
Article:
Hargreaves, Jessica Kate orcid.org/0000-0002-7173-7902, Knight, Marina Iuliana orcid.org/0000-0001-9926-6092, Pitchford, Jonathan William orcid.org/0000-0002-8756-0902 et al. (2 more authors) (2018) Clustering nonstationary circadian plant rhythms using locally stationary wavelet representations. SIAM Multiscale modeling and simulation. pp. 184-214.

https://doi.org/10.1137/16M1108078

CLUSTERING NONSTATIONARY CIRCADIAN RHYTHMS USING LOCALLY STATIONARY WAVELET REPRESENTATIONS*

JESSICA K. HARGREAVES †, MARINA I. KNIGHT †, JON W. PITCHFORD ‡, RACHAEL J. OAKENFULL §, AND SETH J. DAVIS§

Abstract. Rhythmic processes are found at all biological and ecological scales, and are fundamental to the efficient functioning of living systems in changing environments. The biochemical mechanisms underpinning these rhythms are therefore of importance, especially in the context of anthropogenic challenges such as pollution or changes in climate and land use. Here we develop and test a new method for clustering rhythmic biological data with a focus on circadian oscillations. The method combines locally stationary wavelet time series modelling with functional principal components analysis and thus extracts the time-scale patterns arising in a range of rhythmic data. We demonstrate the advantages of our methodology over alternative approaches, by means of a simulation study and real data applications, using both a published circadian dataset and a newly generated one. The new dataset records plant response to various levels of stress induced by a soil pollutant, a biological system where existing methods which assume stationarity are shown to be inappropriate. Our method successfully clusters the circadian data in an interesting way, thereby facilitating wider ranging analyses of the response of biological rhythms to environmental changes.

Key words. evolutionary wavelet spectrum, nondecimated wavelet transform, nonstationary processes, unsupervised learning, plant circadian clock

AMS subject classifications. 62P10

1. Introduction.

The earth rotates on its axis every 24 hours resulting in a day and night cycle. Correspondingly, almost all species exhibit changes in their behaviour between day and night (Bell-Pedersen et al., 2005). These daily rhythms are not only caused by a response to daily changes in the physical environment, but are also the result of an internal timekeeping system or 'biological clock' within the organism (Vitaterna et al., 2001; Minors and Waterhouse, 2013). In particular, most plants are able to anticipate dawn and adjust their biochemistry accordingly. When an organism is deprived of external time cues, these rhythms typically persist qualitatively but may change in detail; the study of these changes can reveal the biochemical reactions underpinning the circadian clock and, at a larger scale, can provide valuable insight into the possible consequences of environmental change (McClung, 2006; Bujdoso and Davis, 2013).

Experiments recording plant response to light entrainment result in datasets that, from a statistical point of view, can be considered as time series realisations. Period and phase estimation (see Figure S1 in Appendix A for a visual interpretation of this terminology) are the fundamental elements of most circadian analyses. The current standard uses BRASS (Biological Rhythm Analysis Software System (Edwards et al., 2010)) to estimate the period of each time series using Fourier analysis (see Moore et al. (2014) or Zielinski et al. (2014) for a complete description of the under-

*Submitted to the editors December 13, 2016.

Funding: This work was supported by the EPSRC. Circadian work in the SJD group is currently funded by the BBSRC awards BB/M000435/1 and BB/N018540/1.

†Department of Mathematics, University of York, York, YO10 5GE, UK (jkh516@york.ac.uk, marina.knight@york.ac.uk).

‡Departments of Mathematics and Biology, University of York, York, YO10 5GE, UK (jon.pitchford@york.ac.uk).

§Department of Biology, University of York, York, YO10 5GE, UK (rachael.oakenfull@york.ac.uk, seth.davis@york.ac.uk).

lying period analysis methods). Data stationarity is an implicit assumption within the underlying methodology – put simply, its statistical characteristics are assumed constant over time. However, in reality, nonstationary behaviour is common in biological systems (Zielinski et al., 2014). Here we propose, develop and test methods that are capable of detecting changes of period over time by drawing on the plant time-frequency signature as quantified by its spectrum.

The methodology developed here is general, but our concrete example concerns (i) identifying if a plant's clock is affected under exposure to different concentrations of ammonium cerium nitrate, (ii) establishing which concentrations produce similar effects and (iii) subsequently characterising these effects. The answers to these questions have important implications, not only for the understanding of the mechanism of the plant's circadian clock, but also for the environmental impact associated with soil pollution (Yang et al., 2016).

In order to answer the above questions, we propose to estimate the spectral behaviour of our time series under the formal framework of locally stationary wavelet (LSW) processes (Nason et al., 2000), which are able to account for data nonstationarity. Wavelets are ideal for identifying discriminant local time and scale (frequency) features, and time-frequency (scale) patterns are known to be indicative of the plant response to various stimuli (Zielinski et al., 2014). A functional principal components analysis on the spectral data treated as an 'image' (as suggested in a Fourier context by Holan et al. (2010)) is then used to reduce the data dimensionality and allows the extraction of important behavioural features. Furthermore, this functional representation is also used to inform a clustering method that facilitates quantifying the effects induced by different concentrations of ammonium cerium nitrate.

This article is organized as follows. Section 2 outlines the novel circadian dataset and establishes its nonstationary behaviour; it also reviews state-of-the-art circadian data analysis tools present in the current literature. Section 3 develops our proposed novel locally stationary wavelet-based clustering method. The findings of an extensive simulation study are presented in Section 4. Section 5.1 demonstrates the additional insight our clustering method can provide when applied to a published circadian plant dataset. Section 5.2 presents the results of clustering the novel circadian plant dataset using the proposed methodology and examines them in the context of several relevant biological questions. Section 6 concludes with a brief discussion and suggests topics for further investigation.

2. Motivation.

In this section we briefly outline the experimental details that led to a novel circadian plant dataset and assess the prominent features of the circadian plant rhythms under analysis, namely their lack of stationarity. This result, along with several others recorded in the literature (e.g. Price et al. (2008), Leise et al. (2013)) motivates the development of analysis techniques that can account for nonstationarity. Furthermore, we also discuss the phenomenon of individual-level variability in plant response to stimuli, despite their sharing identical genetic characteristics (Doyle et al., 2002). The presence of multiple behaviours within the same treatment group motivates our development of a clustering procedure that can detect these different characteristics and analyse them separately. For completeness, we also report the results of the analysis a circadian biologist would typically use.

2.1. Experimental details.

The novel circadian dataset (henceforth referred to as the cerium dataset) was obtained by the Davis Lab (Biology, University of York) following a similar method to Hanano et al. (2006). For a detailed description of these methods see Appendix B. Briefly, for each plant, gene expression levels are

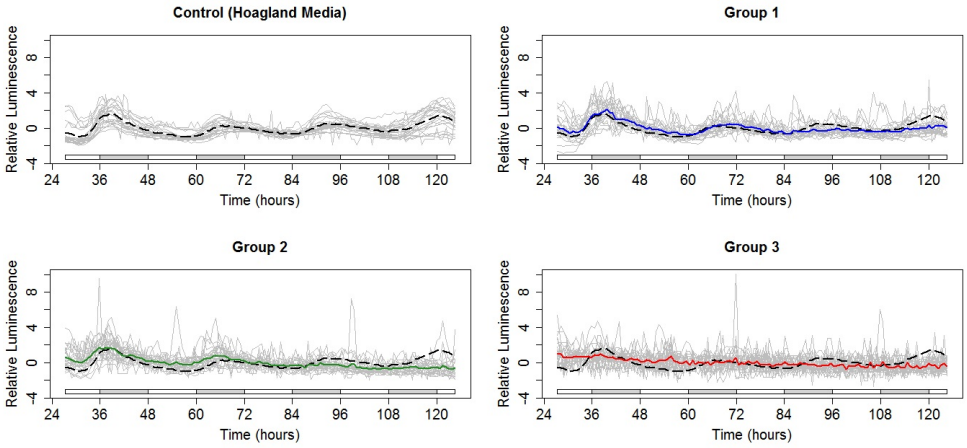

FIG. 1. *Luminescence evolution over time for plants subjected to a control and 3 different ammonium cerium nitrate concentrations. Time is measured in hours relative to zeitgeber time (time of last external temporal cue: the dawn signal of lights-on). Top left: Each plant signal from the control group (in grey) along with the group average (dashed black). Other panels: Each realisation from the groups (in grey) along with the group average and the control group average (dashed black). Group 1:* $100\mu M$ *ammonium cerium nitrate with average in blue. Group 2:* $150\mu M$ *ammonium cerium nitrate with average in green. Group 3:* $200\mu M$ *ammonium cerium nitrate with average in red. (Each time series has been normalised to have mean zero.) Note: the free run started from time 24; shaded bars below each graph indicate the subjective darkness that plants expected to experience during the 'normal' day.*

measured (using a firefly luciferase reporter system) at regular intervals resulting in an individual time series. In this experiment, the gene of interest was 'cold and circadian regulated and RNA binding 2', known as CCR2 (Doyle et al., 2002).

The cerium dataset consists of a total 96 plant signals (time series) recorded at 128 time points, with the control and groups 1–3 (each corresponding to a different concentration of ammonium cerium nitrate) all containing 24 plants. The control group is grown in Hoagland's media (Hoagland et al., 1950), which contains essential nutrients required for plant growth, and is not exposed to any additional levels of ammonium cerium nitrate. To examine the effects of cerium on the circadian clock, the other three groups, while also grown in Hoagland's media, were additionally exposed to varying additional concentrations of ammonium cerium nitrate– 100μM for Group 1, 150μM for Group 2 and 200μM for Group 3. A plot of individual luminescence time series, the average expression at each time point, for each of the treatment groups, is shown in Figure 1. Note that time is measured in hours relative to *zeitgeber* time, which is the time of the last external temporal cue: the dawn signal of lights-on.

2.2. BRASS analysis. In the circadian community, analysis of this data would typically be performed by the Microsoft Excel macro BRASS. Table 1 provides a summary of the output of the analysis of the cerium dataset in BRASS. In particular, it shows the mean period estimate (obtained using FFT-NLLS analysis (Plautz et al., 1997) considering only period estimates between 15 and 40 hours), the number of plants that could not be analysed by BRASS and the mean Relative Amplitude Error (RAE) for each of the 4 groups. RAE is a value between 0 and 1 and gives information about the goodness of fit of the model (a value of 0 indicates a perfect fit). Circadian

Group	Hoagland's	Group 1 (100μM)	Group 2 (150μM)	Group 3 (200μM)
Average period estimate (in hours)	27	27	26	24
Number of plants excluded by BRASS	7	10	12	21
Average RAE	0.23	0.44	0.41	0.74

TABLE 1

Summary of the output of the analysis of the circadian dataset in BRASS. The 'number of plants excluded by BRASS' is the number of time series for which BRASS was not able to return a period estimate. 'RAE' (Relative Amplitude Error) is a value between 0 and 1 and gives information about the goodness of fit of the model (a value of 0 indicates a perfect fit). Recall: there are 24 plants in each of the groups.

biologists often visualise these results in a scatter plot of relative amplitude error against period length for the plants analysed by BRASS (see e.g. Hanano et al. (2006)) and such a plot for this dataset is given in Figure S2, Appendix A.

On examining Table 1, note that not all data is used to produce the period estimate reported by BRASS– in particular, the 'number of plants excluded by BRASS' is the number of time series for which the FFT–NLLS algorithm (Plautz et al., 1997) was not able to return a period estimate, possibly due to a loss of rhythmicity. Thus, under the assumption of stationarity (and the above constraints), these methods are not able to analyse all data produced by this experiment, indicating that this dataset is not suitably modelled using Fourier methods. Furthermore, by just reporting the results of this analysis, the biologist would conclude that adding 100μM or 150μM ammonium cerium nitrate produces no detectable effect on the circadian clock (as these period estimates are similar). However, visual examination of Figure 1 shows that ammonium cerium nitrate appears to have a strong effect on these plants, providing further evidence that more statistically advanced approaches are needed.

2.3. Nonstationarity in circadian rhythms. Price et al. (2008) asserted that data arising from circadian experiments is nonstationary and discussed the features which support this claim, namely a progressively dampened signal with a changing period. The authors advocated the use of wavelets to analyse circadian data and developed a technique for characterising the modal periods present in circadian data using a continuous wavelet decomposition (this is disseminated in the `waveclock` package in R, currently on CRAN archive). Later, Harang et al. (2012) also supported the circadian data nonstationarity view, and furthermore claimed that circadian analysis under nonstationary behaviour by means of traditional Fourier methods can lead to inaccurate results. Harang et al. (2012) thus recommended the use of wavelets, which allow the changes in period to be tracked through time, and developed 'WAVOS'- a wavelet-based MATLAB toolkit that allows for analysis of nonstationary circadian data.

Leise et al. (2013) discussed the appropriateness of traditional methods to determine period length from experimental datasets that assume a rhythm of fixed period and amplitude, proposing that most biological rhythms exhibit changes in both period and amplitude. Therefore, the authors extended wavelet methods to measure how biological rhythms vary over time and developed MATLAB scripts to implement their analysis using both continuous and discrete wavelet transforms.

