining after accounting for the temporal and water hydrological changes with time and becomes more persistent and composite at the yearly scale. Therefore, lagged variables and more multi-variate interactions are added to explain the increased variability and account for the persistence of temporal correlations at the larger timescales.

4 Discussion and Conclusion

It is evident that although high-frequency data provide information which was previously inaccessible, they pose various challenges to statistical modeling and analysis. We evidence this here using as an illustrative dataset a high-resolution time series of $EpCO_2$. Exploring and modeling these high-resolution sensor data was very complex and challenging because of the differences in the behavior of the variable of interest over the different timescales, the complex multi-variate relationships which are time and scale dependent and the persistent temporal correlation characterizing such high-frequency data.

The primary EDA showed that $EpCO_2$ is non-stationary and exhibits variations over a wide range of timescales. $EpCO_2$ is generally higher in summer than winter and more variable during summer due to the greater catchment productivity in the summer when more CO_2, or sources of, are available, and greater in-stream processing of C results in CO_2 consumption (during day-time) and production (during night-time). This processing can be seen in the intra-daily cycle of $EpCO_2$, which is lowest close to midday (maximum solar radiation to support photosynthesis) and highest just after midnight, when respiration has occurred for longest and so CO_2 concentration is highest.

Wavelet analysis helped identify temporal variability, including intra-daily, seasonal and inter-annual variations. These variations arise due to changes in the relative strength of external (e.g. climatological) and internal (biological processing) drivers of resultant $EpCO_2$.

The MRA indicated that the intra-daily cycle is the major contributor to the variability of the $EpCO_2$ series. This intra-daily cycle reflects the dark-light-dark cycle within the day. The amplitude of this diel cycle is not constant throughout the year but larger variability occurs during summer when a pronounced diurnal cycle is present. It is also evident that the variability resulting from the daylight cycle changes from one year to another, again reflecting different balances of external and internal drivers of $EpCO_2$.

The hierarchal GAMs fitted over a day, a month and a year showed that the variability of $EpCO_2$ and its relationship with water hydrology are time and scale dependent. The GAMs allow temporal variations and the mechanisms controlling $EpCO_2$ across the different timescales to be accommodated. These temporal variations and multi-variate relationships change across the different timescales and become more complex as the model is extended to cover a longer time period within the hydrological year. These GAMs showed that the $EpCO_2$ exhibits a 24 hour dark-light-dark cycle reaching the minimum value at noon. The magnitude of this day/night cycle changes along the year and is more apparent during the summer where the $EpCO_2$ reaches its maximum levels. It was also obvious that the hydrology has an influence on the level of $EpCO_2$. At low flow, DIC concentration is highest [20] and biological activity is greatest (as temperature tends to be higher); event flow, whilst flushing out soil CO_2 so increasing the pool size, ultimately dilutes the DIC pool and so lowers saturation of dissolved carbon dioxide. Turbulent waters and colder temperature reduce biological activity. As such CO_2 over-saturation is reduced and $EpCO_2$ decreases. The diel cycle can still exist, but variability is reduced in winter. Seasonality in flow thus has a significant effect on the $EpCO_2$. The diel cycle appears to dominate the $EpCO_2$ variations in summer during the absence of hydrological events and during low flows (evidenced by higher SC), while high flow events dampen the diel cycle in winter. Hence, the contribution of the temporal and hydrological variations changes with season and timescale. Consequently, the fitted GAMs encountered some problems in uniquely identifying the sources of variability and the contribution of each variable to the variability in $EpCO_2$.

