

This is a repository copy of *Clustering nonstationary circadian plant rhythms using locally stationary wavelet representations*.

White Rose Research Online URL for this paper:
<https://eprints.whiterose.ac.uk/123769/>

Version: Accepted Version

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>

Reuse

Items deposited in White Rose Research Online are protected by copyright, with all rights reserved unless indicated otherwise. They may be downloaded and/or printed for private study, or other acts as permitted by national copyright laws. The publisher or other rights holders may allow further reproduction and re-use of the full text version. This is indicated by the licence information on the White Rose Research Online record for the item.

Takedown

If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

1 **CLUSTERING NONSTATIONARY CIRCADIAN RHYTHMS USING**
2 **LOCALLY STATIONARY WAVELET REPRESENTATIONS***

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

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

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

20 **AMS subject classifications.** 62P10

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

33 Experiments recording plant response to light entrainment result in datasets that,
34 from a statistical point of view, can be considered as time series realisations. Period
35 and phase estimation (see Figure S1 in Appendix A for a visual interpretation of
36 this terminology) are the fundamental elements of most circadian analyses. The cur-
37 rent standard uses BRASS (Biological Rhythm Analysis Software System (Edwards
38 et al., 2010)) to estimate the period of each time series using Fourier analysis (see
39 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).

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

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

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

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

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

85 **2.1. Experimental details.** The novel circadian dataset (henceforth referred
86 to as the cerium dataset) was obtained by the Davis Lab (Biology, University of
87 York) following a similar method to Hanano et al. (2006). For a detailed description
88 of these methods see Appendix B. Briefly, for each plant, gene expression levels are

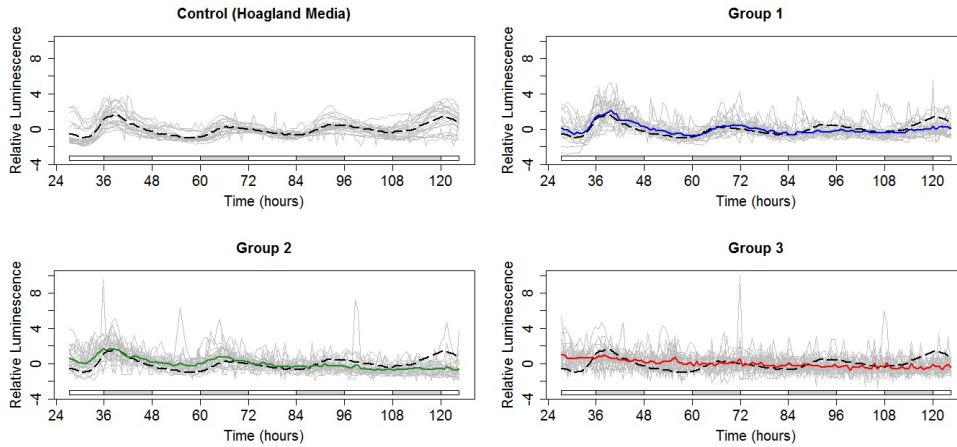


FIG. 1. Luminescence evolution over time for plants subjected to a control and 3 different ammonium cerium nitrate concentrations. Time is measured in hours relative to zeitgeber time (time of last external temporal cue: the dawn signal of lights-on). Top left: Each plant signal from the control group (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: $100\mu\text{M}$ ammonium cerium nitrate with average in blue. Group 2: $150\mu\text{M}$ ammonium cerium nitrate with average in green. Group 3: $200\mu\text{M}$ 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.

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

92 The cerium dataset consists of a total 96 plant signals (time series) recorded at
 93 128 time points, with the control and groups 1–3 (each corresponding to a different
 94 concentration of ammonium cerium nitrate) all containing 24 plants. The control
 95 group is grown in Hoagland's media (Hoagland et al., 1950), which contains essential
 96 nutrients required for plant growth, and is not exposed to any additional levels of
 97 ammonium cerium nitrate. To examine the effects of cerium on the circadian clock, the
 98 other three groups, while also grown in Hoagland's media, were additionally exposed
 99 to varying additional concentrations of ammonium cerium nitrate— $100\mu\text{M}$ for Group
 100 1, $150\mu\text{M}$ for Group 2 and $200\mu\text{M}$ for Group 3. A plot of individual luminescence time
 101 series, the average expression at each time point, for each of the treatment groups,
 102 is shown in Figure 1. Note that time is measured in hours relative to *zeitgeber* time,
 103 which is the time of the last external temporal cue: the dawn signal of lights-on.

104 **2.2. BRASS analysis.** In the circadian community, analysis of this data would
 105 typically be performed by the Microsoft Excel macro BRASS. Table 1 provides a
 106 summary of the output of the analysis of the cerium dataset in BRASS. In particular,
 107 it shows the mean period estimate (obtained using FFT-NLLS analysis (Plautz et al.,
 108 1997) considering only period estimates between 15 and 40 hours), the number of
 109 plants that could not be analysed by BRASS and the mean Relative Amplitude Error
 110 (RAE) for each of the 4 groups. RAE is a value between 0 and 1 and gives information
 111 about the goodness of fit of the model (a value of 0 indicates a perfect fit). Circadian

Group	Hoagland's	Group 1 (100 μ M)	Group 2 (150 μ M)	Group 3 (200 μ M)
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.

112 biologists often visualise these results in a scatter plot of relative amplitude error
113 against period length for the plants analysed by BRASS (see e.g. Hanano et al.
114 (2006)) and such a plot for this dataset is given in Figure S2, Appendix A.

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

127 **2.3. Nonstationarity in circadian rhythms.** Price et al. (2008) asserted that
128 data arising from circadian experiments is nonstationary and discussed the features
129 which support this claim, namely a progressively dampened signal with a changing
130 period. The authors advocated the use of wavelets to analyse circadian data and devel-
131 oped a technique for characterising the modal periods present in circadian data using
132 a continuous wavelet decomposition (this is disseminated in the `waveclock` package
133 in R, currently on CRAN archive). Later, Harang et al. (2012) also supported the
134 circadian data nonstationarity view, and furthermore claimed that circadian analysis
135 under nonstationary behaviour by means of traditional Fourier methods can lead to
136 inaccurate results. Harang et al. (2012) thus recommended the use of wavelets, which
137 allow the changes in period to be tracked through time, and developed ‘WAVOS’- a
138 wavelet-based MATLAB toolkit that allows for analysis of nonstationary circadian
139 data.

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

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

Group	Hoagland's	Group 1 (100 μ M)	Group 2 (150 μ M)	Group 3 (200 μ M)
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.

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

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

163 **2.4. Individual-level variability in circadian rhythms.** We noticed in our
 164 dataset the presence of individual-level variability in plant responses to the same
 165 stimuli, despite their sharing identical genetic characteristics (Doyle et al., 2002). For
 166 example, different types of behaviour can be seen in the control group of Figure 1.
 167 This is particularly noticeable at the beginning (prior to time $T = 36$) and end (after
 168 time $T = 96$) of the experiment where the plant signals displayed one of two different
 169 amplitudes. This variability highlights the issues caused by taking an average period
 170 estimate for each group and comparing the results, or comparing the average raw time
 171 series for each group. Although all plants in each treatment group share identical
 172 genetic characteristics and have been treated in identical conditions, they respond
 173 differently. In such situations, looking at average behaviour masks the individual
 174 differences and is conducive to misleading conclusions, as also acknowledged in other
 175 fields (Fiecas and Ombao, 2016). This motivates our choice to cluster the circadian
 176 plant data using their time-frequency (scale) patterns and further accounts for their
 177 proven (see Section 2.3) nonstationary features.