For our novel circadian dataset, we investigated whether the individual plant

Group	Hoagland's	Group 1 (100μM)	Group 2 (150μM)	Group 3 (200μM)
Number of nonstationary plants	22	19	19	8

TABLE 2

Results for the Priestley-Subba Rao test of stationarity, implemented in the `fractal` *package in R and available from the CRAN package repository. Number of nonstationary plants indicates the number of time series (in each group) with enough evidence to reject the null hypothesis of stationarity at the 1% significance level. Recall: there are 24 plants in each of the groups.*

signals are (second-order) stationary via hypothesis testing. We employed two tests for stationarity– a Fourier-based test (Priestley and Rao, 1969) and a wavelet-based test (Nason, 2013). The Fourier-based test we used was the Priestley-Subba Rao (PSR) test. The results, which can be found in Table 2, show that over 70% of the plant signals provided enough evidence to reject the null hypothesis of stationarity. This conclusion is backed-up by the wavelet-based spectrum test for stationarity. Additionally, this test also indicates where the nonstationarities are located in the series. (A visual representation for each group can be found in Figure S3, Appendix A.)

Therefore, in agreement with previous observations in circadian literature, both tests suggest that our circadian data also displays nonstationary features. In order to assess the impact of different concentrations of ammonium cerium nitrate, we propose a novel clustering technique that combines the use of wavelets (ideal for analysing nonstationary behaviour) with rigorous statistical (process) modelling. Additionally, to mitigate against individual plant variability, our technique proposes the use of time-scale patterns as explained next.

2.4. Individual-level variability in circadian rhythms. We noticed in our dataset the presence of individual-level variability in plant responses to the same stimuli, despite their sharing identical genetic characteristics (Doyle et al., 2002). For example, different types of behaviour can be seen in the control group of Figure 1. This is particularly noticeable at the beginning (prior to time $T = 36$) and end (after time $T = 96$) of the experiment where the plant signals displayed one of two different amplitudes. This variability highlights the issues caused by taking an average period estimate for each group and comparing the results, or comparing the average raw time series for each group. Although all plants in each treatment group share identical genetic characteristics and have been treated in identical conditions, they respond differently. In such situations, looking at average behaviour masks the individual differences and is conducive to misleading conclusions, as also acknowledged in other fields (Fiecas and Ombao, 2016). This motivates our choice to cluster the circadian plant data using their time-frequency (scale) patterns and further accounts for their proven (see Section 2.3) nonstationary features.

3. Proposed clustering method. Our proposed methodology combines the use of wavelets, as recommended (but not implemented) by Zielinski et al. (2014) in their review of period estimation methods for circadian data, with rigorous stochastic nonstationary time series modelling. We exploit the locally stationary wavelet processes of Nason et al. (2000), arriving at a novel and general approach for clustering circadian signals according to their leading time-scale spectral patterns, as extracted by functional principal components analysis.

3.1. Modelling nonstationary time series. Many of the statistically rigorous approaches to modelling nonstationary time series are based on the Cramér-Rao representation of stationary processes: all zero-mean discrete time second-order stationary time series $\{X_t\}_{t\in\ \mathbb{Z}}$ can be represented as

$$X_t = \int_{-\pi}^{\pi} A(\omega)\exp(i\omega t)d\xi(\omega), \tag{1}$$

where $A(\omega)$ is the amplitude of the process and $d\xi(\omega)$ is an orthonormal increments process (Priestley, 1982).

In the representation in equation (1) above, we note that, for stationary processes, the amplitude $A(\omega)$ does not depend on time (i.e. the frequency behaviour is the same across time). However, for many real time series, including the cerium dataset, this assumption is not realistic and a model where the frequency behaviour can vary with time would therefore be preferable. One way of introducing time dependence into a model is by replacing the amplitudes $A(\omega)$ with a time-dependent form. Priestley (1965) introduced a time-frequency model with the amplitude replaced by $A(\omega, t)$, while Dahlhaus (1997) introduced the locally stationary modelling philosophy and developed the locally stationary Fourier (LSF) model. In this setting, the time-dependent amplitude function is defined on 'rescaled time' to enable asymptotic considerations.

Later, Nason et al. (2000) introduced a locally stationary wavelet model, where the Fourier building blocks (present in the LSF model) are replaced by families of discrete nondecimated *wavelets*. This statistical modelling framework allows the process to have time-dependent amplitudes that in their turn induce a time-dependent second-order structure (e.g. time-dependent evolutionary wavelet spectrum). The advantage of wavelets is that they are localised in both time and scale (frequency) and are therefore well-suited to modelling second-order characteristics that evolve over time. Therefore, the locally stationary wavelet model combines the advantages of a wavelet analysis with rigorous stochastic nonstationary time series modelling. (We refer the interested reader to Daubechies (1992) and Nason (2010) for detailed texts on wavelets and their applications in statistics.)

In our work we adopt the locally stationary wavelet (LSW) process framework, under which a time series $\{X_{t;T}\}_{t=0}^{T-1}$, $T = 2^J \geq 1$ is defined to be a sequence of (doubly-indexed) stochastic processes with the following representation

$$X_{t,T} = \sum_{j=1}^{J}\sum_{k\in\mathbb{Z}} w_{j,k;T}\psi_{j,k}(t)\xi_{j,k}, \tag{2}$$

where $\{\xi_{j,k}\}$ is a random orthonormal increment sequence, $\{\psi_{j,k}(t) = \psi_{j,t-k}\}_{j,k}$ is a set of discrete non-decimated wavelets and $\{w_{j,k;T}\}$ is a set of amplitudes, each of which at a scale j and time k.

The properties of the random increment sequence $\{\xi_{j,k}\}$ ensure that $\{X_{t,T}\}$ is a zero-mean process– in practice, it is customary to detrend a process with non-zero mean, and this is our approach here.

Estimation under the LSW framework is made possible by controlling the speed of evolution of the amplitudes $\{w_{j,k;T}\}$ using a condition of the form $\sup_k|w_{j,k;T} - W_j(k/T)| \leq C_j/T$, where $W_j(z), z \in (0,1)$ is a 'limiting' amplitude function with

various smoothness constraints and $\{C_j\}_j$ is a set of constants with $\sum_{j=1}^{\infty} C_j < \infty$ (Nason et al., 2000).

The definition of the LSW process in Equation (2) requires the data to be of dyadic length ($T = 2^J$). In many practical applications, this is not realistic and there are a number of approaches to address this situation. For example, the practitioner could truncate the time series and analyse a segment of the data (of length $T = 2^J$), and this is our approach here. Alternatively, it is possible to extend the data to the next greater power of two by artificially appending values. In particular, common approaches include padding the data with zeros, replicating a data value (such as the final value) or reflecting the dataset about an end point. Another approach is to interpolate data values to produce a new data set of the required length (Ogden, 1997). However, preconditioning the data could lead to misleading results. Therefore, we do not artificially extend the data in this paper.

An analogous quantity to the spectrum of a stationary process, which quantifies the contribution of a frequency (ω) to the process variance, is introduced in the LSW setting. This quantity, commonly referred to as the evolutionary wavelet spectrum (EWS), quantifies the power distribution in an LSW process over *time and scale* and is formally defined as

$$S_j(z) = |W_j(z)|^2, \tag{3}$$

at each scale $j \in \overline{1, J}$ and rescaled time $z = k/T \in (0, 1)$.

An unbiased estimator of the EWS $\{S_j(z)\}$ is obtained by correcting the raw wavelet periodogram $I^j_{k,T} = |d_{j,k;T}|^2$, where $d_{j,k;T} = \sum_{t=0}^{T} X_{t,T}\psi_{j,k}(t)$ are the empirical nondecimated wavelet coefficients. The correction is attained by premultiplying the raw wavelet periodogram vector $\mathbf{I}(z) := (I^j_{\lceil zT \rceil,T})_{j=1}^{J}$ by the inverse of the autocorrelation wavelet inner product $(J \times J)$ matrix, $A_J = (\sum_\tau \Psi_j(\tau)\Psi_l(\tau))_{j,l}$, where $\Psi_j(\tau) = \sum_k \psi_{j,k}(0)\psi_{j,k}(\tau)$ is the autocorrelation wavelet.

Thus, the corrected wavelet periodogram is

$$\mathbf{L}(z) = A_J^{-1}\mathbf{I}(z), \text{ for all } z \in (0, 1). \tag{4}$$

As in the stationary setting, the wavelet periodogram is not a consistent estimator of the wavelet spectrum (Nason, 2010). One method to overcome this is to smooth the raw wavelet periodogram as a function of (rescaled) time within each scale j, and then to apply the correction above. Various smoothing approaches have been proposed in the literature, see e.g. smoothing using variance stabilisation of Fryzlewicz and Nason (2006).

In what follows, let us denote the corrected and smoothed periodogram of a time series (plant signal) $\{X_{t,T}\}_{t=0}^{T-1}$ as $\{\hat{S}_j(z)\}_j$, for rescaled time $z \in (0, 1)$.

3.2. Overview of current clustering/classification techniques that account for nonstationarity. The problem of clustering and classification for nonstationary data has received a good deal of attention in the statistical literature, thanks to its relevance in many applied fields. In the context of monitoring potential nuclear testing, Shumway (2003) considered the use of time-varying spectra for the classification and clustering of nonstationary time series by means of locally stationary Fourier models and Kullback-Leibler discrimination measures. Also in this context, Fryzlewicz and Ombao (2009) developed a procedure for the *classification* of nonstationary time series. The observed data were modelled as realisations of locally

stationary wavelet processes and their corresponding wavelet spectra were estimated and used as the signal classification signature. In the context of an industrial experiment, Krzemieniewska et al. (2014) further developed this method by proposing an alternative divergence index to the simple squared quadratic distance of Fryzlewicz and Ombao (2009) for comparing the spectra of two time series. Note that the above techniques are underpinned by rigorous process modelling but the focus is on classification into known groups, rather than on clustering. When classifying animal communication signals, known to have a nonstationary character, Holan et al. (2010) achieved dimension reduction by treating each windowed Fourier spectrum as an 'image' and performing a functional principal components analysis. In this context, the authors proposed to classify nonstationary time series by means of a generalised linear model that incorporated the (dimension-reduced) spectrogram of a short-time Fourier transform into the model as a predictor.

For clustering applications, the maximum covariance analysis (MCA) on wavelet representations of *two series* has been proposed in previous works. MCA has the advantage of extracting common time-scale (frequency) patterns while also reducing the dimension of the data. Rouyer et al. (2008) used MCA to yield a quantitative measure of the common time-scale content in squared wavelet coefficients for pairs of time series. This subsequently yields a distance matrix used to obtain a cluster tree that groups signals according to their spectral time-scale patterns. In the context of an energy application, Antoniadis et al. (2013) also used an MCA over the wavelet coefficients obtained via a continuous wavelet transform and quantify signal similarity by comparing the evolution in time of each pair of leading patterns. This builds a distance matrix which is then used within classical clustering algorithms to differentiate among high dimensional populations.

Formally, consider two time series, $\{X_t^{(i)}\}$ and $\{X_t^{(j)}\}$. Both Antoniadis et al. (2013) and Rouyer et al. (2008) obtained a time-scale decomposition of each time series (the wavelet transform and its squared version, respectively). Regardless of the usage of wavelet coefficients or their squared version, denote these new quantities in the wavelet domain by $Q^{(i)}$ and $Q^{(j)}$, for the $\{X_t^{(i)}\}$ and $\{X_t^{(j)}\}$ signals respectively, and define the time-scale covariance matrix by

$$R^{(i,j)} = Q^{(i)} Q^{(j)H}, \tag{5}$$

where $Q^{(j)H}$ denotes the conjugate transpose and $R^{(i,j)}$ is a $J \times J$ matrix with possibly complex values. Performing a singular value decomposition of $R^{(i,j)}$ gives the following decomposition:

$$R^{(i,j)} = U^{(i)} \Lambda^{(i,j)} V^{(j)H} \tag{6}$$

where the columns of $U^{(i)}$ and $V^{(j)}$ are the orthonormal singular vectors of $Q^{(i)}$ and $Q^{(j)}$ respectively, and $\Lambda^{(i,j)}$ is a diagonal matrix with the singular values of the decomposition arranged in decreasing order. Denote the k-th pair of the singular vectors of $U^{(i)}$ and $V^{(j)}$ as u_k and v_k respectively. We can then define the k-th leading pattern as the projections of $Q^{(i)}$ and $Q^{(j)}$ over their respective k-th singular vectors:

$$P_k^{(i)} = u_k^H Q^{(i)} \text{ and } P_k^{(j)} = v_k^H Q^{(j)}. \tag{7}$$

This process is then repeated for each pair of time series to produce the leading patterns and singular vectors which are then used with various distance measures

(described in Section 3.4.1) to obtain the dissimilarity matrix which forms the input of classical clustering algorithms.