Serial autocorrelation is one of the characteristics of high-resolution time series. The residuals of the fitted GAMs displayed a periodic autocorrelation structure that persists over a large number of lags. The complexity of the dependence structure increases from daily to yearly timescales. Therefore, modeling HFD by assuming independence is no longer valid. A two-stage fitting procedure has been used here, where some lagged dependent variables are added to the model and then an AR process is fitted to the adjusted model residuals. An alternative method would be to incorporate the correlation structure through using the structure of a generalized additive mixed model (GAMM) using the `gamm` function in the `mgcv` library in `R` [22]. The GAMM simultaneously fits a GAM and a mixed effect model that accounts for the autocorrelation between the model residuals. Incorporating auto-correlation through GAMM results in higher smoothing parameters being selected and hence the remaining structure to be accounted for in the residuals increases. This increases the complexity of modeling required for the residuals. Whereas in the TSP, the first stage results in optimally selecting lower smoothing parameters assuming independent errors, which reduces the complexity of modeling required for the residuals in the second stage. After incorporating lagged terms and a simple correlation structure, only a small amount of structure is still remaining between the residuals of the yearly models. Future work could include investigating models such as ARCH/GARCH models to account for the remaining structure in the residuals. Such models are able to capture the varying variation in the residual process. Although care has to be taken since autocorrelation can influence smoothing parameter se-

lection from automatic approaches, GAMMs are computationally inefficient with large time series and numerically unstable because of the confounding between correlation and non-linearity. Therefore, alternative methods to fit GAMs with correlated data are required.

In conclusion, these HFD have illustrated the complex long-term and short-term dynamics of $EpCO_2$ which were previously inaccessible with lower frequency data. However, these HFD encounter various challenges in terms of statistical modeling and analysis. The challenges facing the description and analysis of such a complex high-resolution datasets must be overcome to avoid limiting insight into, e.g. catchment processes. Among these challenges are (i) the great volumes of data, (ii) the complex multi-variate interactions between the covaiates and the response variable, (iii) the complex correlation structures persisting over a large number of lags between observations due to the high-frequency nature of the data, and (iv) the identifiability problems in allocating the existing large variability to the signal or noise as a result of the confounding between correlation and non-linearity. Therefore, advanced statistical tools and models are needed to analyze such complex HFD.

Acknowledgements AE is grateful to the Glasgow University sensor studentship for funding. SW generated the $EpCO_2$ profiles through Natural Environment Research Council Advanced Fellowship, NER/J/S/2001/00793. Stephanie Evers is thanked for field assistance, and SEPA for providing discharge records. SW is most grateful to the Fasque Estate, particularly Archie Dykes, for site access and providing accommodation.

References

1. Bowman A, Azzalini A (1997) Applied Smoothing Techniques for Data Analysis:The Kernel Approach with S-plus Illustrations, 1st edn, Oxford Statistical Science Series, Oxford University Press, Oxford
2. Bowman A, Giannitrapani M, Scott M (2009) Spatiotemporal smoothing and sulphur dioxide trends over Europe, Journal of the Royal Statistical Society, 58, 737–752
3. Butman D, Raymond PA (2011) Significant efflux of carbon dioxide from streams and rivers in the United States, Nature Geoscience,
4. Cole JJ, Caraco NF, Kling GW, Kratz TK (1994) Carbon dioxide supersaturation in the surface water of lakes, Science, 265, 1568–1570
5. Dawson J, Soulsby C, Hrachowitz MS, Telzlaff D (2009) Seasonality of EpCO2 at different scales along an integrated river continuum within the Dee Basin NE Scotland, Hydrological Processes, 14, 2929–2942
6. United States Environmental Protection Agency EPA (2012), Water: Monitoring and Assessment, http://water.epa.gov/type/rsl/monitoring/vms59.cfm. Accessed 27 May 2014.
7. Hastie TJ, Tibshirani RJ (1990) Generalized Additive Models, 1st edn, Monographs on Statistics and Applied Probability, Chapman and Hall, London
8. Intergovernmental Panel on Climate Change IPCC (2013) Climate Change 2013: The Physical Science Basis, http://www.climatechange2013.org/report/. Accessed 20 March 2014.
9. Kirchner J, Fang X, Neal C, Robson A (2004) The fine structure of water quality dynamics: the (high-frequency) wave of the future, Hydrological Processes, 18, 1353–1359
10. Li S, Lu XX, Bush RT (2013) CO2 partial pressure and CO2 emission in the Lower Mekong River, Journal of Hydrology, 504, 40–56
11. Li S, Lu XX, He M et al (2012) Daily CO2 partial pressure and CO2 outgassing in the upper Yangtze River basin: a case study of Longchuanjiang, Journal of Hydrology, 466-467, 141–150
12. Moraetis D, Efstathiou D, Stamati F et al (2010) High-frequency monitoring for the identification of hydrological and bio-geochemical processes in a Mediterranean river basin, Journal of Hydrology, 389, 127–136
13. Nason GP (2008) Wavelets Methods in Statistics With R, 1st edn, Springer, Use R!
14. Neal C (1998) Determination of dissolved CO2 in upland streamwater, Journal of Hydrology, 99, 127–142
15. Neal C, Reynolds B, Rowland P et al (2012) High-frequency water quality time series in precipitation and streamflow: From fragmentary signals to scientific challenge, Science of the Total Environment, 434, 3–12