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

185 **3.1. Modelling nonstationary time series.** Many of the statistically rigor-
 186 ous approaches to modelling nonstationary time series are based on the Cramér-Rao
 187 representation of stationary processes: all zero-mean discrete time second-order sta-
 188 tionary time series $\{X_t\}_{t \in \mathbb{Z}}$ can be represented as

$$189 \quad (1) \quad X_t = \int_{-\pi}^{\pi} A(\omega) \exp(i\omega t) d\xi(\omega),$$

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

192 In the representation in equation (1) above, we note that, for stationary pro-
 193 cesses, the amplitude $A(\omega)$ does not depend on time (i.e. the frequency behaviour
 194 is the same across time). However, for many real time series, including the cerium
 195 dataset, this assumption is not realistic and a model where the frequency behaviour
 196 can vary with time would therefore be preferable. One way of introducing time depen-
 197 dence into a model is by replacing the amplitudes $A(\omega)$ with a time-dependent form.
 198 Priestley (1965) introduced a time-frequency model with the amplitude replaced by
 199 $A(\omega, t)$, while Dahlhaus (1997) introduced the locally stationary modelling philoso-
 200 phy and developed the locally stationary Fourier (LSF) model. In this setting, the
 201 time-dependent amplitude function is defined on ‘rescaled time’ to enable asymptotic
 202 considerations.

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

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

$$217 \quad (2) \quad X_{t;T} = \sum_{j=1}^J \sum_{k \in \mathbb{Z}} w_{j,k;T} \psi_{j,k}(t) \xi_{j,k},$$

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

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

224 Estimation under the LSW framework is made possible by controlling the speed
 225 of evolution of the amplitudes $\{w_{j,k;T}\}$ using a condition of the form $\sup_k |w_{j,k;T} -$
 226 $W_j(k/T)| \leq C_j/T$, where $W_j(z)$, $z \in (0, 1)$ is a ‘limiting’ amplitude function with

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

229 The definition of the LSW process in Equation (2) requires the data to be of
 230 dyadic length ($T = 2^J$). In many practical applications, this is not realistic and there
 231 are a number of approaches to address this situation. For example, the practitioner
 232 could truncate the time series and analyse a segment of the data (of length $T = 2^J$),
 233 and this is our approach here. Alternatively, it is possible to extend the data to the
 234 next greater power of two by artificially appending values. In particular, common
 235 approaches include padding the data with zeros, replicating a data value (such as
 236 the final value) or reflecting the dataset about an end point. Another approach is
 237 to interpolate data values to produce a new data set of the required length (Ogden,
 238 1997). However, preconditioning the data could lead to misleading results. Therefore,
 239 we do not artificially extend the data in this paper.

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

$$245 \quad (3) \quad S_j(z) = |W_j(z)|^2,$$

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

247 An unbiased estimator of the EWS $\{S_j(z)\}$ is obtained by correcting the raw
 248 wavelet periodogram $I_{k,T}^j = |d_{j,k;T}|^2$, where $d_{j,k;T} = \sum_{t=0}^T X_{t,T} \psi_{j,k}(t)$ are the empiri-
 249 cal nondecimated wavelet coefficients. The correction is attained by premultiplying
 250 the raw wavelet periodogram vector $\mathbf{I}(z) := (I_{[zT],T}^j)_{j=1}^J$ by the inverse of the auto-
 251 correlation wavelet inner product ($J \times J$) matrix, $A_J = (\sum_{\tau} \Psi_j(\tau) \Psi_l(\tau))_{j,l}$, where
 252 $\Psi_j(\tau) = \sum_k \psi_{j,k}(0) \psi_{j,k}(\tau)$ is the autocorrelation wavelet.

253 Thus, the corrected wavelet periodogram is

$$254 \quad (4) \quad \mathbf{L}(z) = A_J^{-1} \mathbf{I}(z), \text{ for all } z \in (0, 1).$$

255 As in the stationary setting, the wavelet periodogram is not a consistent estimator
 256 of the wavelet spectrum (Nason, 2010). One method to overcome this is to smooth the
 257 raw wavelet periodogram as a function of (rescaled) time within each scale j , and then
 258 to apply the correction above. Various smoothing approaches have been proposed in
 259 the literature, see e.g. smoothing using variance stabilisation of Fryzlewicz and Nason
 260 (2006).

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

263 **3.2. Overview of current clustering/classification techniques that ac-**
 264 **count for nonstationarity.** The problem of clustering and classification for non-
 265 stationary data has received a good deal of attention in the statistical literature,
 266 thanks to its relevance in many applied fields. In the context of monitoring poten-
 267 tial nuclear testing, Shumway (2003) considered the use of time-varying spectra for
 268 the classification and clustering of nonstationary time series by means of locally sta-
 269 tionary Fourier models and Kullback-Leibler discrimination measures. Also in this
 270 context, Fryzlewicz and Ombao (2009) developed a procedure for the *classification* of
 271 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

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

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:

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

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:

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

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-Loève 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 ($\text{Var}(\alpha_m) = \lambda_m$) 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 $i = 1, \dots, N$. 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 $i = 1, \dots, N$, where $t = 0, \dots, T - 1$ 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-Loève 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 $\hat{S}^{(1)}, \dots, \hat{S}^{(N)}$. For each plant signal (each $i = 1, \dots, N$), 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,

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

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

$$370 \quad (9) \quad R = U \Lambda U^T,$$

371 where U is an orthonormal matrix whose columns are the eigenvectors of R (also
 372 known as the principal directions of the data; here, we can conceptualise these as
 373 representing ‘images’) and Λ is a diagonal matrix whose diagonal elements are eigen-
 374 values of R (positive real numbers arranged in decreasing order of magnitude; these
 375 are proportional to the variance accounted for by each direction). We can achieve size
 376 reduction by choosing to represent our data in fewer dimensions. The usual practice
 377 is to use the set of $p < n$ eigenvectors of R corresponding to the p largest eigenvalues
 378 and aggregate these in an $n \times p$ matrix, U_{PCA} , which performs the PCA projection.
 379 Therefore, for each eigenvector, we can find a corresponding projection in the princi-
 380 pal component space by computing QU_{PCA} . In this transformed space, each process
 381 is now represented by a p -dimensional vector, i.e. the principal co-ordinates of the
 382 i -th process are given by the i -th row of the matrix QU_{PCA} , denoted from now on as
 383 $\text{Score}^{(i)}$ (p -dimensional vector).

384 **3.4. Proposed clustering method.** Our proposal is to construct a clustering
 385 method that assesses time series similarity/ dissimilarity on the basis of their spectral
 386 content as distilled in the scores developed in Section 3.3 above. Next we shall intro-
 387 duce potential distance measure candidates and assess various methods to determine
 388 the number of principal components to retain and the optimal number of clusters.

389 **3.4.1. Distance measures.** The success of any clustering algorithm depends on
 390 the adopted dissimilarity measure. In this section, we propose four possible distance
 391 measures and discuss their advantages and disadvantages. The proposed distance
 392 measures consist of developments of those adopted in the work reviewed in Section
 393 3.2. In our simulation studies (Section 4), we compare the performance of clustering
 394 algorithms embedding the different distance measures outlined below.

395 The simplest choice for the dissimilarity measure is the squared quadratic (SQ)
 396 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
 397 adopted by Fryzlewicz and Ombao (2009) who quote its advantages of good practical
 398 performance and computational ease. In our context it is defined as the sum of the
 399 squared differences between the scores relating to the p principal components retained

$$400 \quad (10) \quad SQ(X_{t,T}^{(i)}, X_{t,T}^{(j)}) = \sum_{k=1}^p \left[\text{Score}_k^{(i)} - \text{Score}_k^{(j)} \right]^2,$$

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

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

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

$$410 \quad (11) \quad WSQ(X_{t,T}^{(i)}, X_{t,T}^{(j)}) = \frac{\sum_{k=1}^p \lambda_k [\text{Score}_k^{(i)} - \text{Score}_k^{(j)}]^2}{\sum_{k=1}^p \lambda_k},$$

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

414 We now outline the distance measures as adopted in [Antoniadis et al. \(2013\)](#)
 415 and [Rouyer et al. \(2008\)](#). Both approaches hinge on the singular vectors and leading
 416 patterns for each time series pair. Specifically, [Antoniadis et al. \(2013\)](#) compared the
 417 time evolution of each pair of leading patterns. In particular, for the k -th pair of
 418 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
 419 take the first difference (Δ) and measure energy by means of its modulus

$$420 \quad (12) \quad d_k(i, j) = |\Delta(P_k^{(i)} - P_k^{(j)})|.$$

421 Finally, the most significant p directions are aggregated using a weighted combi-
 422 nation with weights given by the squared singular values:

$$423 \quad (13) \quad D(i, j) = \frac{\sum_{k=1}^p \lambda_k d_k^2(i, j)}{\sum_{k=1}^p \lambda_k}.$$

424 The last comparison metric is

$$425 \quad (14) \quad DT(i, j) = \frac{\sum_{k=1}^p \lambda_k (RD(P_k^{(i)}, P_k^{(j)}) + RD(\mathbf{u}_k^{(i)}, \mathbf{u}_k^{(j)}))}{\sum_{j=1}^p \lambda_k},$$

426 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
 427 RD denotes the measure from [Rouyer et al. \(2008\)](#), adapted from [Keogh and Pazzani \(1998\)](#).
 428 This metric compares two vectors by measuring the angle between each pair
 429 of corresponding segments (a segment is defined as a pair of consecutive points of
 430 a vector) and is a method for measuring parallelism between curves. The overall
 431 distance is then computed as a weighted mean of the distance for each of the p pairs
 432 of leading patterns and singular vectors retained (with the weights being equal to the
 433 amount of covariance explained by each axis).