Contrasting with the classification techniques described above, these clustering approaches are not underpinned by rigorous statistical modelling, and while they propose respectively the usage of wavelet coefficients or their squares, the reasoning that should drive this choice is not discussed by either Rouyer et al. (2008) or Antoniadis et al. (2013).

3.3. Proposed functional principal components analysis for the wavelet spectral content. In this work we propose to combine the rigorous modelling framework provided by the locally stationary wavelet (LSW) processes that allows for the reliable (unbiased and consistent) estimation of the spectral time-scale features specific to each plant, with the dimension reduction afforded through the use of a functional principal components analysis (FPCA).

In our biological problem of interest, the time-scale representation of the signal is high-dimensional. Since any useful biological information is likely to relate to the low-dimensional mechanisms known to regulate the clock (Bujdoso and Davis, 2013), this motivates our proposal to use a FPCA to perform dimension reduction over the spectral content. In the spirit of Holan et al. (2010), we treat our LSW spectral estimate as an 'image' and the spectral coefficients as time-scale 'pixels'. The pixels are not independent– in fact, the spectrum presents coherent patterns that should be accounted for. This motivates the use of the Karhunen-Loéve representation (at the heart of FPCA) which, in our context, for a continuous spectrum $\{S(\mathbf{v}) : \mathbf{v} = (j, z), \mathbf{v} \in \mathbb{R} \times (0, 1)\}$ allows for its covariance function $C_S(\mathbf{v}, \mathbf{v}')$ to be decomposed via an eigen-decomposition (Ramsay and Silverman, 2005). Consequently, the spectra may be decomposed as $S(\mathbf{v}) = \sum_{m \geq 1} \alpha_m \phi_m(\mathbf{v})$, with scores $(\alpha_m)_m$ independent random variables whose variance is given by the corresponding eigenvalues ($\mathrm{Var}(\alpha_m) = \lambda_m$) and $\phi_m(\mathbf{v})$ orthonormal eigenvectors that capture the variability in the spectral domain.

Assuming we observed N plant signals at $T = 128$ equally spaced time points, we model the i-th plant signal as an LSW process $\{X_{t,T}^{(i)}\}_{t=0}^{T-1}$ for each $i = 1, \ldots, N$. As biological evidence points towards the relevance of the plant spectral signature in understanding its response to stimuli, we estimate the wavelet spectrum by means of its corresponding corrected and smoothed periodogram, $\{\hat{S}_j^{(i)}(t/T)\}_{j=1}^{J}$ for each time series $i = 1, \ldots, N$, where $t = 0, \ldots, T-1$ and $J = \log_2(T)$. The estimated spectra, viewed as continuous functions $\{\hat{S}^{(i)}(\mathbf{v})\}$ with $\mathbf{v} = (j, z = t/T) \in \mathbb{R} \times (0, 1)$, are then treated as input observations in a FPCA. Their corresponding estimated covariance function $\hat{C}(\mathbf{v}, \mathbf{v}')$ thus summarises the dependence of plants across time *and* scale.

Although the continuous Karhunen-Loéve representation is often the most realistic from the point of view of modelling a biological process, due to the discrete nature of observations resulting from most experiments, it is rarely considered in applications. In practice, we use its empirical version, also known as empirical orthogonal function analysis, as is common in e.g. spatial statistics and geophysics (Cressie and Wikle, 2015). In particular, the estimated spectral coefficients can be arranged in N matrices, each of size $J \times T$, which we denote $\hat{S}^{(1)}, \ldots, \hat{S}^{(N)}$. For each plant signal (each $i = 1, \ldots, N$), vectorise the matrix $\hat{S}^{(i)}$, i.e. concatenate the rows of the matrix $\hat{S}^{(i)}$ to produce a vector $\hat{\mathbf{s}}^{(i)}$ with length $J \times T = n$. These N vectors are combined to form a data matrix Q of size $N \times n$, where each row of Q represents the spectral content of a plant. Formally,

$$Q = \left[\hat{\mathbf{s}}^{(1)}, \ldots, \hat{\mathbf{s}}^{(N)}\right]^T. \tag{8}$$

Note that in practice, this analysis is equivalent to performing a classical principal components analysis on the mean centred data, which we still denote by Q in order not to further clutter the notation. The spectral decomposition of the sample covariance matrix $R = Q^T Q$ is given by

$$R = U \Lambda U^T, \tag{9}$$

where U is an orthonormal matrix whose columns are the eigenvectors of R (also known as the principal directions of the data; here, we can conceptualise these as representing 'images') and Λ is a diagonal matrix whose diagonal elements are eigenvalues of R (positive real numbers arranged in decreasing order of magnitude; these are proportional to the variance accounted for by each direction). We can achieve size reduction by choosing to represent our data in fewer dimensions. The usual practice is to use the set of $p < n$ eigenvectors of R corresponding to the p largest eigenvalues and aggregate these in an $n \times p$ matrix, U_{PCA}, which performs the PCA projection. Therefore, for each eigenvector, we can find a corresponding projection in the principal component space by computing QU_{PCA}. In this transformed space, each process is now represented by a p-dimensional vector, i.e. the principal co-ordinates of the i-th process are given by the i-th row of the matrix QU_{PCA}, denoted from now on as $\text{Score}^{(i)}$ (p-dimensional vector).

3.4. Proposed clustering method. Our proposal is to construct a clustering method that assesses time series similarity/ dissimilarity on the basis of their spectral content as distilled in the scores developed in Section 3.3 above. Next we shall introduce potential distance measure candidates and assess various methods to determine the number of principal components to retain and the optimal number of clusters.

3.4.1. Distance measures. The success of any clustering algorithm depends on the adopted dissimilarity measure. In this section, we propose four possible distance measures and discuss their advantages and disadvantages. The proposed distance measures consist of developments of those adopted in the work reviewed in Section 3.2. In our simulation studies (Section 4), we compare the performance of clustering algorithms embedding the different distance measures outlined below.

The simplest choice for the dissimilarity measure is the squared quadratic (SQ) distance between two time series, $\{X_{t,T}^{(i)}\}_{t=0}^{T-1}$ and $\{X_{t,T}^{(j)}\}_{t=0}^{T-1}$. This distance measure is adopted by Fryzlewicz and Ombao (2009) who quote its advantages of good practical performance and computational ease. In our context it is defined as the sum of the squared differences between the scores relating to the p principal components retained

$$SQ(X_{t,T}^{(i)}, X_{t,T}^{(j)}) = \sum_{k=1}^{p} \left[\text{Score}_k^{(i)} - \text{Score}_k^{(j)}\right]^2, \tag{10}$$

where $\text{Score}_k^{(i)}$ denotes the score associated to the k-th principal component of time series $\{X_{t,T}^{(i)}\}$, as explained above. The value $SQ(i,j)$ is the (i,j)th entry of the dissimilarity matrix, D.

Our proposal is to develop this simplistic measure by aggregating the scores in the most significant p directions using a *weighted* combination with weights given by the squared singular values. We refer to this measure as the weighted squared quadratic (WSQ) distance and define the WSQ distance between two time series, $\{X_{t,T}^{(i)}\}_{t=0}^{T-1}$

and $\{X_{t,T}^{(j)}\}_{t=0}^{T-1}$ as the weighted sum of the squared differences between their scores in p directions. Formally

$$WSQ(X_{t,T}^{(i)}, X_{t,T}^{(j)}) = \frac{\sum_{k=1}^{p} \lambda_k \left[\text{Score}_k^{(i)} - \text{Score}_k^{(j)}\right]^2}{\sum_{k=1}^{p} \lambda_k}, \tag{11}$$

where $\text{Score}_k^{(i)}$ is as in equation (10) and λ_k denotes the corresponding k-th squared singular value. The value $WSQ(i,j)$ is the (i,j)th entry of the dissimilarity matrix, D.

We now outline the distance measures as adopted in Antoniadis et al. (2013) and Rouyer et al. (2008). Both approaches hinge on the singular vectors and leading patterns for each time series pair. Specifically, Antoniadis et al. (2013) compared the time evolution of each pair of leading patterns. In particular, for the k-th pair of leading patterns corresponding to time series $\{X_{t,T}^{(i)}\}_{t=0}^{T-1}$ and $\{X_{t,T}^{(j)}\}_{t=0}^{T-1}$, the authors take the first difference (Δ) and measure energy by means of its modulus

$$d_k(i,j) = |\Delta(P_k^{(i)} - P_k^{(j)})|. \tag{12}$$

Finally, the most significant p directions are aggregated using a weighted combination with weights given by the squared singular values:

$$D(i,j) = \frac{\sum_{k=1}^{p} \lambda_k d_k^2(i,j)}{\sum_{k=1}^{p} \lambda_k}. \tag{13}$$

The last comparison metric is

$$DT(i,j) = \frac{\sum_{k=1}^{p} \lambda_k (RD(P_k^{(i)}, P_k^{(j)}) + RD(\mathbf{u}_k^{(i)}, \mathbf{u}_k^{(j)}))}{\sum_{j=1}^{p} \lambda_k}, \tag{14}$$

where $\mathbf{u}_k^{(i)}$ and $\mathbf{u}_k^{(j)}$ are the k-th singular vectors of $X_{t,T}^{(i)}$ and $X_{t,T}^{(j)}$ respectively, and RD denotes the measure from Rouyer et al. (2008), adapted from Keogh and Pazzani (1998). This metric compares two vectors by measuring the angle between each pair of corresponding segments (a segment is defined as a pair of consecutive points of a vector) and is a method for measuring parallelism between curves. The overall distance is then computed as a weighted mean of the distance for each of the p pairs of leading patterns and singular vectors retained (with the weights being equal to the amount of covariance explained by each axis).

Note that in the simulation study (Section 4), when comparing our method with the methods outlined in Antoniadis et al. (2013) and Rouyer et al. (2008), we cluster the data using their specified time-scale decomposition and distance measure.

3.4.2. Determining the number of principal components to retain. Recall the aim to reduce the dimensionality of our problem; for each of the distance metrics above, we must decide how many axes, p, to retain. Antoniadis et al. (2013) and Rouyer et al. (2008) both decided to use the number of axes that correspond to a fixed percentage of the total covariance (as is common in principal components analysis). A different approach is to select the number of components based on a screeplot. This displays the proportion of variance explained by the (ordered) eigenvalues, and p is then selected by looking for an elbow in the screeplot. Cho et al. (2013) proposed selecting this value based on the dimension of the correlation between two curves, r.

They showed that retaining r principal components gave a good approximation and also provided a method of estimating the correlation dimension using an information criterion. We do not adopt the method of Cho et al. (2013) in this work. Instead, we choose to select the number of components either based on a screeplot or by retaining the number of axes that correspond to a fixed percentage of the total covariance, as these two methods carry less computational burden.

3.4.3. Determining the number of clusters. One of the most difficult tasks in clustering is determining the number of clusters (Antoniadis et al., 2013). This can be informed through a number of statistical techniques (Kaufman and Rousseeuw, 2009) as well as by scientific expert knowledge. For example, the 'elbow method' examines the percentage of variance explained as a function of the number of clusters; the number of clusters is then chosen by looking for an elbow in the plot of this function. Tibshirani et al. (2001) developed this methodology by estimating the number of clusters in a dataset via the gap statistic. Alternatively, the 'silhouette method' (Rousseeuw, 1987) can be used. The 'silhouette' of a data point is a number between -1 and 1, with values of 1 indicating correct clustering, and optimization techniques are then used to determine the number of clusters that gives rise to the largest 'silhouette' (Kaufman and Rousseeuw, 2009).

3.4.4. Proposed LSW-PCA clustering algorithm. Our proposed clustering method, which we shall refer to as LSW-PCA clustering, is outlined in Algorithm 1 below. We perform a partitioning around medoids (PAM) that admits a general dissimilarity matrix as input and is known to be more robust than other alternatives such as k-means (Antoniadis et al., 2013). Each of the proposed choices, i.e. spectral information, number of principal components retained (p) and distance measure, are informed by the findings of the simulation study (see Section 4 and Appendix C).

Algorithm 1 Proposed LSW-PCA clustering algorithm

Assume that each of the N observed (e.g. circadian) signals is a realisation of a locally stationary LSW process $\{X_{t,T}^{(i)}\}_{t=0}^{T-1}$, with $i = 1, 2, \ldots, N$.

1. *Spectral estimation*: estimate the spectral content of each process by using a model-based LSW corrected estimator and aggregate all information in a matrix (see Section 3.3).
2. *Dimension reduction*: achieve dimension reduction by projecting the spectral information of each process in a functional principal component space and obtain the scores associated to each signal. The number of principal components retained (p) is decided by means of the screeplot of percentage variance explained (see Section 3.4.2).
3. *Spectral distance matrix*: quantify the spectral differences between two signals by using the (weighted) squared quadratic distance measure (see Section 3.4.1).
4. *Cluster the data*: by performing a partitioning around medoids (PAM) with the distance matrix above as input.