16. Neal C, Reynolds B, Kirchner J et al (2013) High-frequency water quality time series in precipitation and streamflow: From fragmentary signals to scientific challenge, Hydrological Processes, 27, 2531–2539
17. Percival D, Walden A (2006) Wavelets Methods for Time Series Analysis, 1st edn, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, Cambridge
18. Raymond PA, Caraco NF, Cole JJ (1997) Carbon dioxide concentration and Atmoshperic flux in the Hudson River, Estuaries, 20, 381–390
19. Richey JE, Melack JM, Aufdenkampe AK et al (2002) Outgassing from Amazonian Rivers and wetlands as a large tropical source of atmospheric CO2, Nature, 416, 617–620
20. Waldron S, Scott M, Soulsby C (2007) Stable isotope analysis reveals lower-order river dissolved inorganic carbon pools are highly dynamic, Enviornmental Science Technology, 41, 6156–6162
21. Waldron S, Scott M, Vihermaa LE, and Newton J (2014) Quantifying precision and accuracy of measurements of dissolved inorganic carbon stable isotopic composition using continuous-flow isotope-ratio mass spectrometry, Rapid Communications in Mass Spectrometry , 28 (10), 1117-1126
22. Wood SN (2006) Generalized Additive Models - An introduction with R, 1st edn, Text in Statistical Science Series Chapman and Hall, London
23. Wood SN (2011) Fast stable REML and ML estimation of semiparametric GLMs, JRSSB, 73, 1–34
24. Yao G, Gao Q, Wang Z et al (2007) Dynamics of CO2 partial pressure and CO2 outgassing in the lower reaches of the Xijiang River, a subtropical monsoon river in China, Science of the Total Environment, 376, 255–266
25. Yick J, Mukherjee B, Ghosal D (2008) Wireless sensor network survey, Computer Networks, 52, 2292–2230

Amira Elayouty is currently a Ph.D student in Statistics, in the School of Mathematics and Statistics at the University of Glasgow, Glasgow, UK, G12 8QW. Her research interests include spatio-temporal models, non-parametric regression and additive models with a particular focus on environmental statistics.

Professor Marian Scott is a Professor of Environmental Statistics, in the School of Mathematics and Statistics at the University of Glasgow, Glasgow, UK, G12 8QW (marian.scott@glasgow.ac.uk). Her research interests include varying-coefficient and additive models, spatiotemporal models, quantile regression and functional data analysis.

Dr Claire Miller is a Senior Lecturer in Statistics, in the School of Mathematics and Statistics at the University of Glasgow, Glasgow, UK, G12 8QW (claire.miller@glasgow.ac.uk). Her research interests include nonparametric, varying-coefficient and additive models, environmetrics, spatiotemporal models and functional data analysis.

Prof. Susan Waldron holds a personal chair in Biogeochemistry and heads the Carbon Landscape Research Group (www.carbonlandscapes.org). A geologist by training, her research focuses on the carbon cycle particularly transfer of C from the terrestrial environment to aquatic systems and from there to the atmosphere. A key focus of her research is environmental resilience and response to hosting energy production and land-based renewable and using sensor technology to capture the detail of the environmental response.