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

437 **3.4.2. Determining the number of principal components to retain.** Re-
 438 call the aim to reduce the dimensionality of our problem; for each of the distance
 439 metrics above, we must decide how many axes, p , to retain. [Antoniadis et al. \(2013\)](#)
 440 and [Rouyer et al. \(2008\)](#) both decided to use the number of axes that correspond to a
 441 fixed percentage of the total covariance (as is common in principal components analy-
 442 sis). A different approach is to select the number of components based on a screeplot.
 443 This displays the proportion of variance explained by the (ordered) eigenvalues, and
 444 p is then selected by looking for an elbow in the screeplot. [Cho et al. \(2013\)](#) proposed
 445 selecting this value based on the dimension of the correlation between two curves, r .

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

452 **3.4.3. Determining the number of clusters.** One of the most difficult tasks
 453 in clustering is determining the number of clusters ([Antoniadis et al., 2013](#)). This can
 454 be informed through a number of statistical techniques ([Kaufman and Rousseeuw,](#)
 455 [2009](#)) as well as by scientific expert knowledge. For example, the ‘elbow method’
 456 examines the percentage of variance explained as a function of the number of clus-
 457 ters; the number of clusters is then chosen by looking for an elbow in the plot of
 458 this function. [Tibshirani et al. \(2001\)](#) developed this methodology by estimating the
 459 number of clusters in a dataset via the gap statistic. Alternatively, the ‘silhouette
 460 method’ ([Rousseeuw, 1987](#)) can be used. The ‘silhouette’ of a data point is a number
 461 between -1 and 1 , with values of 1 indicating correct clustering, and optimization
 462 techniques are then used to determine the number of clusters that gives rise to the
 463 largest ‘silhouette’ ([Kaufman and Rousseeuw, 2009](#)).

464 **3.4.4. Proposed LSW-PCA clustering algorithm.** Our proposed clustering
 465 method, which we shall refer to as LSW-PCA clustering, is outlined in [Algorithm 1](#)
 466 below. We perform a partitioning around medoids (PAM) that admits a general
 467 dissimilarity matrix as input and is known to be more robust than other alternatives
 468 such as k-means ([Antoniadis et al., 2013](#)). Each of the proposed choices, i.e. spectral
 469 information, number of principal components retained (p) and distance measure, are
 470 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 $i = 1, 2, \dots, N$.

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.
-

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

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

480 **4.1. Simulated data.** The basic structure of each simulated experiment can
 481 be described as follows. A dataset of $N = 100$ (50 simulations from each of the
 482 two groups) was generated using the LSW representation (see equation (2)) with
 483 Daubechies' extremal phase wavelet with one vanishing moment and a Gaussian or-
 484 thonormal increment sequence with mean zero and unit variance (the `locits` R pack-
 485 age was used). Each periodogram was level smoothed by log transform, followed
 486 by translation invariant global universal thresholding and then the inverse transform
 487 was applied. For each scale of the wavelet periodogram, only levels 3 and finer were
 488 thresholded. Using the estimated spectral information, we obtained a dissimilarity
 489 matrix for each of the methods under investigation. This matrix was the input of a
 490 PAM algorithm (performed in the `cluster` R package) which clustered the data into
 491 two groups. We then compared the clusters with the known group memberships and
 492 recorded the correctly clustered percentage. The above procedure was then repeated
 493 100 times and the results for each method were averaged.

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

$$499 \quad (15) \quad S_j^{(1)}(z) = \begin{cases} 4 \cos^2(4\pi z), & \text{for } j = 2, z \in (1/64, 16/64) \\ 4 \cos^2(2\pi z), & \text{for } j = 3, z \in (17/64, 1) \\ 0, & \text{otherwise;} \end{cases}$$

500 and

$$501 \quad (16) \quad S_j^{(2)}(z) = \begin{cases} 4 \cos^2(2\pi z), & \text{for } j = 2, z \in (17/64, 1) \\ 4 \cos^2(4\pi z), & \text{for } j = 3, z \in (1/64, 1/2) \\ 0, & \text{otherwise;} \end{cases}$$

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

506 **Case 2: Gradual period change.** For our second study, we assume each time series
 507 is a realisation from one of 3 possible groups, each with different spectral characteris-
 508 tics. In particular, each group represents a time series that gradually changes period
 509 from 24 to: 25 (Group 1), 26 (Group 2) and 27 (Group 3) over (approximately) two
 510 days, before continuing with the relevant period for a further two days. The purpose
 511 of this simulation study is to replicate a typical circadian experiment with changes
 512 that could not be captured by standard analyses that assume stationarity and report
 513 an average period value. Therefore, we will take $T = 256$ which is equivalent to a
 514 free-running period of 4 days with equally spaced observations every 22.5 minutes.
 515 Figure 3 shows the wavelet spectra which represent the gradually changing periods
 516

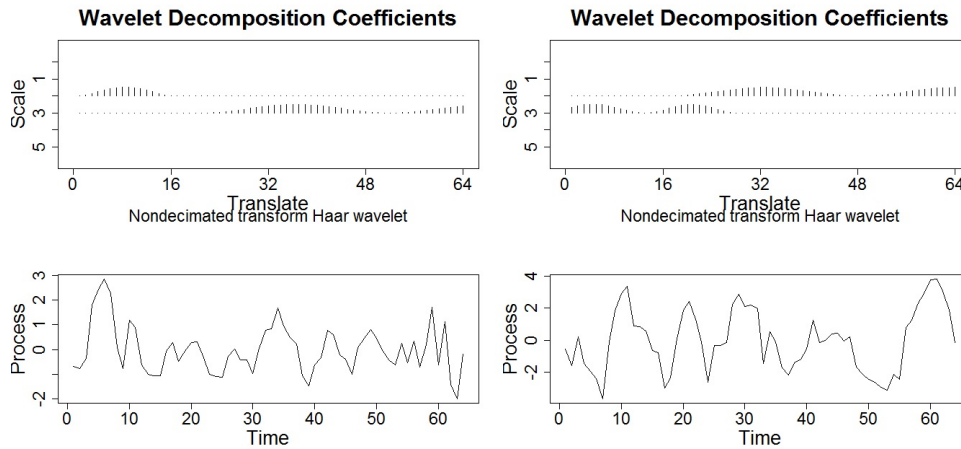


FIG. 2. Case 1. Top left: Group 1 wavelet spectrum; Top right: Group 2 wavelet spectrum; Bottom left: Group 1 realisation and Bottom right: Group 2 realisation.