4. Simulation study. The goals of our simulation study are twofold. First, we investigate the impact of the wavelet information choice (e.g. wavelet coefficients versus model-based spectral estimate), distance measure choice and methods to determine the number of principal components to retain. Secondly, we assess the comparative performance of our proposed procedure with other methods. Since our work

is motivated by an application in the field of circadian biology, we have designed our simulated scenarios to display typical characteristics of circadian rhythms and also to reflect the limitations of empirical work in the life sciences, where the resolution and length of the time series would be limited in practice.

4.1. Simulated data. The basic structure of each simulated experiment can be described as follows. A dataset of $N = 100$ (50 simulations from each of the two groups) was generated using the LSW representation (see equation (2)) with Daubechies' extremal phase wavelet with one vanishing moment and a Gaussian orthonormal increment sequence with mean zero and unit variance (the `locits` R package was used). Each periodogram was level smoothed by log transform, followed by translation invariant global universal thresholding and then the inverse transform was applied. For each scale of the wavelet periodogram, only levels 3 and finer were thresholded. Using the estimated spectral information, we obtained a dissimilarity matrix for each of the methods under investigation. This matrix was the input of a PAM algorithm (performed in the `cluster` R package) which clustered the data into two groups. We then compared the clusters with the known group memberships and recorded the correctly clustered percentage. The above procedure was then repeated 100 times and the results for each method were averaged.

Case 1: Defined spectra. For this study, we assume each time series is a realisation from one of $g = 1, 2$ possible groups, each with different spectral characteristics. Define the evolutionary wavelet spectrum of each group $\{S_j^{(g)}(z)\}_{j=1}^{J}$ with $J = \log_2(T)$ for all $z \in (0, 1)$ and $T = 64$ by

$$S_j^{(1)}(z) = \begin{cases} 4\cos^2(4\pi z), & \text{for } j = 2, z \in (1/64, 16/64) \\ 4\cos^2(2\pi z), & \text{for } j = 3, z \in (17/64, 1) \\ 0, & \text{otherwise;} \end{cases} \tag{15}$$

and

$$S_j^{(2)}(z) = \begin{cases} 4\cos^2(2\pi z), & \text{for } j = 2, z \in (17/64, 1) \\ 4\cos^2(4\pi z), & \text{for } j = 3, z \in (1/64, 1/2) \\ 0, & \text{otherwise;} \end{cases} \tag{16}$$

The choice above encompasses changes in amplitude and period through time, akin to those of interest to the circadian biologist. Figure 2 provides a visualisation of the wavelet spectra above (top row) and an example of a signal realisation from each of the two groups (bottom row).

Case 2: Gradual period change. For our second study, we assume each time series is a realisation from one of 3 possible groups, each with different spectral characteristics. In particular, each group represents a time series that gradually changes period from 24 to: 25 (Group 1), 26 (Group 2) and 27 (Group 3) over (approximately) two days, before continuing with the relevant period for a further two days. The purpose of this simulation study is to replicate a typical circadian experiment with changes that could not be captured by standard analyses that assume stationarity and report an average period value. Therefore, we will take $T = 256$ which is equivalent to a free-running period of 4 days with equally spaced observations every 22.5 minutes. Figure 3 shows the wavelet spectra which represent the gradually changing periods

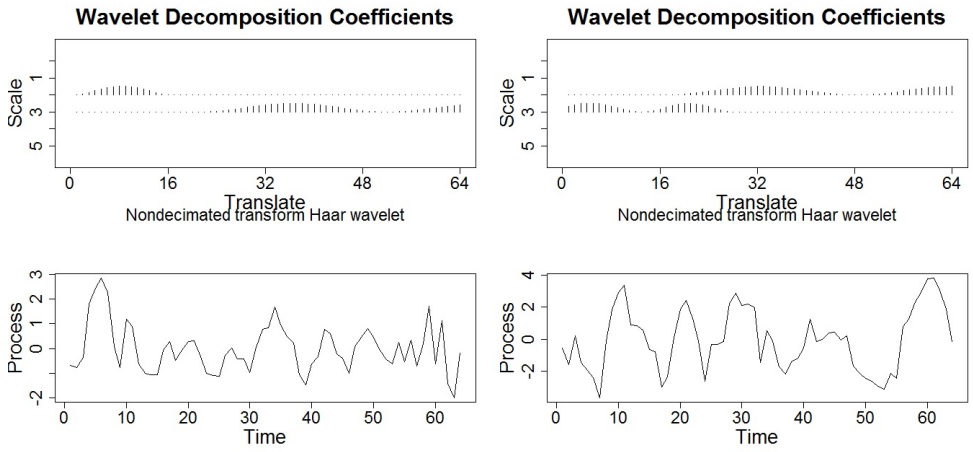

FIG. 2. *Case 1. Top left: Group 1 wavelet spectrum; Top right: Group 2 wavelet spectrum; Bottom left: Group 1 realisation and Bottom right: Group 2 realisation.*

that define each of the 3 groups above. Notice that the increased period is shown by the movement up through the resolution levels and the gradual increase in period of the wavelet coefficients. To determine which changes can be discriminated by the methods, we perform two studies within this setting (i) Case 2A: simulations from Group 1 and Group 2, and (ii) Case 2B: simulations from Group 1 and Group 3.

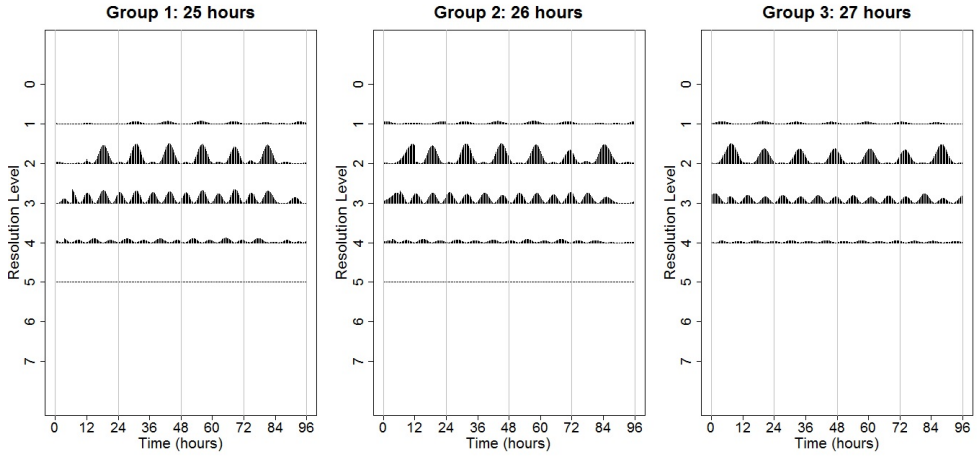

FIG. 3. *Case 2. Left: Group 1 wavelet spectrum (gradual period change from 24 to 25 hours); Centre: Group 2 wavelet spectrum (gradual period change from 24 to 26 hours); Right: Group 3 wavelet spectrum (gradual period change from 24 to 27 hours).*

Case 3: Different rates of change. For our final study, let us assume each time series is a realisation from one of 3 possible groups, each with different spectral characteristics. In particular, each group represents a time series that gradually changes

period from 24 to period 27 over 2 days (Group 1), 3 days (Group 2), 5 days (Group 3) and then continues with period 27 for the remainder of the experiment. The purpose of this simulation study is to replicate a circadian experiment with changes that could not be captured by standard analyses that assume stationarity and report an average period value. Therefore, we also take $T = 256$ which is equivalent to a free-running period of 4 days with equally spaced observations every 22.5 minutes. Figure 4 shows the wavelet spectra which represent the characteristics that define each of the 3 groups above. To determine which changes can be discriminated by the methods, we perform three studies within this setting: (i) Case 3A: simulations from Group 1 and Group 2, (ii) Case 3B: simulations from Group 1 and Group 3, and (iii) Case 3C: simulations from Group 2 and Group 3.

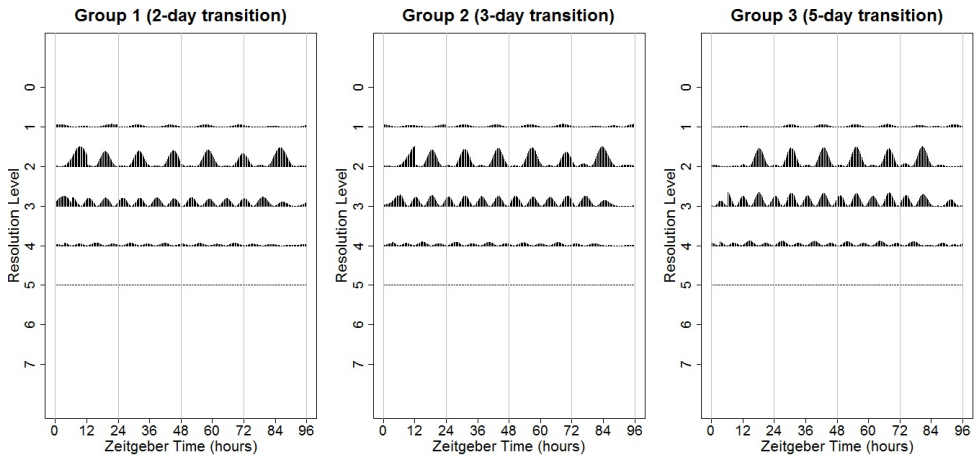

FIG. 4. *Case 3. Left: Group 1 wavelet spectrum (2-day transition); Centre: Group 2 wavelet spectrum (3-day transition); Right: Group 3 wavelet spectrum (5-day transition).*

4.2. Results. For each of our simulation studies outlined above, we investigate the impact of the wavelet information choice (e.g. wavelet coefficients versus model-based spectral estimate), distance measure choice and methods to determine the number of principal components to retain. We report our findings next, with detailed results for Case 1 presented in Appendix C.

Distance measure choice. To examine the effect of the choice of distance measure on our proposed clustering method, we performed the simulation studies as outlined above using all four distance measures defined in Section 3.4.1. We found that our method is fairly robust to the choice of distance measure, although the squared and weighted quadratic distances (SQ, respectively WSQ), appear to give superior results to the distance choices in Antoniadis et al. (2013) and Rouyer et al. (2008).

Dimension choice. We also examined the different methods outlined in Section 3.4.2 to select the number of principal components to retain for our LSW-PCA clustering method. We thus compared determining the number of principal components to retain by examining the screeplot with the situation where we retain the minimal number of components that correspond to 90% of the total covariance. Once again

we found that the LSW-PCA clustering method is robust to the way in which we choose the number of principal components to retain. Based on these results, we suggest using the LSW-PCA clustering method with the squared quadratic distance (see equation (10)), and retaining principal components by examining the screeplot. However, note that our algorithm is robust to an automatic choice based on a set percentage of the total covariance.

Wavelet information choice. In Section 3.2 we noted that other wavelet-based clustering approaches in the literature, while non-model based techniques (unlike our proposed LSW-PCA), extract the information by means of wavelet coefficients (Antoniadis et al., 2013) or squared wavelet coefficients (Rouyer et al., 2008). Therefore, using the Case 1 setting, to investigate the impact of wavelet information choice, we performed a simulation study with the following input data: original signals (thus extracting time-dependent information only), wavelet coefficients (time-scale information), squared wavelet coefficients (second-order time scale information) and finally the LSW corrected wavelet periodogram (to consistently estimate the spectrum under the LSW modelling, but without the FPCA stage). We found that clustering based on the raw data and the raw wavelet transform gave poor results (54% correctly clustered compared to 63% for squared wavelet coefficients and 69% for the corrected periodogram) which supports the assertion that clustering based on the second-moment information is preferable. Also note that using the FPCA approach further improves the results, from 69% correctly clustered to 76% (see Table 3).

Performance comparison. Finally, we compare the LSW-PCA method with the competitor methods proposed by Rouyer et al. (2008) and Antoniadis et al. (2013) (outlined in Section 3.2). Both of these benchmark methods do well in practice and represent the state-of-the-art among procedures for clustering nonstationary time series. The results are summarised in Table 3. These simulation studies provide empirical evidence that our proposed LSW-PCA method works very well and outperforms its competitors for clustering nonstationary time series. Again we see that (for this particular application) methods based on the second-order information (our LSW-PCA method and the Rouyer et al. (2008) method) perform better than the method based on the wavelet transform (Antoniadis et al., 2013). Moreover, our method, which utilises an LSW model to obtain an unbiased, consistent estimator of the underlying spectral information, performs considerably better still than the method which uses the raw wavelet periodogram. These results also show that our proposed method, which performs a FPCA on the estimated spectral coefficients of the entire dataset, outperforms the pairwise methods of Rouyer et al. (2008) and Antoniadis et al. (2013). However, note that in Cases 2A, 3A and 3C, the LSW-PCA method also has difficulty discriminating between the defined groups. These results may be due to the resolution of the data. Therefore, if the analyst predicted that a treatment effect would be characterised by this behaviour, we would recommend increasing the length of the experiment and taking observations at shorter intervals which would improve the resolution of all methods.

