

## WAVELET SPECTRAL TESTING: APPLICATION TO NONSTATIONARY CIRCADIAN RHYTHMS<sup>1</sup>

BY JESSICA K. HARGREAVES, MARINA I. KNIGHT, JON W. PITCHFORD,  
RACHAEL J. OAKENFULL, SANGEETA CHAWLA, JACK MUNNS AND  
SETH J. DAVIS

*University of York*

Rhythmic data are ubiquitous in the life sciences. Biologists need reliable statistical tests to identify whether a particular experimental treatment has caused a significant change in a rhythmic signal. When these signals display nonstationary behaviour, as is common in many biological systems, the established methodologies may be misleading. Therefore, there is a real need for new methodology that enables the formal comparison of nonstationary processes. As circadian behaviour is best understood in the spectral domain, here we develop novel hypothesis testing procedures in the (wavelet) spectral domain, embedding replicate information when available. The data are modelled as realisations of locally stationary wavelet processes, allowing us to define and rigorously estimate their evolutionary wavelet spectra. Motivated by three complementary applications in circadian biology, our new methodology allows the identification of three specific types of spectral difference. We demonstrate the advantages of our methodology over alternative approaches, by means of a comprehensive simulation study and real data applications, using both published and newly generated circadian datasets. In contrast to the current standard methodologies, our method successfully identifies differences within the motivating circadian datasets, and facilitates wider ranging analyses of rhythmic biological data in general.

**1. Introduction.** Almost all species exhibit changes in their behaviour between day and night (Bell-Pedersen et al. (2005)). These daily rhythms (known as ‘circadian rhythms’) are the result of an internal timekeeping system, in response to daily changes in the physical environment (Vitaterna, Takahashi and Turek (2001), Minors and Waterhouse (2013)). The ‘circadian clock’ enhances survival by directing anticipatory changes in physiology synchronised with environmental fluctuations. When an organism is deprived of external time cues, its circadian rhythms typically persist qualitatively but may change in detail; the study of these changes can reveal the biochemical reactions underpinning the circadian clock and, at a larger scale, can provide valuable insight into the possible consequences of environmental and ecological challenges (McClung (2006), Bujdosó and Davis (2013)).

---

Received November 2018.

<sup>1</sup>Supported by EPSRC. Circadian work in the SJD group is currently funded by BBSRC awards BB/M000435/1 and BB/N018540/1.

*Key words and phrases.* Wavelets, spectral decomposition, hypothesis testing, circadian rhythms.

1.1. *Motivation.* In many scientific applications, the available data consist of signals with known group memberships and scientists are interested in establishing whether these groups display statistically different behaviour. Our work is motivated by a general problem: biologists need reliable statistical tests to identify whether a particular experimental treatment has caused a significant change in the circadian rhythm. If the changes are limited to period and/or phase then existing Fourier-based theory may be adequate. However, when the changes to the circadian clock are less straightforward, for example, involving nonstationarity or changes at multiple scales (Hargreaves et al. (2018)), the application of these established methods may be conducive to misleading conclusions. The value of our approach is illustrated by three complementary examples, encompassing the effect of various salt stresses on plants, the identification of mutations inducing rapid rhythms, and the response of nematode clocks to pharmacological treatment, as described in the following sections. The biological experimental details for each dataset appear in Appendix A in the Supplementary Material (Hargreaves et al. (2019)).

1.1.1. *Lead nitrate dataset (Davis Lab, Biology, University of York).* This dataset (henceforth referred to as the ‘Lead dataset’) is from a broad investigation of whether plant circadian clocks are affected by industrial and agricultural pollutants (Foley et al. (2005), Hargreaves et al. (2018), Senesil et al. (1998), Nicholson et al. (2003)). Specifically, this experiment asks whether lead affects the *Arabidopsis thaliana* circadian clock and, if so, when and how? Figure 1 displays the luminescence profiles for both untreated *A. thaliana* plants, as well as for those exposed to lead nitrate.