517 that define each of the 3 groups above. Notice that the increased period is shown
 518 by the movement up through the resolution levels and the gradual increase in period
 519 of the wavelet coefficients. To determine which changes can be discriminated by the
 520 methods, we perform two studies within this setting (i) Case 2A: simulations from
 521 Group 1 and Group 2, and (ii) Case 2B: simulations from Group 1 and Group 3.
 522

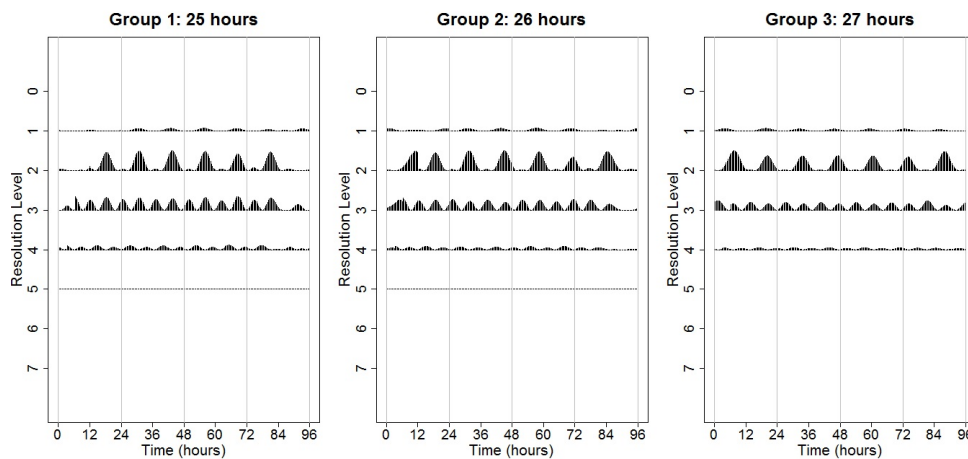


FIG. 3. Case 2. Left: Group 1 wavelet spectrum (gradual period change from 24 to 25 hours); Centre: Group 2 wavelet spectrum (gradual period change from 24 to 26 hours); Right: Group 3 wavelet spectrum (gradual period change from 24 to 27 hours).

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

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

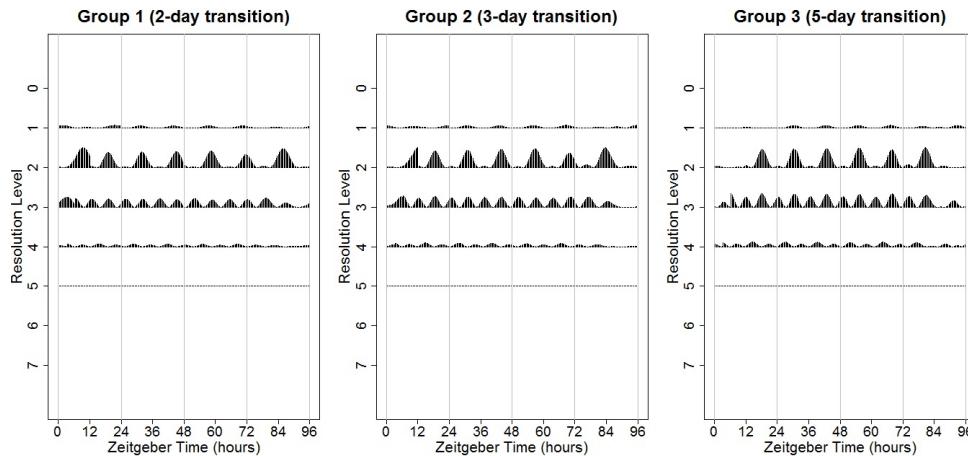


FIG. 4. Case 3. Left: Group 1 wavelet spectrum (2-day transition); Centre: Group 2 wavelet spectrum (3-day transition); Right: Group 3 wavelet spectrum (5-day transition).

537 **4.2. Results.** For each of our simulation studies outlined above, we investi-
 538 gate the impact of the wavelet information choice (e.g. wavelet coefficients versus
 539 model-based spectral estimate), distance measure choice and methods to determine
 540 the number of principal components to retain. We report our findings next, with
 541 detailed results for Case 1 presented in Appendix C.

542

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

549

550 **Dimension choice.** We also examined the different methods outlined in Section
 551 3.4.2 to select the number of principal components to retain for our LSW-PCA clus-
 552 tering method. We thus compared determining the number of principal components
 553 to retain by examining the screeplot with the situation where we retain the minimal
 554 number of components that correspond to 90% of the total covariance. Once again

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

561

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

577

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

599

5. Real data analysis.

600

601 **5.1. Previously published circadian data.** In this section, we apply our
602 method to an already published circadian dataset, which tested the effects of cop-
603 per on plants in a method similar to our cerium dataset. Our aim is to demonstrate
604 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.

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

607 The copper dataset was also obtained using a firefly luciferase reporter system as
 608 described in Appendix B. However, this experiment used a different gene of interest
 609 GIGANTEA (GI). For a detailed description of these experimental methods see Ap-
 610 pendix D and [Perea-García et al. \(2016a,b\)](#). Briefly, plants were grown under different
 611 copper regimes: ‘Deficiency’ (no CuSO₄), ‘Sufficiency’ or ‘Control’ (1 μM CuSO₄),
 612 and ‘Excess’ (10 μM CuSO₄). The copper dataset consists of a total of 74 plant sig-
 613 nals (time series) recorded at 151 time points, with the ‘Deficiency’ group containing
 614 19 plants; the ‘Control’ or ‘Sufficiency’ group, 26 plants and the ‘Excess’ group, 29
 615 plants. [Perea-García et al. \(2016a\)](#) conducted an analysis in BRASS (see Section 2.2)
 616 and concluded that the period did not seem to be affected by copper deficiency or
 617 excess. In particular, the average period estimates for each group were reported not
 618 statistically significantly different. Therefore, it was concluded that changes in avail-
 619 able copper were not readily detected by BRASS, even though qualitative differences
 620 were easily noted. These findings provide supportive evidence that more statistically
 621 advanced approaches are needed to analyse these types of data.

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

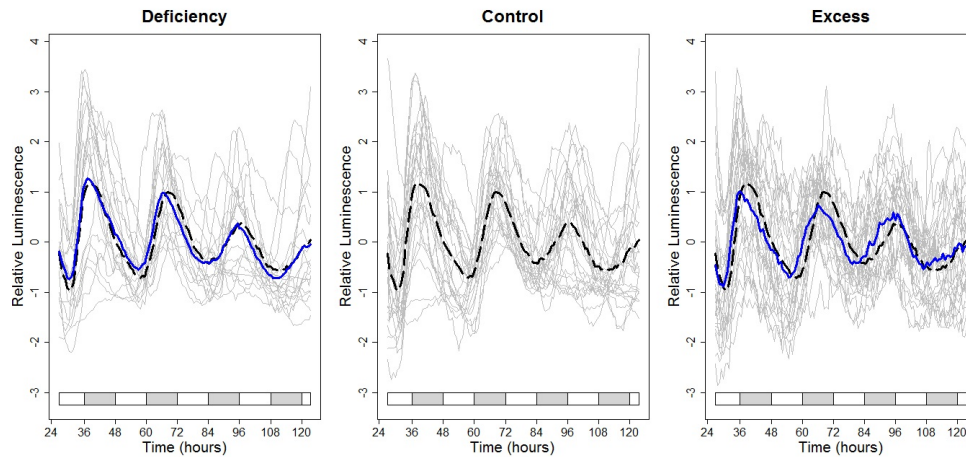


FIG. 5. Luminescence evolution over time for plants subjected to a control and 2 different copper regimes. Time is measured in hours relative to zeitgeber time (time of last external temporal cue: the dawn signal of lights-on). Centre: Each plant signal from the ‘Control’ group (in grey) along with the 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 ($10 \mu\text{M}$ 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.

639 In determining the optimal number of clusters, we used the ‘elbow method’ and
 640 then validated this result via the ‘silhouette method’ (implemented in the `fpc` R pack-
 641 age) and consultations with experimental scientists, as outlined in Section 3.4.3. All
 642 approaches indicated that we should cluster the data into two groups, which suggests
 643 the presence of two distinct groups within this dataset, each with different time-
 644 frequency behaviour. This is in contrast to the results in Perea-García et al. (2016a),
 645 which found no detectable difference in period. This illustrates the point in Section
 646 2.4, that although plants in each treatment group share identical genetic character-
 647 istics and have been treated in identical conditions, they can respond differently and
 648 average behaviour assessment can mask these differences.