Dr Maria Franco-Villoria is a research assistant in Statistics, in the Department of Economics and Statistics at the University of Turin, Italy (maria.francovilloria@unito.it). Her research interests include environmental statistics and functional data analysis.

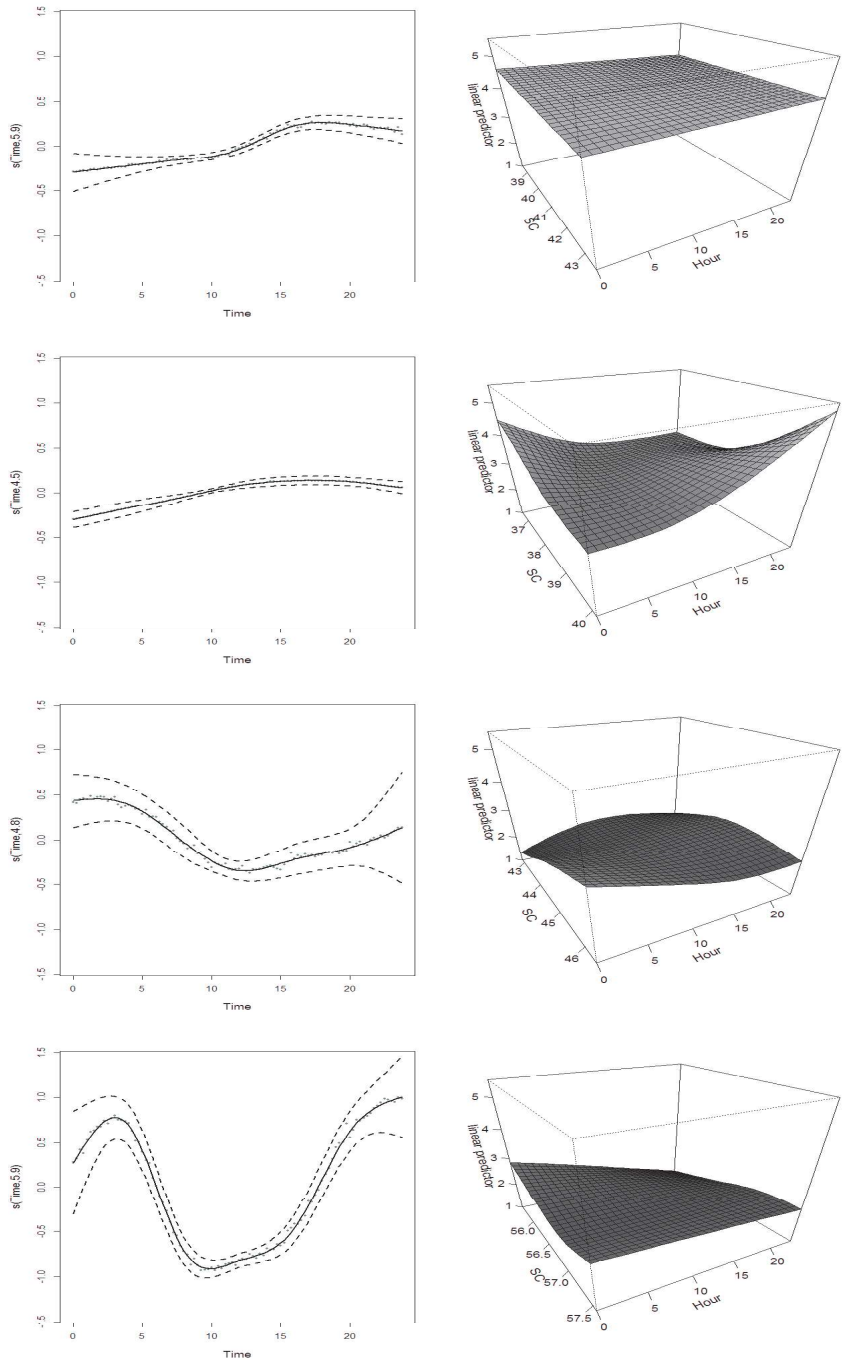

Fig. 6 The fitted smooth functions of Time (left) and the interaction between SC and Hour of day (right) of the daily GAM for the days 14/10/2005 (top), 14/1/2006, 14/4/2006 and 14/7/2006 (bottom). The dashed lines in the left panels are the $\pm$ 2 s.e. bands.