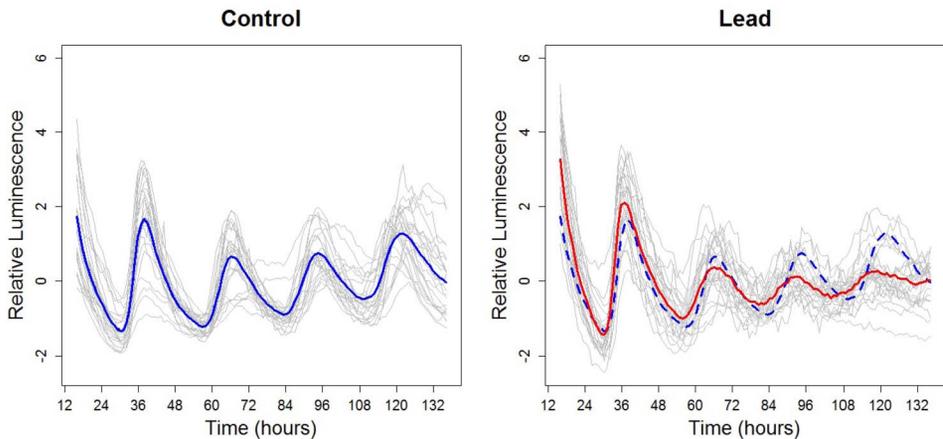


FIG. 1. Lead dataset: Luminescence profiles over time for untreated *A. thaliana* plants (Control) and those exposed to lead nitrate (Lead). Left: Individuals in the control group (in grey) along with the group average (bold). Right: Individuals in the lead treatment group (in grey) along with the treatment group average (bold) and the control group average (dashed). Each time series has been re-centred around zero.

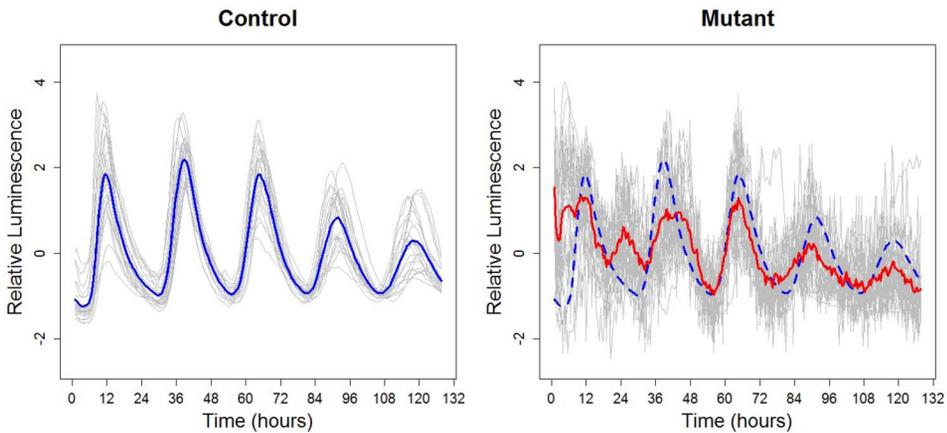


FIG. 2. Ultradian dataset: *Luminescence profiles over time for control and mutant A. thaliana plants.* Left: *Individuals in the control group (in grey) along with the group average (bold).* Right: *Individuals in the mutant group (in grey) along with the mutant group average (bold) and the control group average (dashed).* Each time series has been recentred around zero.

1.1.2. *Ultradian dataset (Millar Lab, Biology, University of Edinburgh).* In order to understand the clock mechanism, a common approach is to mutate a gene and examine the resulting behaviour in response to a variety of stimuli. Figure 2 depicts the luminescence profiles recording plant response to light for both the control and genetically mutated *A. thaliana* plants (Millar et al. (2015)). Researchers are interested in establishing whether a specific genetic mutation induced high-frequency behaviour (known as ‘ultradian rhythms’) in the laboratory model plant *A. thaliana*.

1.1.3. *Nematode dataset (Chawla Lab, Biology, University of York).* The free-living nematode *Caenorhabditis elegans* is an animal widely used in neuroscience and genetics, but its circadian clock is still poorly understood. To increase understanding of the nematode clock, and potentially uncover rhythmicity not detected by conventional approaches, researchers applied a pharmacological treatment to *C. elegans*, based on evidence that it causes aberrant circadian rhythms in other established mammalian and insect circadian models (Kon et al. (2015), Dusik et al. (2014)). Figure 3 depicts the luminescence profiles for both untreated and treated *C. elegans*.

On examining Figures 1 and 2, it is visually clear that changes in period and amplitude between the control and test groups occur in both datasets. Figure 3 reveals apparently similar luminescence profiles for both untreated and treated *C. elegans*. Nevertheless, in each experiment, less easily quantified or subtle differences between these groups may also exist.