649

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

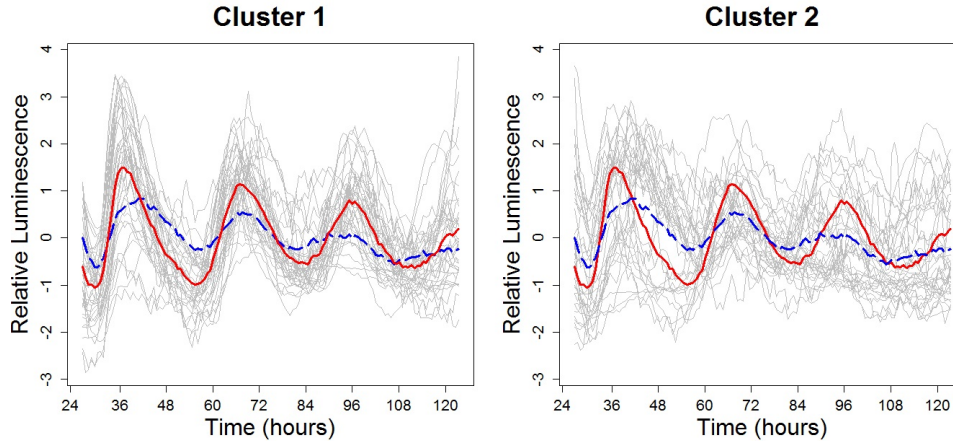


FIG. 6. Results of clustering the copper dataset into two clusters using the proposed LSW-PCA method. The individual signals (grey) along with the cluster average in: red for Cluster 1 and (dashed) blue for Cluster 2.

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

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

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

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

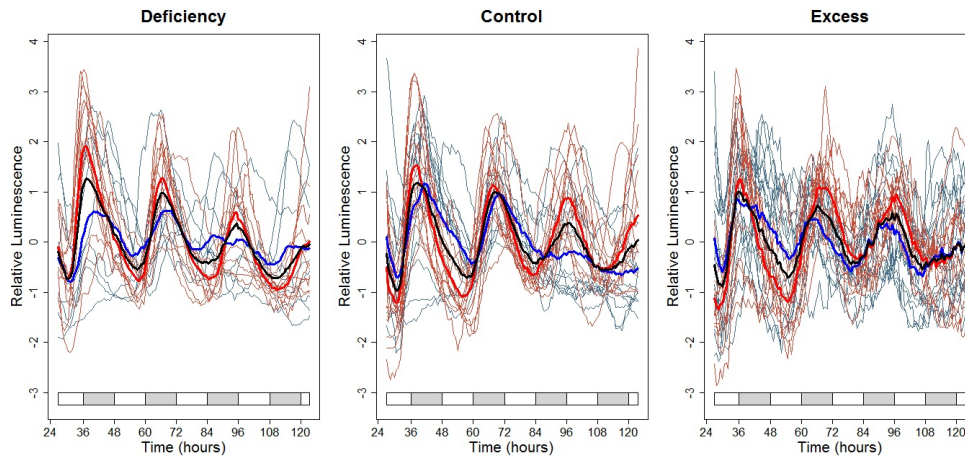


FIG. 7. Results of clustering the copper dataset into two clusters using the proposed LSW-PCA method. For each treatment group the individual signals are plotted in: red for Cluster 1 and blue for Cluster 2. The average of each treatment group is shown in black. Within each treatment group, the Cluster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

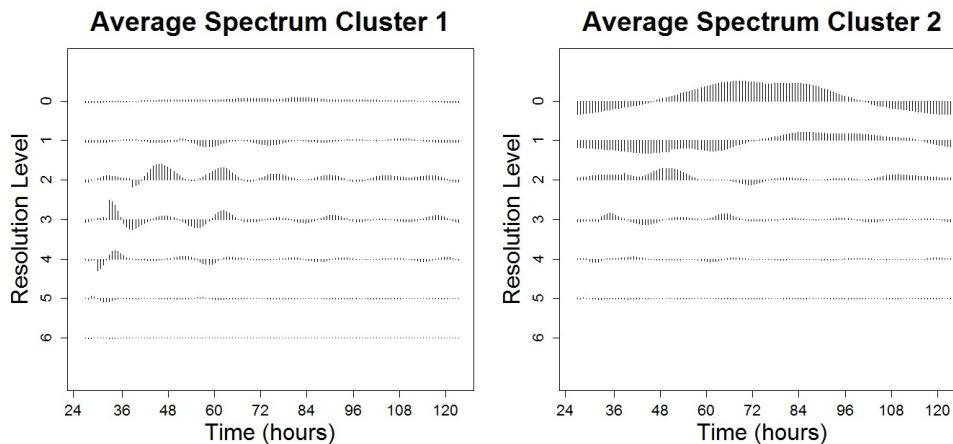


FIG. 8. Cluster average estimated spectra on the copper dataset using the proposed LSW-PCA method.

684 Furthermore, note the nonstationary behaviour that characterises both clusters
 685 (changing period and amplitude). The presence of these nonstationary characteris-
 686 tics supports our assertion that the existing methods (which assume stationarity) are
 687 inappropriate for such datasets and cannot capture this behaviour. Figure 7 shows
 688 that, although all plants in each treatment group share identical genetic characteris-
 689 tics and have been treated in identical conditions, they respond in two different ways.
 690 Note that the treatment group averages (in black) lie between the two (within treat-
 691 ment group) cluster averages. This is particularly noticeable in the ‘Deficiency’ group.
 692 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.

693 groups has resulted in similar average behaviour.

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

697 **5.2. Novel circadian plant data.** We now return to the circadian data that
 698 motivated this work and apply our proposed LSW-PCA clustering method to analyse
 699 the novel cerium data. As the LSW model is underpinned by wavelets and requires the
 700 data to be of dyadic length ($T = 2^J$), in our analysis we chose a segment of length $T =$
 701 128 out of the original dataset. This truncation was decided upon after consultation
 702 with the experimental scientists, as in Section 5.1. For each plant we estimated the
 703 wavelet spectrum by means of the corrected wavelet periodogram estimate (with the
 704 same setting as described in the simulation study in Section 4). On examining the
 705 screeplot (see Figure S4 in Appendix A) and for ease of interpretation, we retained
 706 two principal components to cluster the data. The proposed LSW-PCA clustering
 707 method yielded the results detailed in Table 5.

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

720

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

- 723 1. Cluster 1: contains mostly plants in the Control dataset (Hoagland’s), and
 724 very few plants subjected to lower-medium concentrations of ammonium
 725 cerium nitrate (100 μM and 150 μM)– conceptualised as essentially ‘Control’;
- 726 2. Cluster 2: contains mostly plants with lower concentration of ammonium
 727 cerium nitrate (100 μM) and a few plants from the Control dataset– concep-
 728 tualised as ‘Low concentration’;
- 729 3. Cluster 3: identifies similar behaviour to plants mostly exposed to medium-
 730 high concentrations (150 μM , 200 μM), but interestingly also contains a few
 731 plants from the Control and 100 μM concentration.

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

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

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

752 In contrast, Clusters 1 and 2 (approximately corresponding to the control and
753 low concentration groups respectively) display more similar, rhythmic behaviour. On
754 examining Figure 9, the rhythmic periods of the cluster averages seem approximately
755 equal. However, there are also clear differences between the two groups. Firstly, there
756 is a difference in the amplitudes of the two cluster averages. Cluster 1 has a larger
757 peak at approximately $t = 36$ and an even larger peak at $t = 120$. This can be seen in
758 the large coefficients around these time points in resolution levels 1-4 in the average
759 spectrum of Cluster 1. Alternatively, Cluster 2 seems to have a very large peak at
760 $t = 36$ followed by a distinct reduction in the amplitude of the other peaks. This can
761 also be seen in the large coefficients in resolution levels 2-4 in the average spectrum
762 of Cluster 2 in Figure 10.