5. Real data analysis.

5.1. Previously published circadian data. In this section, we apply our method to an already published circadian dataset, which tested the effects of copper on plants in a method similar to our cerium dataset. Our aim is to demonstrate the additional insights provided by our proposed method. The dataset from Perea-

Sim. Study	Rouyer et al. (2008)	Antoniadis et al. (2013)	LSW-PCA Method
Case 1	66%	61%	76%
Case 2A	56%	54%	65%
Case 2B	58%	55%	76%
Case 3A	54%	54%	61%
Case 3B	55%	55%	75%
Case 3C	55%	54%	63%

TABLE 3

Comparison of the proposed LSW-PCA clustering method with the methods proposed by Rouyer et al. (2008) and Antoniadis et al. (2013) for the simulation studies. Percentages show correct clustering rates.

García et al. (2016a,b) examined circadian rhythms in high concentrations of copper as well as copper deficiency. This previously published circadian data will henceforth be referred to as the copper dataset.

The copper dataset was also obtained using a firefly luciferase reporter system as described in Appendix B. However, this experiment used a different gene of interest GIGANTEA (GI). For a detailed description of these experimental methods see Appendix D and Perea-García et al. (2016a,b). Briefly, plants were grown under different copper regimes: 'Deficiency' (no CuSO4), 'Sufficiency' or 'Control' (1 μM CuSO4), and 'Excess' (10 μM CuSO4). The copper dataset consists of a total of 74 plant signals (time series) recorded at 151 time points, with the 'Deficiency' group containing 19 plants; the 'Control' or 'Sufficiency' group, 26 plants and the 'Excess' group, 29 plants. Perea-García et al. (2016a) conducted an analysis in BRASS (see Section 2.2) and concluded that the period did not seem to be affected by copper deficiency or excess. In particular, the average period estimates for each group were reported not statistically significantly different. Therefore, it was concluded that changes in available copper were not readily detected by BRASS, even though qualitative differences were easily noted. These findings provide supportive evidence that more statistically advanced approaches are needed to analyse these types of data.

We analysed the circadian copper data by means of the proposed LSW-PCA clustering method (outlined in Algorithm 1) to establish and characterise the effect copper has on GI within the *Arabidopsis* circadian clock. As the LSW model is underpinned by wavelets and requires the data to be of dyadic length ($T = 2^J$), in our analysis we chose a segment of length $T = 128$ out of the copper dataset. This truncation was decided upon after consultation with the experimental scientists, who confirmed that the selected segments contained the times during which the plant transfered from entrained cycles into 'free-running conditions' (constant light). Figure 5 shows each individual luminescence time series from each treatment group (in grey) along with the group average (in bold) for our truncated demeaned dataset. The average of the 'Control' group is also shown in (dashed) black in each plot for comparison. For each plant we estimated the wavelet spectrum by means of the corrected wavelet periodogram estimate (with the same setting as described in the simulation study). After examining the screeplot, and for ease of interpretation, we retained two principal components to use for clustering. Using a dissimilarity matrix obtained by computing the squared quadratic distance between the first two scores of each time series, the proposed LSW-PCA clustering method yielded the results detailed in Table 4.

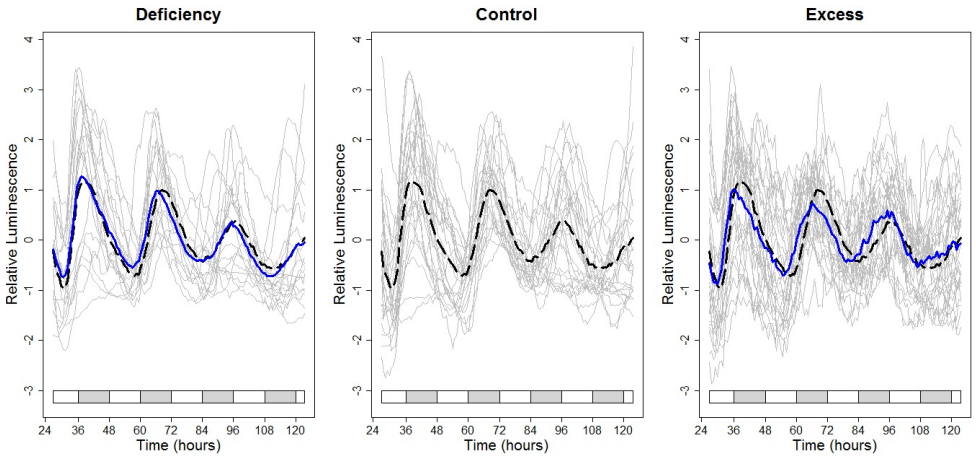

FIG. 5. *Luminescence evolution over time for plants subjected to a control and 2 different copper regimes. Time is measured in hours relative to zeitgeber time (time of last external temporal cue: the dawn signal of lights-on). Centre: Each plant signal from the* ***'Control'*** *group (in grey) along with the group average (dashed black). Other panels: Each realisation from the groups (in grey) along with the group average (in blue) and the control group average (dashed black). Left:* ***'Deficiency'*** *Group (1/2 MS). Right:* ***'Excess'*** *group (10 μM $CuSO_4$). (Each time series has been normalised to have mean zero.) The grey and white bars indicate the subjective night and day, respectively.*

Number of plants	Deficiency	Control	Excess	Total
Cluster 1	**11**	**14**	13	38
Cluster 2	8	12	**16**	36
Total	19	26	29	74

TABLE 4

Results of clustering the copper dataset into two clusters using the proposed LSW-PCA method. The modal cluster for each copper regime is highlighted in bold.

In determining the optimal number of clusters, we used the 'elbow method' and then validated this result via the 'silhouette method' (implemented in the `fpc` R package) and consultations with experimental scientists, as outlined in Section 3.4.3. All approaches indicated that we should cluster the data into two groups, which suggests the presence of two distinct groups within this dataset, each with different time-frequency behaviour. This is in contrast to the results in Perea-García et al. (2016a), which found no detectable difference in period. This illustrates the point in Section 2.4, that although plants in each treatment group share identical genetic characteristics and have been treated in identical conditions, they can respond differently and average behaviour assessment can mask these differences.

Discussion of findings. On examining Table 4, we can see that the LSW-PCA clustering method has clustered the behaviour of the data into the following two groups: Cluster 1 identifies similar behaviour of plants in the 'Control' and copper 'Deficiency' groups, and Cluster 2 is the modal cluster of the copper 'Excess' group. These results are in agreement with Figure 5 which provides visual evidence that the

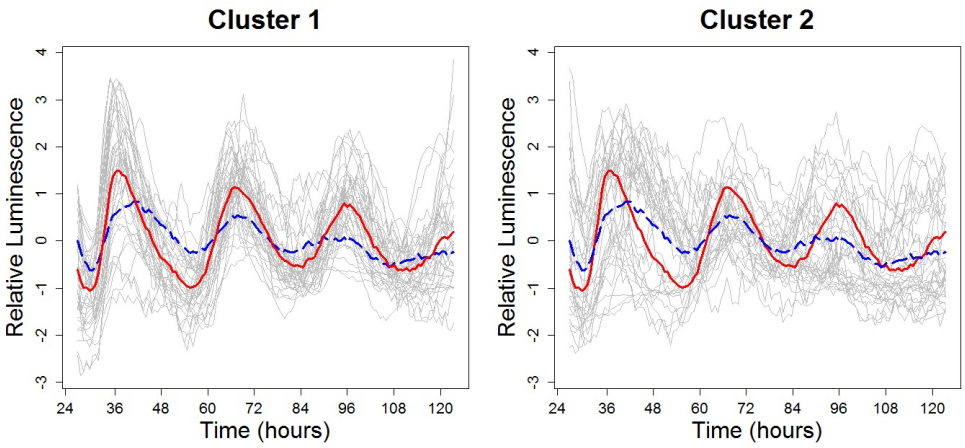

FIG. 6. *Results of clustering the copper dataset into two clusters using the proposed LSW-PCA method. The individual signals (grey) along with the cluster average in: red for Cluster 1 and (dashed) blue for Cluster 2.*

plants in the copper 'Excess' group seemed to display distinct behaviour from the other groups. However, the Cluster 2 'Excess' behaviour can also be seen in some plants in the other two groups, particularly in the 'Control' group. The presence of 'Control' and 'Deficiency' treated plants in the cluster associated mostly with 'Excess' levels of copper, highlights individual-level variability in plant response to stimuli, despite their sharing identical genetic characteristics (Doyle et al., 2002). This result may be due to the individual plants in some instances showing a stress response, particularly those individuals from the 'Deficiency' group in Cluster 2. Alternatively, this may be due to stress induced by the experimental method itself. Thus, although both types of behaviour are present in each treatment group, increased levels of copper increase the likelihood of a Cluster 2-type response.

Our proposed method also allows us to characterise the behaviour associated with each cluster. The signals within each cluster are shown (in grey) along with the cluster average (in bold) in Figure 6. Figure 7 shows the final cluster each individual time series was assigned to: the individual signals are plotted in red for Cluster 1 and blue for Cluster 2, for each treatment group. The cluster estimated average spectra appear in Figure 8.

Note in Figure 6 that Cluster 1 is characterised by a gradual increase in period throughout the experiment and gradual amplitude dampening with time. The amplitude dampening can also clearly be seen in the decreasing coefficients in resolution levels 2–4 (and particularly in level 2) in the average spectrum of Cluster 1 in Figure 8. The gradual increase in period can be seen as the activity in the spectrum begins in resolution level 4 and moves into levels 3 and 2 with time.

Cluster 2 is characterised by low frequency behaviour throughout the experiment (a longer period) and marked amplitude dampening with time, resulting in a rhythmicity loss. Indeed, this behaviour is also identified by the average spectrum in Figure 8. The increased period is reflected in the large coefficients at coarsest levels and the increased period of the wavelet coefficients in resolution levels 2 and 3. The dampening is apparent as the magnitude of the spectral coefficients decreases as time progresses.

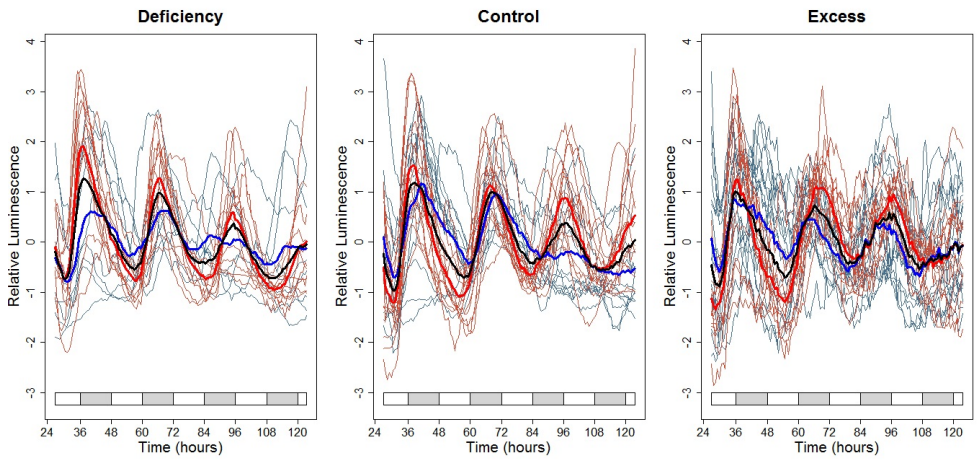

FIG. 7. *Results of clustering the copper dataset into two clusters using the proposed LSW-PCA method. For each treatment group the individual signals are plotted in: red for Cluster 1 and blue for Cluster 2. The average of each treatment group is shown in black. Within each treatment group, the Cluster 1 average is shown in bold red and the Cluster 2 average in bold blue.*

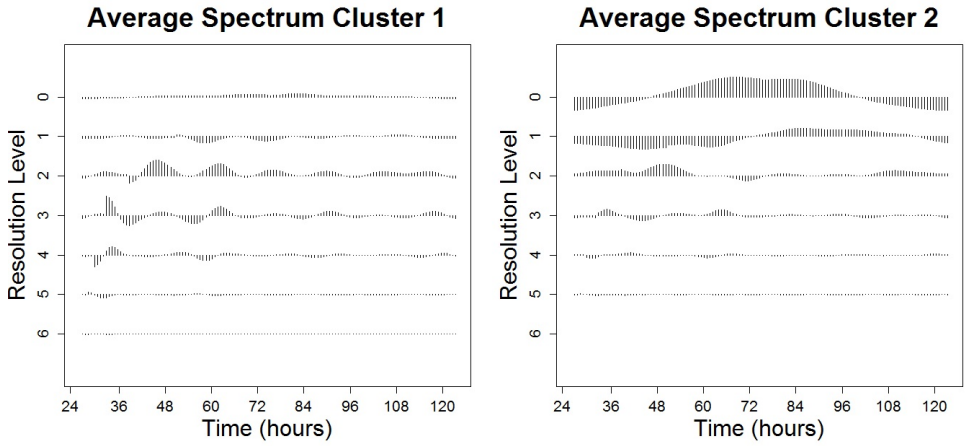

FIG. 8. *Cluster average estimated spectra on the copper dataset using the proposed LSW-PCA method.*

Furthermore, note the nonstationary behaviour that characterises both clusters (changing period and amplitude). The presence of these nonstationary characteristics supports our assertion that the existing methods (which assume stationarity) are inappropriate for such datasets and cannot capture this behaviour. Figure 7 shows that, although all plants in each treatment group share identical genetic characteristics and have been treated in identical conditions, they respond in two different ways. Note that the treatment group averages (in black) lie between the two (within treatment group) cluster averages. This is particularly noticable in the 'Deficiency' group. Therefore, the presence of both types of behaviour in each of the original treatment

Number of plants	Hoagland's	100 μM	150 μM	200 μM	Total
Cluster 1	**13**	2	3	0	18
Cluster 2	6	**14**	0	0	20
Cluster 3	5	8	**21**	**24**	58
Total	24	24	24	24	96

TABLE 5

Results of clustering the (normalised, truncated) cerium dataset into three groups using the proposed LSW-PCA method. The modal cluster for each concentration is highlighted in bold.

groups has resulted in similar average behaviour.