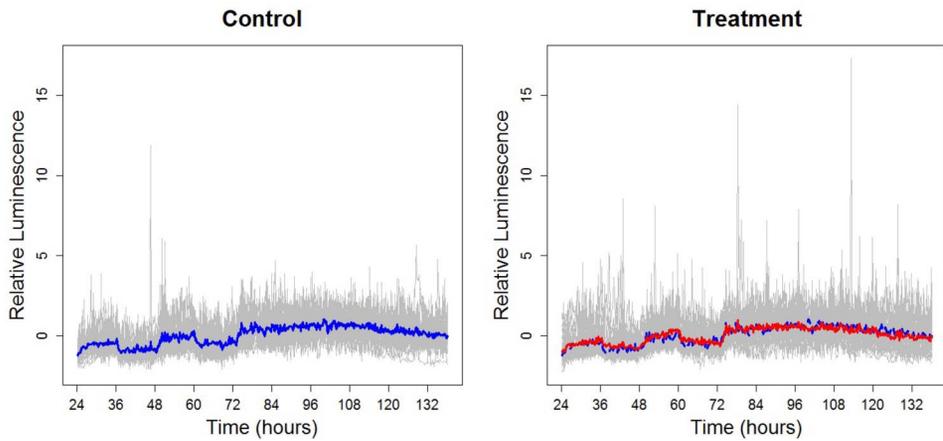


FIG. 3. Nematode dataset: *Luminescence profiles over time for untreated *C. elegans* (Control) and those subjected to a pharmacological treatment (Treatment). Left: Individuals in the control group (in grey) along with the group average (bold). Right: Individuals in the treatment group (in grey) along with the treatment group average (bold) and the control group average (dashed). Each time series has been recentred around zero.*

1.2. *Aims and structure of the paper.* Period estimation is central to the analysis of circadian data, with the current standard achieving this using Fourier analysis (Zielinski et al. (2014), Costa et al. (2011)) via software packages, such as BRASS (Biological Rhythm Analysis Software System (Edwards et al. (2010))) or BioDare (Moore, Zielinski and Millar (2014)). The practitioner estimates the period of the control and treatment groups respectively, and then tests for statistically significant differences (see, e.g., Perea-García et al. (2015), Costa et al. (2011)). Crucially, in all of our motivating examples, such established Fourier-based tests found no significant difference between groups (see Table S1 in Appendix B.1 in the Supplementary Material (Hargreaves et al. (2019))).

One obvious limitation of this analysis is that the employed methodology does not typically evaluate the crucial underpinning assumption of data stationarity. In the context examined here, assuming stationarity can be inappropriate (Hargreaves et al. (2018), Leise et al. (2013)), a feature shared by many biological systems (Zielinski et al. (2014)). For our motivating example datasets, we investigated whether the individual time series are (second-order) stationary via hypothesis testing. We employed two tests for stationarity—a Fourier-based test (the Priestley–Subba Rao (PSR) (Priestley and Subba Rao (1969)) test) and a wavelet-based test (Nason (2013)). The results (Table S2 in Appendix B.1 in the Supplementary Material (Hargreaves et al. (2019))) show that, for each of our motivating example datasets, over 80% of the time series provided enough evidence to reject the null hypothesis of stationarity. This result suggests that the application of the current methodology (which assumes data stationarity) would be inappropriate for our

motivating datasets and highlights the urgent need for more statistically advanced approaches.

In the specific context of circadian clock data, wavelets have been recognised as ideally suited to identifying local time and scale features (Leise et al. (2013), Harang, Bonnet and Petzold (2006)), with time-scale patterns known as indicative of the organism response to external stimuli (Zielinski et al. (2014)). A substantial body of circadian literature advocates the use of wavelet (Leise et al. (2013), Zielinski et al. (2014), Harang, Bonnet and Petzold (2006)) and, in particular, spectral representations (Price et al. (2008)) of circadian rhythms. This motivates our choice to formally compare circadian signals in the wavelet spectral domain by using their time-scale signature patterns and thus accounting for their proven nonstationary features. Furthermore, we propose to adopt the locally stationary wavelet (LSW) process model of Nason, von Sachs and Kroisandt (2000), which is capable of accounting for data nonstationarity and crucially has previously demonstrated utility for circadian analysis (Hargreaves et al. (2018)). Modelling nonstationary data within the LSW framework has also proven successful across a wide variety of fields, from climatology (Fryzlewicz, Van Bellegem and von Sachs (2003)) and ocean engineering (Killick, Eckley and Jonathan (2013)) to medicine (Nason and Stevens (2015)) and finance (Fryzlewicz (2005)) corresponding to a multitude of tasks such as forecasting, change-point detection, spectral estimation and modelling, respectively.