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

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

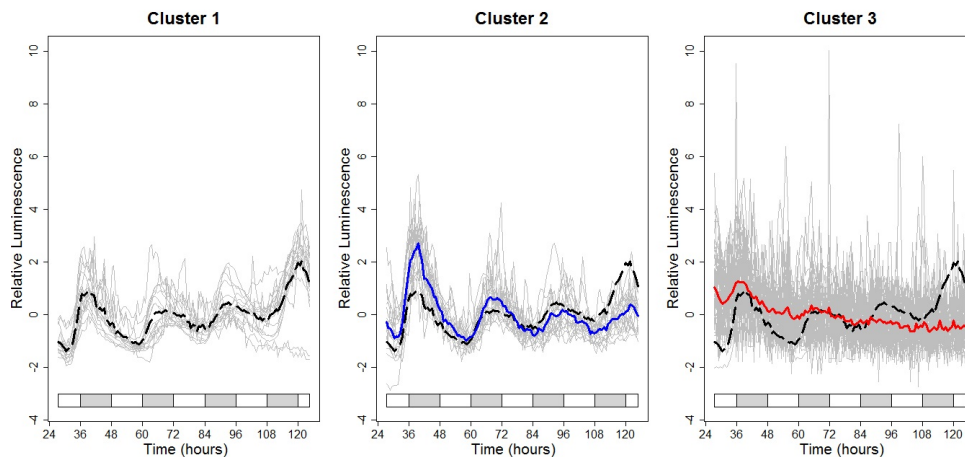


FIG. 9. The results of clustering the cerium dataset into three groups using the proposed LSW-PCA method. The individual signals (grey) along with the cluster average in: (dashed) black for Cluster 1; blue for Cluster 2 and red for Cluster 3. The average of Cluster 1 (conceptualised as essentially ‘Control’)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

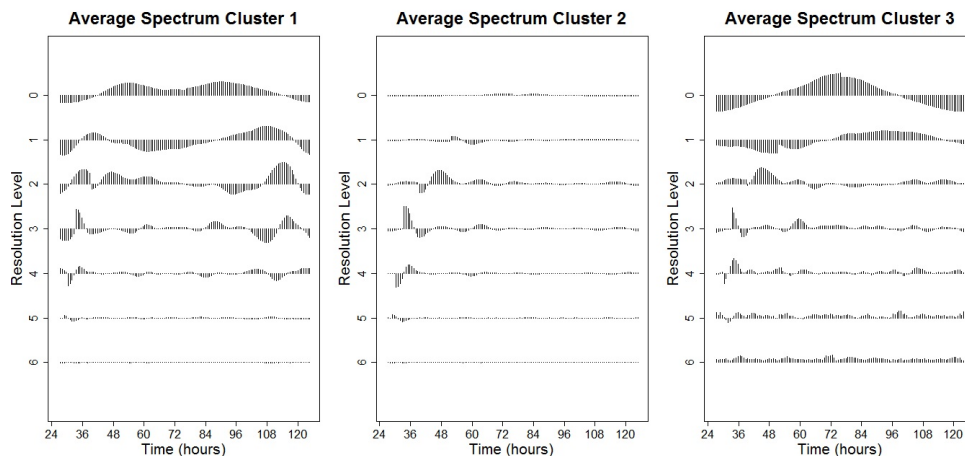


FIG. 10. Cluster average estimated spectra on the cerium dataset using the proposed LSW-PCA method. Cluster 1 approximately corresponds to the ‘Control’ group; Cluster 2 depicts ‘Low concentration’ behaviour ($100 \mu\text{M}$) and Cluster 3 the ‘Higher concentration’ ($150 \mu\text{M}$ and $200 \mu\text{M}$).

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

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

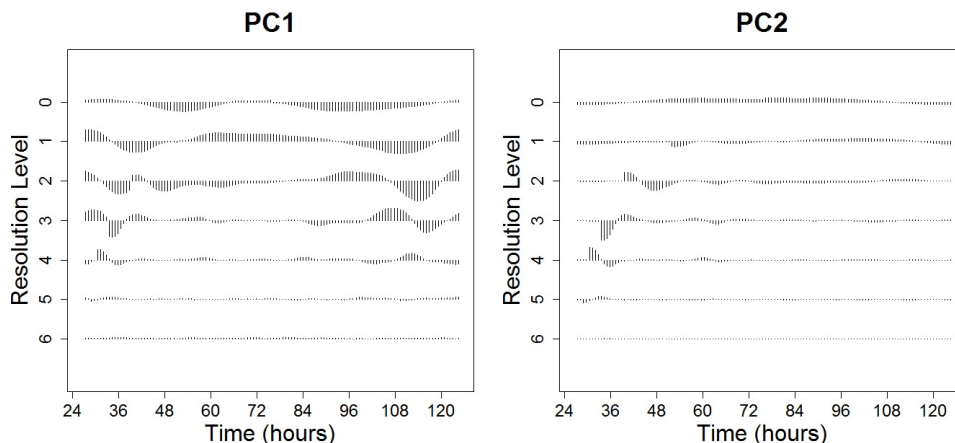


FIG. 11. *First two principal components obtained using the proposed LSW-PCA method on the cerium dataset.*

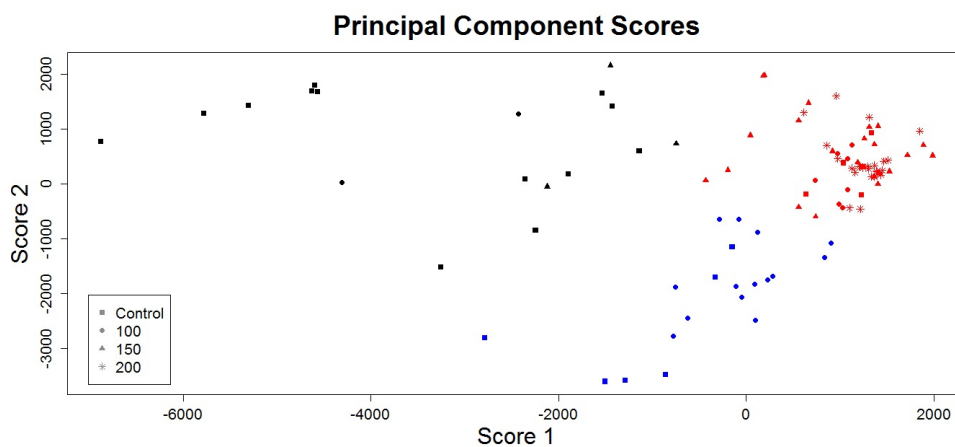


FIG. 12. *The cerium dataset projected onto the first two principal components obtained from the LSW-PCA clustering method. The colours represent the clusters: black for Cluster 1, blue for Cluster 2 and red for Cluster 3. The symbols represent the plant treatments.*

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

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

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

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

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

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

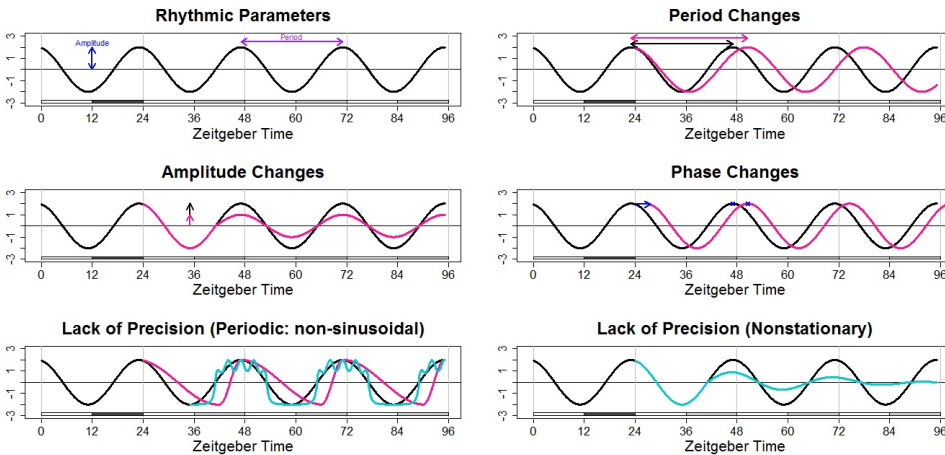


FIG. S1. The defined rhythmic parameters: periodicity, phase, amplitude and clock precision (based on an image from Hanano et al. (2006)).