In conclusion, our LSW-PCA clustering method has detected and characterised the interesting effects excess levels of copper have on the circadian clock, that were not detectable in the original analysis of the copper dataset (Perea-García et al., 2016a).

5.2. Novel circadian plant data. We now return to the circadian data that motivated this work and apply our proposed LSW-PCA clustering method to analyse the novel cerium data. As the LSW model is underpinned by wavelets and requires the data to be of dyadic length ($T = 2^J$), in our analysis we chose a segment of length $T = 128$ out of the original dataset. This truncation was decided upon after consultation with the experimental scientists, as in Section 5.1. For each plant we estimated the wavelet spectrum by means of the corrected wavelet periodogram estimate (with the same setting as described in the simulation study in Section 4). On examining the screeplot (see Figure S4 in Appendix A) and for ease of interpretation, we retained two principal components to cluster the data. The proposed LSW-PCA clustering method yielded the results detailed in Table 5.

The methods outlined in Section 3.4.3 were used to determine the optimal number of clusters. All methods indicated that we should cluster the data into three groups. This was supported by experimental scientists who confirmed that it would be useful to cluster the data into three groups: 'No Change' and two distinct departures from this group. In particular, we hoped to differentiate between and characterise the effects of lower and higher concentrations of cerium. This is because recent research has shown that certain compounds can produce very different effects on plant growth at low and high doses (Yang et al., 2016). Furthermore, this phenomenon seems to be present in our circadian dataset. On examining Figure 1, it appears that plants subjected to higher concentrations of cerium (150μM and 200μM) seem to exhibit similar behaviour, while the control group and concentration 100μM seem to display average behaviour which is distinct from each other and from the higher concentrations.

Discussion of findings. On examining Table 5, we can see that this method has effectively clustered the behaviour of the data into the following three groups:

1. Cluster 1: contains mostly plants in the Control dataset (Hoagland's), and very few plants subjected to lower-medium concentrations of ammonium cerium nitrate (100μM and 150μM)– conceptualised as essentially 'Control';
2. Cluster 2: contains mostly plants with lower concentration of ammonium cerium nitrate (100μM) and a few plants from the Control dataset– conceptualised as 'Low concentration';
3. Cluster 3: identifies similar behaviour to plants mostly exposed to medium-high concentrations (150μM, 200μM), but interestingly also contains a few plants from the Control and 100μM concentration.

These results are in agreement with Figure 1 (which we recall provided visual evidence that the plants subjected to higher concentrations of cerium exhibit similar behaviour, while the control group and concentration 100μM seem to display distinct behaviour). Therefore, this analysis has enabled us to achieve our first goal: to differentiate between the effects of lower and higher concentrations of cerium. Of interest to circadian biologists, however, is the presence of control and low concentration treated plants in the group associated mostly with higher concentrations. This highlights individual-level variability in plant response to stimuli, despite their sharing identical genetic characteristics (Doyle et al., 2002).

Our proposed method also allows us to characterise these groups, both in terms of first and second-order plant behaviour. The signals within each clustered group are shown (in grey) along with the cluster average (in bold) in Figure 9, while the cluster estimated average spectra appear in Figure 10.

On examining Figure 9, notice the different behaviour of Cluster 3 from the other clusters– this effect is characterised by high frequency behaviour throughout the experiment and a marked amplitude dampening with time, resulting in a rhythmicity loss. Indeed, this behaviour is also identified by the average spectrum in Figure 10. The high frequency behaviour is reflected in the large coefficients in resolution level 6. The dampening is apparent as the magnitude of the spectral coefficients decreases as time progresses (particularly in resolution level 2).

In contrast, Clusters 1 and 2 (approximately corresponding to the control and low concentration groups respectively) display more similar, rhythmic behaviour. On examining Figure 9, the rhythmic periods of the cluster averages seem approximately equal. However, there are also clear differences between the two groups. Firstly, there is a difference in the amplitudes of the two cluster averages. Cluster 1 has a larger peak at approximately $t = 36$ and an even larger peak at $t = 120$. This can be seen in the large coefficients around these time points in resolution levels 1-4 in the average spectrum of Cluster 1. Alternatively, Cluster 2 seems to have a very large peak at $t = 36$ followed by a distinct reduction in the amplitude of the other peaks. This can also be seen in the large coefficients in resolution levels 2-4 in the average spectrum of Cluster 2 in Figure 10.

The spectral content extracted in the first two principal components can be found in Figure 11. The projection of the original plant signals onto the principal component plane appears in Figure 12, by cluster and group membership. These indicate that the first principal component represents the departure from the control group after exposure to ammonium cerium nitrate, with larger values indicating a distinct change. The second principal component appears to reflect the spectral behaviour of the 100μM group, in particular the larger amplitude at around $t = 36$. Finally, note that Figure 12 shows that Cluster 1 has the biggest spread, while Cluster 3 is the most tightly packed. This supports biological expectations that plants behave in a similar manner when 'under stress' (Hanano et al., 2006).

6. Conclusions and Further Work. In this manuscript, we have developed a new procedure for clustering inherently nonstationary rhythmic data by modelling them as locally stationary wavelet processes and exploiting their local time-scale spectral properties by means of a functional principal component analysis. Our method combines the advantages of a wavelet analysis with the benefits of rigorous stochastic nonstationary time series modelling and has desirable properties, such as low sensitivity to the choice of distance measure and number of principal components to retain. These characteristics show the method's suitability in organising and under-

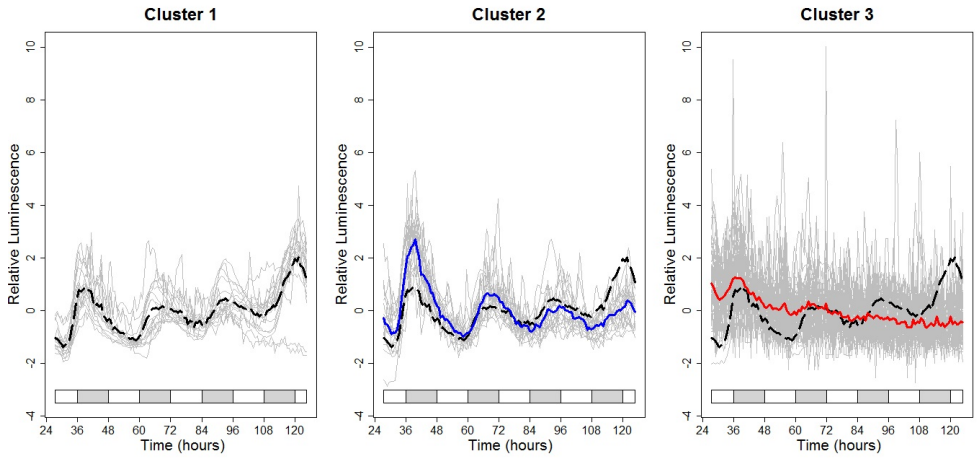

FIG. 9. *The results of clustering the cerium dataset into three groups using the proposed LSW-PCA method. The individual signals (grey) along with the cluster average in: (dashed) black for Cluster 1; blue for Cluster 2 and red for Cluster 3. The average of Cluster 1 (conceptualised as essentially 'Control') is shown (in dashed black) in all plots for reference.*

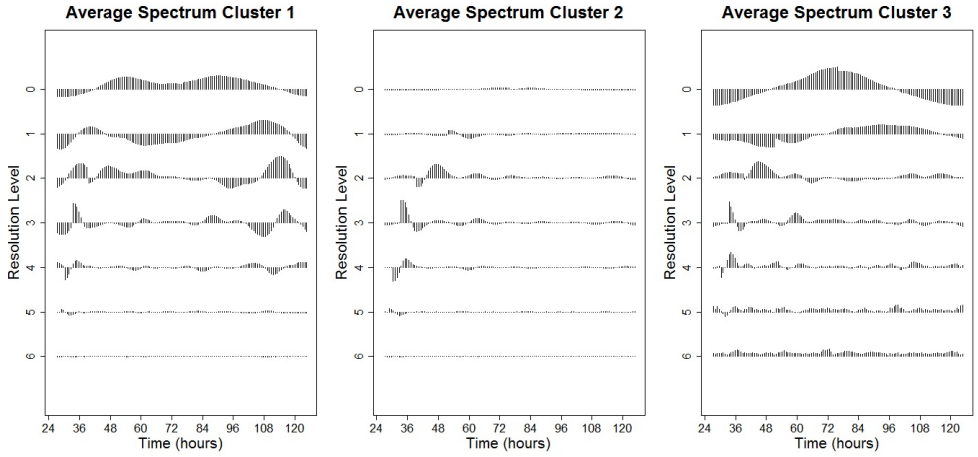

FIG. 10. *Cluster average estimated spectra on the cerium dataset using the proposed LSW-PCA method. Cluster 1 approximately corresponds to the 'Control' group; Cluster 2 depicts 'Low concentration' behaviour (100 μM) and Cluster 3 the 'Higher concentration' (150 μM and 200 μM).*

standing multiple nonstationary time series, such as the gene expression levels in our novel circadian dataset. When compared to competitor (non-model based) methods, we found that our methodology brought clear gains for simulated data (Table 3). Furthermore, when compared to existing methods (which assume stationarity), the LSW-PCA clustering method also displayed advantages for real data (Table 5).

The proposed model-based clusterings can be used to answer questions such as, 'What other concentrations of this compound produce similar effects in plants?' Our approach can also produce visualisations helpful in answering questions such as, 'What

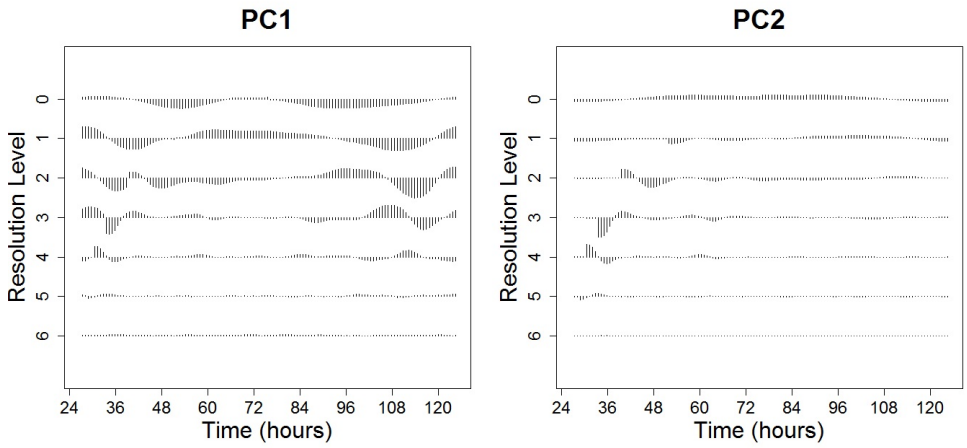

FIG. 11. *First two principal components obtained using the proposed LSW-PCA method on the cerium dataset.*

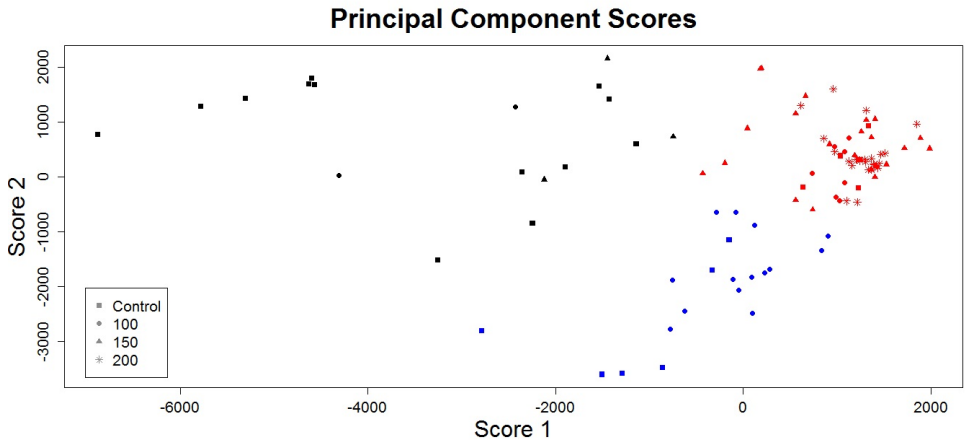

FIG. 12. *The cerium dataset projected onto the first two principal components obtained from the LSW-PCA clustering method. The colours represent the clusters: black for Cluster 1, blue for Cluster 2 and red for Cluster 3. The symbols represent the plant treatments.*

characterises the different types of reactions present in this dataset?' Such answers have important implications for understanding the mechanism of the plant's circadian clock and also environmental implications associated with soil pollution.