The primary contribution of this work is the development of novel wavelet-based hypothesis tests that allow for circadian behaviour comparison while accounting for data nonstationarity. This article is organised as follows. Section 2 reviews the theoretical wavelet-based framework we adopt for modelling nonstationary data and the relevant literature on hypothesis testing in the spectral domain. Our new hypothesis testing procedures are introduced in Section 3. Section 4 provides a comprehensive performance assessment of our new methods via simulation. Section 5 demonstrates the additional insight our techniques provide for the motivating circadian datasets and Section 6 concludes this work.

**2. Overview: Nonstationary processes and hypothesis testing in the spectral domain.**

2.1. *Modelling nonstationary processes.* Many of the statistically rigorous approaches to modelling nonstationary time series stem from the Cramér–Rao representation of stationary processes that states that all zero-mean discrete time second-order stationary time series  $\{X_t\}_{t \in \mathbb{Z}}$  can be represented as

$$(2.1) \quad X_t = \int_{-\pi}^{\pi} A(\omega) \exp(i\omega t) d\xi(\omega),$$

where  $A(\omega)$  is the amplitude of the process and  $d\xi(\omega)$  is an orthonormal increments process (Priestley (1981)). In the representation of stationary processes

above, the amplitude  $A(\omega)$  does not depend on time, that is, the frequency behaviour is the same across time. However, for many real time series, including our circadian datasets, this is unrealistic (Price et al. (2008)) and a model where the frequency behaviour can vary with time is needed.

The LSW paradigm provides precisely such a desired setup, and has also proved to yield superior results when compared to competitor methods in useful tasks such as classification (e.g., Krzemińska, Eckley and Fearnhead (2014) for aerosol spray data) and clustering (e.g., Hargreaves et al. (2018) for circadian rhythms). Fryzlewicz (2005) brings strong arguments for the utility of (linear) Gaussian LSW models for financial data, typically modelled using (nonlinear) models, that allow for time-dependent conditional variance.

In a nutshell, in the LSW framework, the Fourier building blocks in equation (2.1) are replaced by families of discrete nondecimated wavelets and an LSW process  $\{X_{t,T}\}_{t=0}^{T-1}$ ,  $T = 2^J \geq 1$  is represented as follows

$$(2.2) \quad X_{t,T} = \sum_{j=1}^J \sum_{k \in \mathbb{Z}} w_{j,k;T} \psi_{j,k}(t) \xi_{j,k},$$

where  $\{\xi_{j,k}\}$  is a random orthonormal increment sequence,  $\{\psi_{j,k}(t) = \psi_{j,k-t}\}_{j,k}$  is a set of discrete nondecimated wavelets and  $\{w_{j,k;T}\}$  is a set of amplitudes, each of which at a scale  $j$  and time  $k$ . Within each scale  $j$ , the amplitudes  $\{w_{j,k;T}\}_k$  are regulated by a Lipschitz continuous function  $W_j(k/T)$ , which further fulfils some technical assumptions in order to allow estimation. Appendix C in the Supplementary Material (Hargreaves et al. (2019)) provides the background details.

**2.1.1. Practical considerations.** In this paper, we assume the innovations  $\{\xi_{j,k}\}$  to be normally distributed, resulting in modelling the data  $\{X_{t,T}\}$  as a Gaussian LSW process. The normality assumption is typically employed for the (Fourier) circadian testing methodology (Perea-García et al. (2015)). This assumption is also commonly made in time series analysis in general and in LSW modelling in particular (e.g., Oh et al. (2003), Van Belleghem and von Sachs (2008) and Nason and Stevens (2015)), with Nason (2013) arguing for its nonlimiting character in this context. In Appendix B.2 (in the Supplementary Material (Hargreaves et al. (2019))) we show this assumption is tenable for our circadian datasets.

The properties of the random increment sequence  $\{\xi_{j,k}\}$  ensure that  $\{X_{t,T}\}$  is a zero-mean process. In practice, for a process with nonzero mean, it is customary to recentre it around zero (Nason (2008)) and this is our approach here, as the quantity of our primary interest is the process spectral signature.