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

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

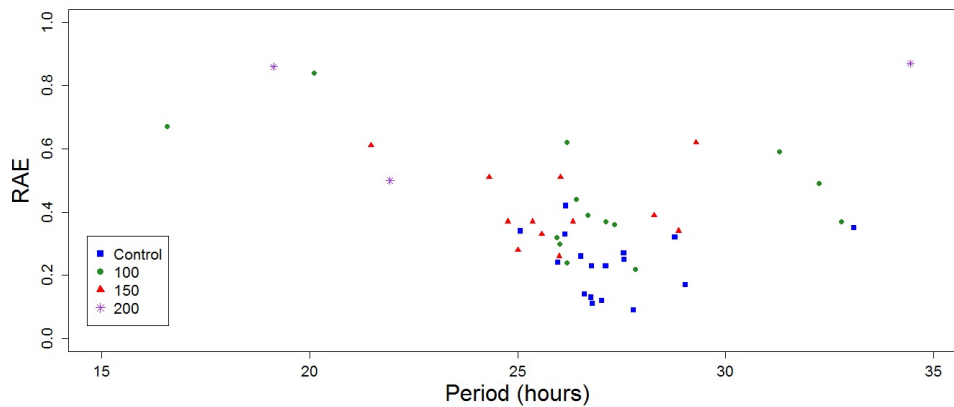


FIG. S2. Summary of the BRASS analysis of the circadian plant signals in response to differing quantities of ammonium cerium nitrate, represented by plots of period estimates plotted against the respective relative amplitude errors (RAE). The colours and symbols represent the plant treatment groups: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

828 the plant to produce quantifiable levels of light. This bioluminescence was measured
 829 using a TopCount NXT scintillation counter (Perkin Elmer), allowing relative gene
 830 expression of CCR2 to be quantified *in vivo* (Plautz et al., 1997; Southern and Mil-
 831 lar, 2005; Perea-García et al., 2016a). These experiments were carried out using the
 832 following methods: *Arabidopsis thaliana* seeds (Ws-CCR2:LUC) were surface steri-
 833 lised and plated onto Hoagland’s media containing 1% sucrose, 1.5% phyto agar
 834 (Hoagland et al., 1950). The seeds were stratified for 2 days at 4°C and transferred to
 835 growth chambers to entrain under 12:12 light/dark cycles at a constant temperature
 836 of 20°C . These conditions were chosen to simulate the ‘normal’ light/dark cycles of
 837 a day. Six-day-old seedlings were transferred to 96 well microtiter plates containing
 838 Hoagland’s 1% sucrose, 1.5% agar (Southern and Millar, 2005) also containing supple-
 839 mental $(\text{NH}_4)_2\text{Ce}(\text{NO}_3)_6$ (ammonium cerium nitrate) at a concentration of $100\mu\text{M}$,
 840 $150\mu\text{M}$ or $200\mu\text{M}$. The plants were then transferred to the TOPCount machine. Mea-
 841 surements were taken at intervals of approximately 45 minutes. Measurement began
 842 after the transition to 12 hours of darkness (known as subjective dusk) on the sev-
 843 enth day of the plants’ life. Therefore, the plant experiences one ‘normal’ day in the
 844 TOPCount machine (known as entrainment). After this, the plant was exposed to
 845 constant light (known as an LL free-run) for approximately four days. In Figure 1,
 846 the shaded bars below the graph represent the light conditions the plants would ex-
 847 perience during the ‘normal’ day. The plants are under constant light throughout the
 848 experiment, however, the grey bars indicate that they would be in darkness during a
 849 ‘normal’ 12 hour light/12 hour dark cycle.

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

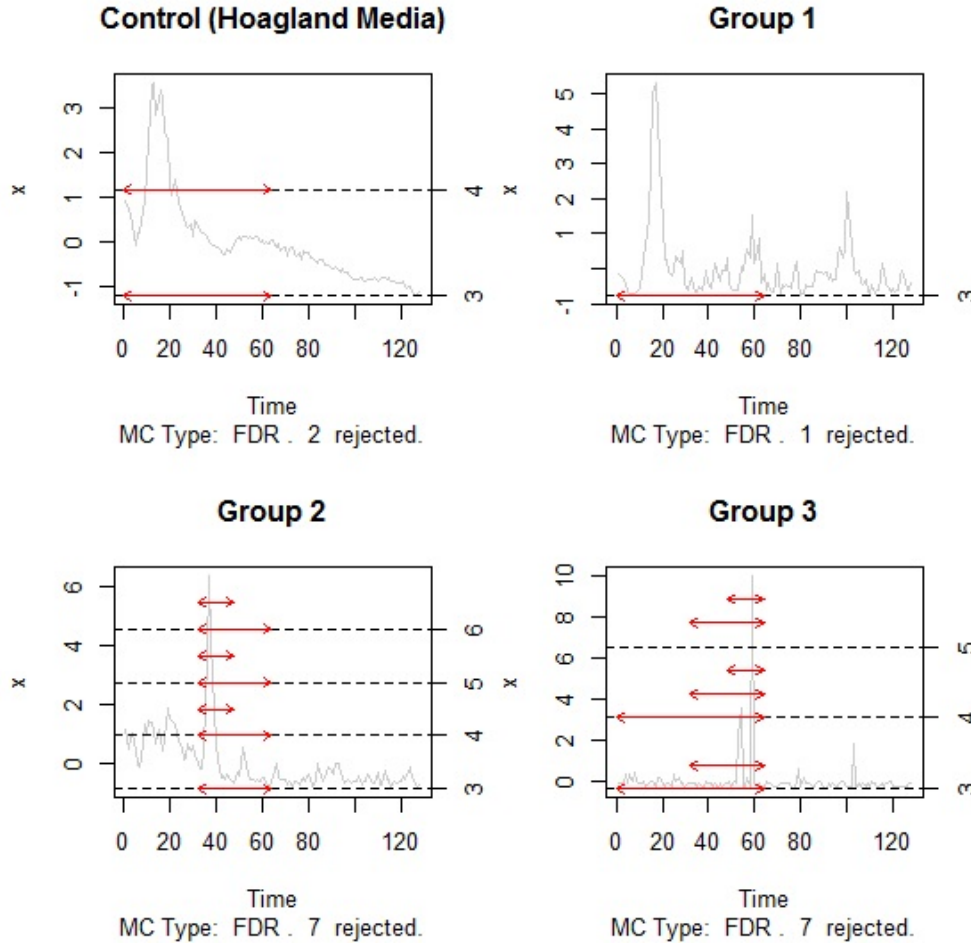


FIG. S3. Plots of the estimated locations of the nonstationarities in the circadian plant signals in response to differing quantities of ammonium cerium nitrate, using the wavelet spectrum test (Nason, 2013), implemented in the `locits` package in R which is available on CRAN. A time series for each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

855 additional levels of ammonium cerium nitrate. To examine the effects of cerium on
 856 the circadian clock, the other three groups, while also grown in the Hoagland's media,
 857 were additionally exposed to varying additional concentrations of ammonium cerium
 858 nitrate— $100\mu\text{M}$ for Group 1, $150\mu\text{M}$ for Group 2 and $200\mu\text{M}$ for Group 3.

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

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

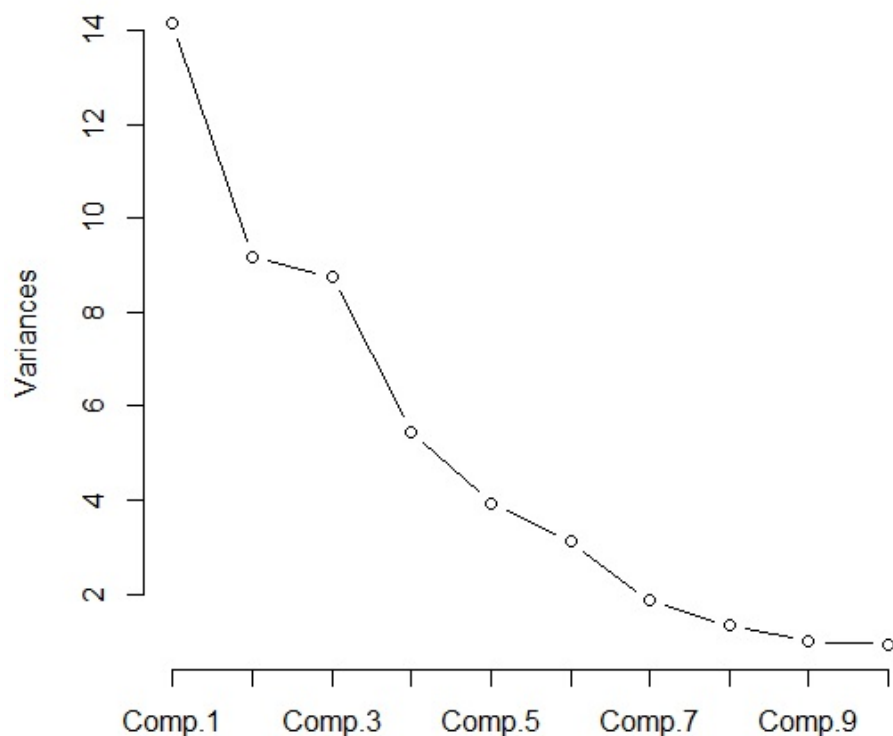


FIG. S4. The screeplot used to inform the selection of the number of principal components to retain for the cerium dataset.