Also note that our proposed algorithm is not restricted to the datasets analysed in this paper; it can be applied to other circadian datasets, as well as to data originating in other fields. The flexibility and computational efficiency of our approach allows more global analyses of plant behaviour to be undertaken which would not be possible within the stationary statistical constraints underlying traditional methods of period estimation. For example, the roles of a wide range of soil pollutants can be assessed within a single statistical framework. By extending this statistical methodology and

empirical protocol to include exposure to other compounds, one could address the question, 'Which other elements in the periodic table, and at which concentrations, produce similar kinds of reactions in plants?' We can also extend the dataset to include plants with deficiencies of elements other than copper. These studies would also enable deeper understanding of the circadian clock mechanisms and its adaptations to change (Perea-García et al., 2016a).

The wavelet system gives a representation for nonstationary time series under which we estimate the wavelet spectrum and subsequently cluster the data. Ideally, we would envisage the use of the wavelet that is best suited to modelling and discriminating between the particular dataset. In simulations we found our method to be fairly robust to the wavelet choice. An area of further work would be to derive a procedure for determining which wavelet system to adopt for any given dataset.

We are aware of the propensity of the recording equipment (see Appendix B) to break down, resulting in gaps in the data. Such failures in hardware are an objective reality of empirical work in the life sciences, and another area of future work is to adapt current methods under the presence of missingness, or 'gappy' data, often arising in experimental data. This estimate could then be used as a classification signature or within our clustering procedure.

Appendix A. Supplementary Figures. In this section we offer visual evidence to support claims in Sections 1, 2 and 5 of the main article. All figures (S1, S2, S3 and S4) are referred to in context as part of the main body of the paper.

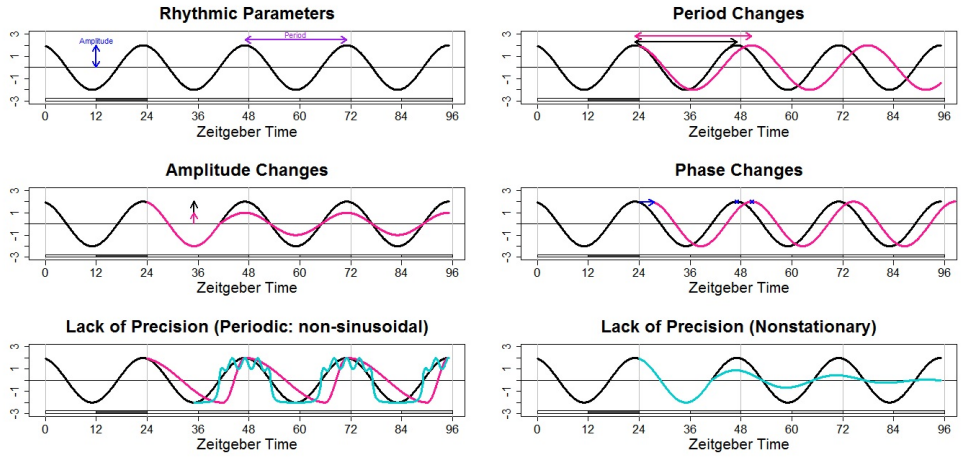

FIG. S1. *The defined rhythmic parameters: periodicity, phase, amplitude and clock precision (based on an image from Hanano et al. (2006)).*

Appendix B. Experimental Details: Novel Circadian Plant Data. In this section we outline the experimental details that led to the novel circadian plant rhythms under analysis (Section 2.1 of the main paper).

To obtain this dataset, the Davis Lab (Biology, University of York) used a firefly luciferase reporter system. This method uses a fusion of the gene of interest to luciferase. In this experiment, the gene of interest was 'cold and circadian regulated and RNA binding 2', known as CCR2 (further details of *CCR2:LUC* can be found in Doyle et al. (2002)). When CCR2 is expressed, luciferase is produced, causing

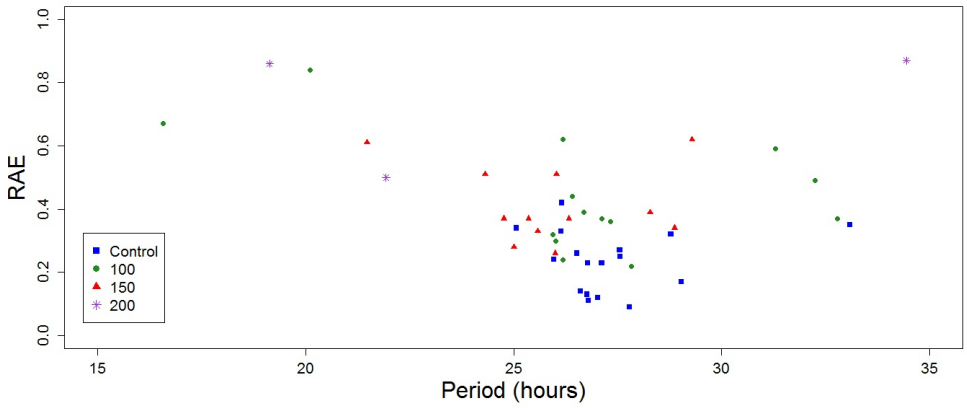

FIG. S2. *Summary of the BRASS analysis of the circadian plant signals in response to differing quantities of ammonium cerium nitrate, represented by plots of period estimates plotted against the respective relative amplitude errors (RAE). The colours and symbols represent the plant treatment groups: blue squares for the Control Group; green circles for Group 1 (*$100\mu M$*); red triangles for Group 2 (*$150\mu M$*) and purple stars for Group 3 (*$200\mu M$*).*

the plant to produce quantifiable levels of light. This bioluminescence was measured using a TopCount NXT scintillation counter (Perkin Elmer), allowing relative gene expression of CCR2 to be quantified *in vivo* (Plautz et al., 1997; Southern and Millar, 2005; Perea-García et al., 2016a). These experiments were carried out using the following methods: *Arabidopsis thaliana* seeds (Ws–*CCR2:LUC*) were surface sterilised and plated onto Hoagland's media containing 1% sucrose, 1.5% phyto agar (Hoagland et al., 1950). The seeds were stratified for 2 days at 4°C and transferred to growth chambers to entrain under 12:12 light/dark cycles at a constant temperature of 20°C. These conditions were chosen to simulate the 'normal' light/dark cycles of a day. Six-day-old seedlings were transferred to 96 well microtiter plates containing Hoagland's 1% sucrose, 1.5% agar (Southern and Millar, 2005) also containing supplemental $(NH_4)_2Ce(NO_3)_6$ (ammonium cerium nitrate) at a concentration of 100μM, 150μM or 200μM. The plants were then transferred to the TOPCount machine. Measurements were taken at intervals of approximately 45 minutes. Measurement began after the transition to 12 hours of darkness (known as subjective dusk) on the seventh day of the plants' life. Therefore, the plant experiences one 'normal' day in the TOPCount machine (known as entrainment). After this, the plant was exposed to constant light (known as an LL free-run) for approximately four days. In Figure 1, the shaded bars below the graph represent the light conditions the plants would experience during the 'normal' day. The plants are under constant light throughout the experiment, however, the grey bars indicate that they would be in darkness during a 'normal' 12 hour light/12 hour dark cycle.

Our dataset therefore consists of a total 96 plant signals (time series) recorded at 128 time points, with each of the control and groups 1–3 (each corresponding to a different concentration of ammonium cerium nitrate) containing 24 plants. In particular, the control group is grown in Hoagland's media (Hoagland et al., 1950) which contains essential nutrients required for plant growth and is not exposed to any

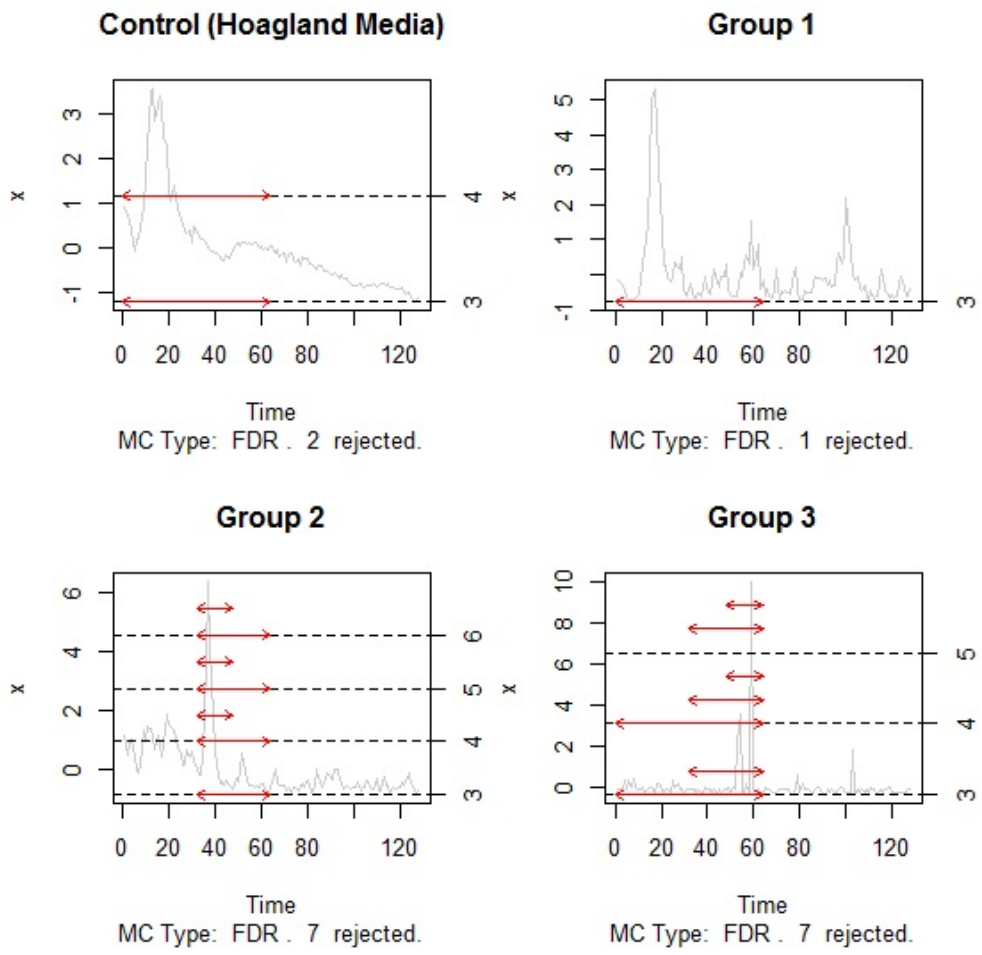

FIG. S3. *Plots of the estimated locations of the nonstationarities in the circadian plant signals in response to differing quantities of ammonium cerium nitrate, using the wavelet spectrum test (Nason, 2013), implemented in the* `locits` *package in R which is available on CRAN. A time series for each of the four groups is shown as an example– Group 1, a time series from the* $100\mu M$ *group; Group 2, a time series from the* $150\mu M$ *group; Group 3, a time series from the* $200\mu M$ *group.*

additional levels of ammonium cerium nitrate. To examine the effects of cerium on the circadian clock, the other three groups, while also grown in the Hoagland's media, were additionally exposed to varying additional concentrations of ammonium cerium nitrate– 100μM for Group 1, 150μM for Group 2 and 200μM for Group 3.

Appendix C. Results of Simulation Study Case 1. In this section we report the findings of the simulation study associated to Case 1 in Section 4.1 of the main paper. These consist of Tables S1 and S2, which further justify the distance and dimension reduction choices adopted for our proposed method.