As is typical for wavelet representations, the data is often required to be of dyadic length,  $T = 2^J$ . In many practical applications, this is not realistic and there are a number of approaches to address this situation (see, e.g., Ogden (1997)). Our approach is to analyse a (dyadic length) segment of the data, with the truncation decided upon careful consultation with the experimental scientists in order to ensure the time-frame of interest is represented.

2.1.2. *The evolutionary wavelet spectrum.* Under the LSW framework, a quantity analogous to the spectrum of a stationary process is the evolutionary wavelet spectrum (EWS)  $S_j(z) := |W_j(z)|^2$ , at each scale  $j \in \overline{1, J}$  and rescaled time  $z = k/T \in (0, 1)$ . The EWS quantifies the power distribution in the process over time and scale. We define the raw wavelet periodogram as  $I_{j,k;T} = |d_{j,k;T}|^2$ , where  $d_{j,k;T} = \sum_{t=0}^T X_{t,T} \psi_{j,k}(t)$  are the empirical nondecimated wavelet coefficients. In the remainder of this paper we drop the explicit dependence on  $T$  for the wavelet coefficients and the periodogram.

The raw wavelet periodogram is an asymptotically unbiased estimator of the quantity

$$(2.3) \quad \beta_j(z) = \sum_{i=1}^J A_{i,j} S_i(z) = (AS)_j(z),$$

where  $A = (A_{i,j})_{i,j=1}^J = (\sum_{\tau} \Psi_i(\tau) \Psi_j(\tau))_{i,j=1}^J$  is the autocorrelation wavelet inner product matrix, with  $\Psi_j(\tau) = \sum_k \psi_{j,k}(0) \psi_{j,k}(\tau)$  the autocorrelation wavelet (Nason, von Sachs and Kroisandt (2000)). The quantity  $\beta_j(z)$  was introduced by Fryzlewicz and Nason (2006) and is often easier to work with theoretically than the spectrum (Nason (2013)). An asymptotically unbiased estimator of the EWS is the empirical wavelet spectrum:

$$(2.4) \quad \mathbf{L}(z) := A^{-1} \mathbf{I}(z),$$

for all  $z \in (0, 1)$ , where  $\mathbf{I}(z) := (I_{j,[zT]})_{j=1}^J$  is the raw wavelet periodogram vector.

The empirical wavelet spectrum is a collection of random variables that are not independent, nor is their (joint or marginal) distribution easy to determine. As each coefficient of the empirical wavelet spectrum is a sum of a (typically logarithmic) number of terms (see equation (2.4)), a mechanism similar to the central limit theorem brings it closer to normality than the raw wavelet periodogram (Fryzlewicz and Ombao (2009)), which is distributed as a scaled  $\chi_1^2$ . As the individual raw periodogram ordinates within each scale are correlated, Fryzlewicz and Nason (2006) model the raw wavelet periodogram as approximately

$$I_{j,k} \sim \beta_j(z) Z_{j,k}^2,$$

where  $z = k/T$  and  $Z_{j,k}^2 \sim \chi_1^2$ , for  $j \in \mathbb{N}, k = 0, \dots, 2^j - 1 = T - 1$ .

A way to ‘correct’ these undesirable features is to employ a transform that brings the raw periodogram ordinates closer to Gaussianity and decorrelates within each scale. We adopt the Haar–Fisz transform (denoted  $\mathcal{F}$ ), introduced (for spectral estimation) by Fryzlewicz and Nason (2006), and apply it separately to each scale  $j = 1, \dots, J$  of the raw wavelet periodogram, denoted  $\mathcal{H}_{j,k;T} := \mathcal{F} I_{j,k;T}$ . Proposition 4 in Fryzlewicz and Nason (2006) then suggests a potential model

$$\mathcal{H}_{j,k} \sim N(\mathcal{B}_j(z), \sigma_j^2),$$

where  $\mathcal{B}_j(z) = \mathcal{F}\beta_j(z)$  with  $z = k/T$  and  $\mathcal{F}Z_{j,k}^2$  are approximately uncorrelated  $N(0, \sigma_j^2)$ , again dropping the explicit dependence on  $T$ . This model, viewed as a nonparametric additive regression model, was also employed by [Nason and Stevens \(2015\)](#) in the context of Bayesian spectral estimation, where its viability was demonstrated.