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.

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

866 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
 867 Circadian Plant Data. However, this experiment uses a different gene of interest
 868 GIGANTEA (GI). Plants were grown on plates as described in Andrés-Colás et al.
 869 (2010), incubated on MS (Murashige and Skoog) medium (Murashige and Skoog,
 870 1962) at half concentration (1/2 MS) [phytoagar 0.8% (w/v) plus 1% sucrose (w/v) in
 871 0.5% MES (w/v)]. WS GI:LUC seedlings were grown under different copper regimes:
 872

Dimension reduction method	90% of total covariance	Screepplot
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.

873 ‘Deficiency’ (1/2 MS), ‘Sufficiency’ or ‘Control’ (1 μ M CuSO₄), and ‘Excess’ (10 μ M
874 CuSO₄). 96 plants were grown in total, 32 under each copper regime. The plants
875 were entrained for 7 days under 12:12 light-dark cycles at a constant temperature of
876 20°C. The plants were then exposed to constant light (LL free-run) for the remainder
877 of the experiment. Bioluminescence was then measured every hour using the same
878 TopCount NXT system as in Appendix B.

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

885 References.

- 886 Andrés-Colás, N., Perea-García, A., Puig, S., and Peñarrubia, L. (2010). Deregu-
887 lated copper transport affects Arabidopsis development especially in the absence of
888 environmental cycles. *Plant physiology*, 153(1):170–184.
- 889 Antoniadis, A., Brossat, X., Cugliari, J., and Poggi, J.-M. (2013). Clustering func-
890 tional data using wavelets. *International Journal of Wavelets, Multiresolution and*
891 *Information Processing*, 11(01):1350003.
- 892 Bell-Pedersen, D., Cassone, V. M., Earnest, D. J., Golden, S. S., Hardin, P. E.,
893 Thomas, T. L., and Zoran, M. J. (2005). Circadian rhythms from multiple oscilla-
894 tors: lessons from diverse organisms. *Nature Reviews Genetics*, 6(7):544–556.
- 895 Bujdoso, N. and Davis, S. J. (2013). Mathematical modeling of an oscillating gene
896 circuit to unravel the circadian clock network of Arabidopsis thaliana. *Frontiers in*
897 *Plant Science*, 4:3.
- 898 Cho, H., Goude, Y., Brossat, X., and Yao, Q. (2013). Modeling and forecasting daily
899 electricity load curves: a hybrid approach. *Journal of the American Statistical*
900 *Association*, 108(501):7–21.
- 901 Cressie, N. and Wikle, C. K. (2015). *Statistics for spatio-temporal data*. John Wiley
902 & Sons.
- 903 Dahlhaus, R. (1997). Fitting time series models to nonstationary processes. *The*
904 *Annals of Statistics*, 25(1):1–37.
- 905 Daubechies, I. (1992). *Ten Lectures on Wavelets*, volume 61. SIAM.
- 906 Doyle, M. R., Davis, S. J., Bastow, R. M., McWatters, H. G., Kozma-Bognár, L., and
907 Nagy, Ferenc and Millar, A. J. and Amasino, R. M. (2002). The ELF4 gene controls
908 circadian rhythms and flowering time in Arabidopsis thaliana. *Nature*, 419(6902):
909 74–77.
- 910 Edwards, K. D., Akman, O. E., Knox, K., Lumsden, P. J., Thomson, A. W., Brown,
911 P. E., Pokhilko, A., Kozma-Bognar, L., Nagy, F., Rand, D. A., and Millar, A
912 (2010). Quantitative analysis of regulatory flexibility under changing environmental

- 913 conditions. *Molecular systems biology*, 6(1): 424.
- 914 Fiecas, M. and Ombao, H. (2016). Modeling the evolution of dynamic brain processes
915 during an associative learning experiment. *Journal of the American Statistical*
916 *Association*, 111:1440–1453.
- 917 Fryzlewicz, P. and Nason, G. P. (2006). Haar–fisz estimation of evolutionary wavelet
918 spectra. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*,
919 68(4):611–634.
- 920 Fryzlewicz, P. and Ombao, H. (2009). Consistent classification of nonstationary time
921 series using stochastic wavelet representations. *Journal of the American Statistical*
922 *Association*, 104:299–312.
- 923 Hanano, S., Domagalska, M. A., Nagy, F., and Davis, S. J. (2006). Multiple phyto-
924 hormones influence distinct parameters of the plant circadian clock. *Genes to Cells*,
925 11(12):1381–1392.
- 926 Harang, R. Bonnet, G. and Petzold, L. R. (2012). WAVOS: a MATLAB toolkit
927 for wavelet analysis and visualization of oscillatory systems *BMC research notes*,
928 *BioMed Central*, 5(1):163.
- 929 Hoagland, D. R. and Arnon, D. I. (1950). The water-culture method for growing
930 plants without soil. *California Agricultural Experiment Station, Circular*, 347.
- 931 Holan, S. H., Wikle, C. K., Sullivan-Beckers, L. E., and Coccoft, R. B. (2010). Mod-
932 eling complex phenotypes: generalized linear models using spectrogram predictors
933 of animal communication signals. *Biometrics*, 66(3):914–924.
- 934 Kaufman, L. and Rousseeuw, P. J. (2009). Finding groups in data: an introduction
935 to cluster analysis. *John Wiley & Sons*, 98:239–243.
- 936 Keogh, E. J. and Pazzani, M. J. (1998). An enhanced representation of time series
937 which allows fast and accurate classification, clustering and relevance feedback.
938 *Proc. of the 4th International Conference of Knowledge Discovery and Data Mining*,
939 *AAAI Press*, 98:239–243.
- 940 Krzemiesiewska, K., Eckley, I. A., and Fearnhead, P. (2014). Classification of non-
941 stationary time series. *Stat*, 3(1):144–157.
- 942 Leise, T. L., Indic, P., Paul, M. J. and Schwartz, W. J. (2013). Wavelet meets
943 actogram. *Journal of biological rhythms*, *SAGE Publications Sage CA: Los Angeles*,
944 *CA*, 28(1):62–68.
- 945 McClung, C. R. (2006). Plant circadian rhythms. *The Plant Cell*, 18(4):792–803.
- 946 Minors, D. S. and Waterhouse, J. M. (2013). *Circadian rhythms and the human*.
947 *Butterworth-Heinemann*
- 948 Moore, A., Zielinski, T., and Millar, A. J. (2014). Online period estimation and deter-
949 mination of rhythmicity in circadian data, using the BioDare data infrastructure.
950 *Methods in Molecular Biology*, 1158:13–44.
- 951 Murashige, T. and Skoog, F. (1962). A revised medium for rapid growth and bio
952 assays with tobacco tissue cultures *Physiologia plantarum*, 15(3):473–497.
- 953 Nason, G. P., Von Sachs, R., and Kroisandt, G. (2000). Wavelet processes and adap-
954 tive estimation of the evolutionary wavelet spectrum. *Journal of the Royal Statis-*
955 *tical Society: Series B (Statistical Methodology)*, 62(2):271–292.
- 956 Nason, G. (2010). *Wavelet methods in statistics with R (use R)*. Springer Science &
957 *Business Media*.
- 958 Nason, G. (2013). *A test for second-order stationarity and approximate confidence*
959 *intervals for localized autocovariances for locally stationary time series*. *Journal*
960 *of the Royal Statistical Society: Series B (Statistical Methodology)*, *Wiley Online*
961 *Library*, 75(5):879–904.
- 962 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.