Appendix D. Experimental Details: Previously Published Circadian Data. In this section we outline the experimental details that led to the previously

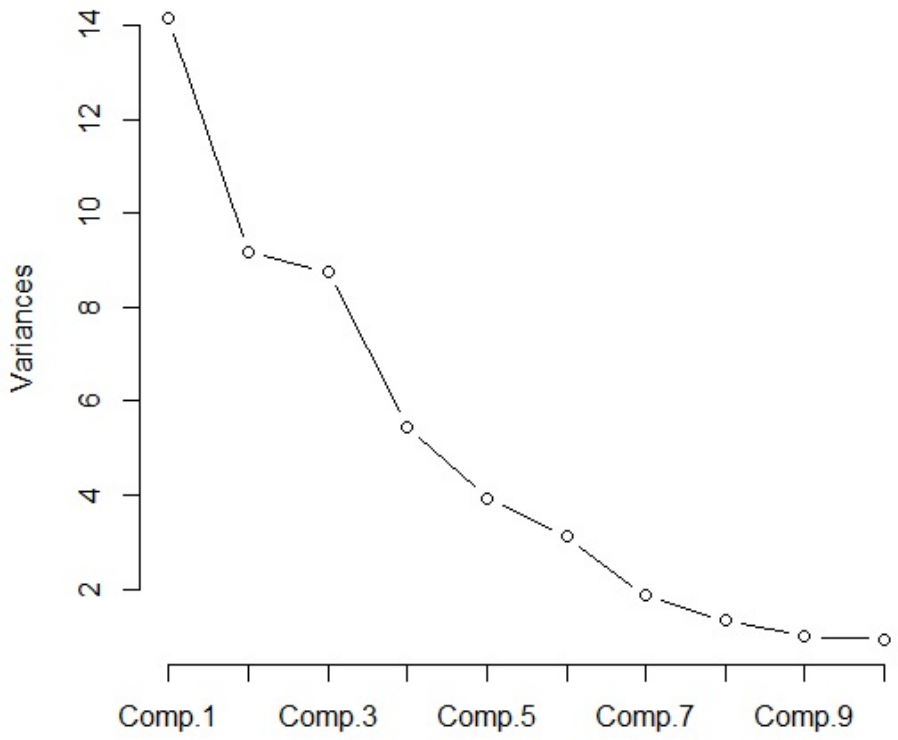

FIG. S4. *The screeplot used to inform the selection of the number of principal components to retain for the cerium dataset.*

Distance Measure	SQ	WSQ	DT	D
Correctly Clustered (%)	76%	70%	69%	65%

TABLE S1
Case 1. Distance measure (Section 3.4.1) comparison for the proposed LSW-PCA method.

published copper dataset (Section 5.1 of the main paper).

This dataset (Perea-García et al., 2016a,b) was also obtained using a firefly luciferase reporter system as described in Appendix B. Experimental Details: Novel Circadian Plant Data. However, this experiment uses a different gene of interest GIGANTEA (GI). Plants were grown on plates as described in Andrés-Colás et al. (2010), incubated on MS (Murashige and Skoog) medium (Murashige and Skoog, 1962) at half concentration (1/2 MS) [phytoagar 0.8% (w/v) plus 1% sucrose (w/v) in 0.5% MES (w/v)]. WS GI:LUC seedlings were grown under different copper regimes:

Dimension reduction method	90% of total covariance	Screeplot
SQ distance	73%	76%
WSQ distance	69%	70%
DT distance	54%	69%

TABLE S2

Case 1. Comparison for selection of principal components for proposed LSW-PCA clustering method. Percentages show correct clustering rates.

'Deficiency' (1/2 MS), 'Sufficiency' or 'Control' (1 μM CuSO4), and 'Excess' (10 μM CuSO4). 96 plants were grown in total, 32 under each copper regime. The plants were entrained for 7 days under 12:12 light-dark cycles at a constant temperature of 20°C. The plants were then exposed to constant light (LL free-run) for the remainder of the experiment. Bioluminescence was then measured every hour using the same TopCount NXT system as in Appendix B.

The dataset analysed in Perea-García et al. (2016a,b) consists of a total 74 plant signals (time series) recorded at 151 time points. Plants with an average luminescence of 40 or below were excluded prior to analysis as luminescence values below this are considered background noise. Therefore, the 'Deficiency' group (1/2 MS) contains 19 plants; the 'Control' or 'Sufficiency' group (1 μM CuSO4) contains 26 plants and the 'Excess' group (10 μM CuSO4) contains 29 plants.

References.

Andrés-Colás, N., Perea-García, A., Puig, S., and Peñarrubia, L. (2010). Deregulated copper transport affects Arabidopsis development especially in the absence of environmental cycles. *Plant physiology*, 153(1):170–184.

Antoniadis, A., Brossat, X., Cugliari, J., and Poggi, J.-M. (2013). Clustering functional data using wavelets. *International Journal of Wavelets, Multiresolution and Information Processing*, 11(01):1350003.

Bell-Pedersen, D., Cassone, V. M., Earnest, D. J., Golden, S. S., Hardin, P. E., Thomas, T. L., and Zoran, M. J. (2005). Circadian rhythms from multiple oscillators: lessons from diverse organisms. *Nature Reviews Genetics*, 6(7):544–556.

Bujdoso, N. and Davis, S. J. (2013). Mathematical modeling of an oscillating gene circuit to unravel the circadian clock network of Arabidopsis thaliana. *Frontiers in Plant Science*, 4:3.

Cho, H., Goude, Y., Brossat, X., and Yao, Q. (2013). Modeling and forecasting daily electricity load curves: a hybrid approach. *Journal of the American Statistical Association*, 108(501):7–21.

Cressie, N. and Wikle, C. K. (2015). *Statistics for spatio-temporal data*. John Wiley & Sons.

Dahlhaus, R. (1997). Fitting time series models to nonstationary processes. *The Annals of Statistics*, 25(1):1–37.

Daubechies, I. (1992). *Ten Lectures on Wavelets*, volume 61. SIAM.

Doyle, M. R., Davis, S. J., Bastow, R. M., McWatters, H. G., Kozma-Bognár, L., and Nagy, Ferenc and Millar, A. J. and Amasino, R. M. (2002). The ELF4 gene controls circadian rhythms and flowering time in Arabidopsis thaliana. *Nature*, 419(6902): 74–77.

Edwards, K. D., Akman, O. E., Knox, K., Lumsden, P. J., Thomson, A. W., Brown, P. E., Pokhilko, A., Kozma-Bognar, L., Nagy, F., Rand, D. A., and Millar, A (2010). Quantitative analysis of regulatory flexibility under changing environmental

conditions. *Molecular systems biology*, 6(1): 424.
Fiecas, M. and Ombao, H. (2016). Modeling the evolution of dynamic brain processes during an associative learning experiment. *Journal of the American Statistical Association*, 111:1440–1453.
Fryzlewicz, P. and Nason, G. P. (2006). Haar–fisz estimation of evolutionary wavelet spectra. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(4):611–634.
Fryzlewicz, P. and Ombao, H. (2009). Consistent classification of nonstationary time series using stochastic wavelet representations. *Journal of the American Statistical Association*, 104:299–312.
Hanano, S., Domagalska, M. A., Nagy, F., and Davis, S. J. (2006). Multiple phytohormones influence distinct parameters of the plant circadian clock. *Genes to Cells*, 11(12):1381–1392.
Harang, R. Bonnet, G. and Petzold, L. R. (2012). WAVOS: a MATLAB toolkit for wavelet analysis and visualization of oscillatory systems *BMC research notes, BioMed Central*, 5(1):163.
Hoagland, D. R. and Arnon, D. I. (1950). The water-culture method for growing plants without soil. *California Agricultural Experiment Station, Circular*, 347.
Holan, S. H., Wikle, C. K., Sullivan-Beckers, L. E., and Cocroft, R. B. (2010). Modeling complex phenotypes: generalized linear models using spectrogram predictors of animal communication signals. *Biometrics*, 66(3):914–924.
Kaufman, L. and Rousseeuw, P. J. (2009). Finding groups in data: an introduction to cluster analysis. *John Wiley & Sons*, 98:239–243.
Keogh, E. J. and Pazzani, M. J. (1998). An enhanced representation of time series which allows fast and accurate classification, clustering and relevance feedback. *Proc. of the* 4^{th} *International Conference of Knowledge Discovery and Data Mining, AAAI Press* , 98:239–243.
Krzemieniewska, K., Eckley, I. A., and Fearnhead, P. (2014). Classification of nonstationary time series. *Stat*, 3(1):144–157.
Leise, T. L.,Indic, P., Paul, M. J. and Schwartz, W. J. (2013). Wavelet meets actogram. *Journal of biological rhythms, SAGE Publications Sage CA: Los Angeles, CA*, 28(1):62–68.
McClung, C. R. (2006). Plant circadian rhythms. *The Plant Cell*, 18(4):792–803.
Minors, D. S. and Waterhouse, J. M. (2013). *Circadian rhythms and the human*. Butterworth-Heinemann
Moore, A., Zielinski, T., and Millar, A. J. (2014). Online period estimation and determination of rhythmicity in circadian data, using the BioDare data infrastructure. *Methods in Molecular Biology*, 1158:13–44.
Murashige, T. and Skoog, F. (1962). A revised medium for rapid growth and bio assays with tobacco tissue cultures *Physiologia plantarum*, 15(3):473–497.
Nason, G. P., Von Sachs, R., and Kroisandt, G. (2000). Wavelet processes and adaptive estimation of the evolutionary wavelet spectrum. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(2):271–292.
Nason, G. (2010). *Wavelet methods in statistics with R (use R)*. Springer Science & Business Media.
Nason, G. (2013). *A test for second-order stationarity and approximate confidence intervals for localized autocovariances for locally stationary time series*. Journal of the Royal Statistical Society: Series B (Statistical Methodology), Wiley Online Library, 75(5):879–904.
Ogden, T. R. (1997). On preconditioning the data for the wavelet transform when

963 the sample size is not a power of two. *Communications in Statistics-Simulation and*
964 *Computation*, 26(2):467–486.
965 Perea-García, A., Andrés-Bordería, A., de Andrés, S. M., Sanz, A., Davis, A. M.,
966 Davis, S. J., Huijser, P., and Peñarrubia, L. (2016a). Modulation of copper defi-
967 ciency responses by diurnal and circadian rhythms in arabidopsis thaliana. *Journal*
968 *of experimental botany*, 67(1):391–403.
969 Perea-García, A., Sanz, A., Moreno, J., Andrés-Bordería, A., de Andrés, S. M., Davis,
970 A. M., Huijser, P., Davis, S. J. and Peñarrubia, L. (2016b). Daily rhythmicity of
971 high affinity copper transport. *Plant signaling & behavior*, 11(3):e1140291.
972 Plautz, J. D., Straume, M., Stanewsky, R., Jamison, C. F., Brandes, C.,Dowse, H. B.,
973 Hall, J. C. and Kay, S. A. (1997). Quantitative analysis of Drosophila period gene
974 transcription in living animals. *Journal of Biological Rhythms, Sage Publications* ,
975 12(3): 204–217.
976 Price, T. S., Baggs, J. E., Curtis, A. M., FitzGerald, G. A. and Hogenesch, J. B.
977 (2008). WAVECLOCK: wavelet analysis of circadian oscillation *Bioinformatics,*
978 *Oxford University Press*, 24(23): 2794–2795.
979 Priestley, M. B. (1965). Evolutionary spectra and non-stationary processes. *Journal*
980 *of the Royal Statistical Society, Series B (Methodological)*, 27:204–237.
981 Priestley, M. and Rao, T. S. (1969). A test for non-stationarity of time-series. *Journal*
982 *of the Royal Statistical Society, Series B (Methodological)*, 31:140–149.
983 Priestley, M. B. (1982). *Spectral analysis and time series*. Academic Press.
984 Ramsay, J. O. and Silverman, B. W. (2005). *Functional Data Analysis*. Springer.
985 Rouyer, T., Fromentin, J.-M., Stenseth, N. C., and Cazelles, B. (2008). Analysing
986 multiple time series and extending significance testing in wavelet analysis. *Marine*
987 *Ecology Progress Series*, 359:11–23.
988 Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and
989 validation of cluster analysis. *Journal of computational and applied mathematics,*
990 *Elsevier*, 20:53–65.
991 Shumway, R. H. (2003). Time-frequency clustering and discriminant analysis. *Statis-*
992 *tics & probability letters*, 63(3):307–314.
993 Southern, M. M. and Millar, A. J. (2005). Circadian genetics in the model higher
994 plant, arabidopsis thaliana. *Methods in enzymology*, 393:23–35.
995 Tibshirani, R., Walther, G. and Hastie, T. (2001). Estimating the number of clusters
996 in a data set via the gap statistic *Journal of the Royal Statistical Society: Series*
997 *B (Statistical Methodology)*, 63(2):411–423.
998 Vitaterna, M. H., Takahashi, J. S., and Turek, F. W. (2001). Overview of circadian
999 rhythms. *Alcohol Research and Health*, 25(2):85–93.
1000 Yang, X., Pan, H., Wang, P. and Zhao, F. (2016). Particle-specific toxicity and
1001 bioavailability of cerium oxide (CeO2) nanoparticles to Arabidopsis thaliana. *Jour-*
1002 *nal of hazardous materials, Elsevier*, 322:292–300.
1003 Zielinski, T., Moore, A. M., Troup, E., Halliday, K. J. and Millar, A. J. (2014).
1004 Strengths and limitations of period estimation methods for circadian data. *PloS*
1005 *one, Public Library of Science*, 9(5):96462.