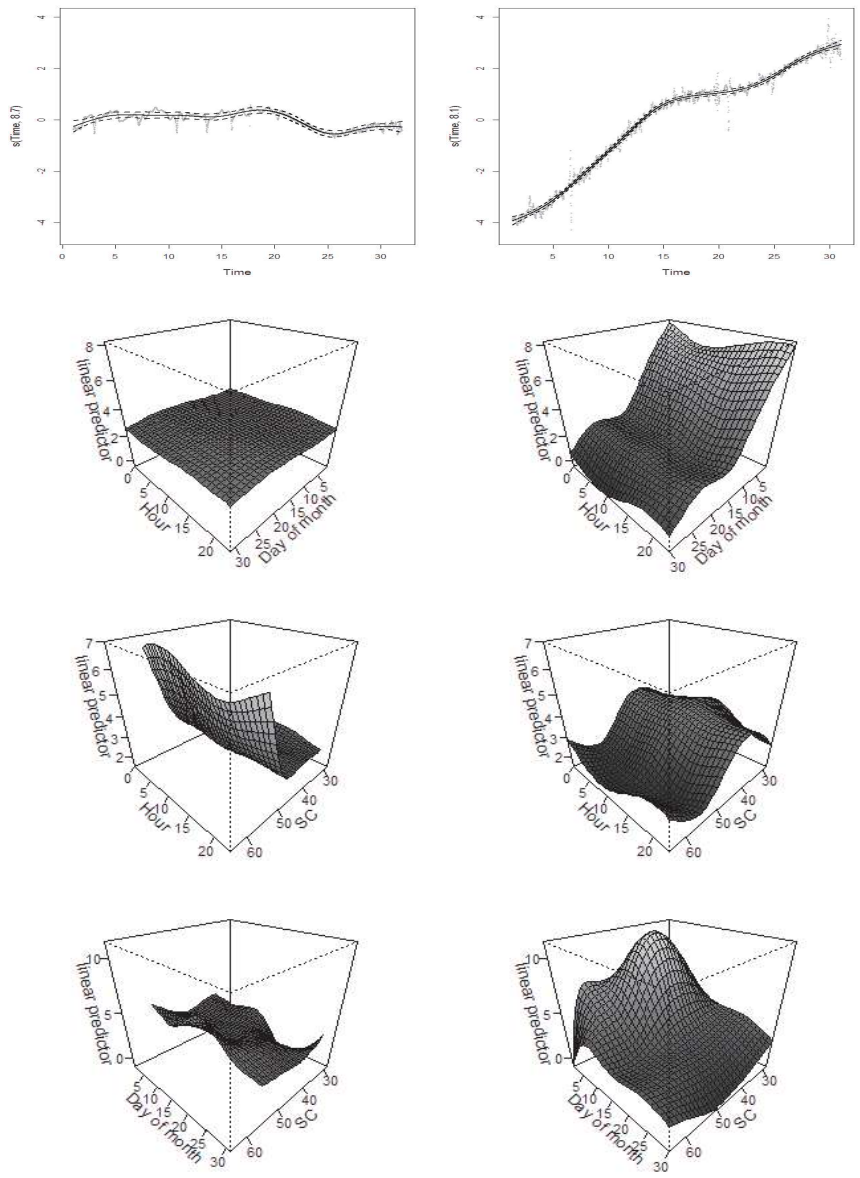

Fig. 7 The fitted smooth functions of Time (top) and the interactions: Hour of Day and Day of month; Hour of Day and SC; and Day of month and SC (bottom) of the monthly GAM for January (left) and June (right) 2005. The dashed lines in the top panels are the adjusted $\pm$ 2 s.e. intervals after accounting for the autocorrelation present in the residuals ε_t.