*2.2. Spectral domain hypothesis testing.* Assuming that the available data consists of multiple nonstationary time series with known group memberships, to the authors' knowledge no hypothesis tests exist to determine whether two groups are significantly different in terms of their associated (evolutionary) wavelet spectra. Wavelet spectral comparison is closest framed as a (consistent) classification method by [Fryzlewicz and Ombao \(2009\)](#), further improved by [Krzemieniewska, Eckley and Fearnhead \(2014\)](#). Spectral comparison, framed as testing for spectral constancy, also appears in connection with testing for time series stationarity and white noise testing. In the Fourier domain, [Priestley and Subba Rao \(1969\)](#) determined (as a hypothesis test) whether the spectrum is time-varying and, hence, whether the process is nonstationary. [von Sachs and Neumann \(2000\)](#) introduced the principle of assessing the constancy of the time-varying Fourier spectrum by examining its Haar wavelet coefficients across time. In the wavelet domain, [Nason \(2013\)](#) developed a test for second-order stationarity which examines the constancy of a wavelet spectrum by also examining its Haar wavelet coefficients. A similar approach is adopted by [Nason and Savchev \(2014\)](#) in the development of white noise tests.

The problem of testing that involves curves is often posed in time series literature as a functional regression problem defined using a functional response and categorical predictors (functional ANOVA; see the monograph of [Ramsay and Silverman \(2005\)](#) for its introduction and the review of [Morris \(2015\)](#) for developments in the field). Functional regression problems are often treated by projection in the Fourier or wavelet domain, where the spectral time series representations become subject to modelling. [Shumway \(1988\)](#) compares groups of curves (with stationary stochastic errors) by testing whether the mean curves have the same Fourier spectrum at each given frequency. [Fan and Lin \(1998\)](#) developed this method by applying the adaptive Neyman test to the (Fourier or wavelet) transformed difference vector (the difference between the two group-average time series). [Vidakovic \(2001\)](#) introduces a wavelet-based functional data analysis, with [McKay et al. \(2012\)](#) developing this as an approach for comparing neurophysiological signals that are functions of time. This approach was also subsequently adopted by [Atkinson et al. \(2017\)](#) to develop model validation using a test statistic based on thresholded wavelet coefficients. [Tavakoli and Panaretos \(2016\)](#) compare pairs of stationary functional time series by developing  $t$ -tests for the equality of their (Fourier) spectral density operators. However, these approaches fail to account for potential nonstationarity in the data. This is mitigated by [Guo et al.](#)

(2003), who propose a smoothing-spline ANOVA on the logarithm of the Fourier spectrum of a locally stationary process that is specifically designed to discriminate between models that contain a linear trend, modulation, time and frequency interaction terms, thus yielding global model comparisons, rather than time- and frequency-specific ones. The closest methodology for spectral comparison while allowing for a localised representation comes from [Martinez et al. \(2013\)](#) who identify regional differences in (the Fourier spectrograms of) bat mating chirps. The statistical modelling of windowed Fourier spectrograms as an image was first proposed by [Holan et al. \(2010\)](#) in a study that aimed to classify animal communication signals. [Martinez et al. \(2013\)](#) apply the higher-dimension functional mixed model of [Morris et al. \(2011\)](#) and use a Bayesian approach to fit a model that incorporates localised chirp Fourier spectrograms as the functional response and categorical regressors that identify bat location (fixed-effects) and independent bat (random)-effects. The observed data is modelled in a (projected) wavelet-domain with several distributional assumptions in place, for example, data Gaussianity, spike Gaussian-slab prior distributions for the wavelet coefficients. However, while their windowed Fourier spectrogram does offer a time-frequency representation of the data, thus potentially capturing nonstationarity, it is sensitive to the choice of kernel and crucially of window-width ([Martinez et al. \(2013\)](#)). In the context of clustering circadian plant rhythms, [Hargreaves et al. \(2018\)](#) demonstrated the superiority of a principled model-based spectral estimator that, in the spirit of [Holan et al. \(2010\)](#), was also used as an image in subsequent modelling. Additionally, we note that our study aims to identify not only (i) time-scale (frequency) group differences (conceptually a task close to [Martinez et al. \(2013\)](#)), but also (ii) to detect global scale-level differences (while still allowing for a development that incorporates potential nonstationarity) and (iii) to identify similar patterns within each scale, rather than exact differences (the reader will find precise details in the next section).

**3. Proposed spectral domain hypothesis tests.** Aligned to our motivating examples, the key goals of our work are to develop novel hypothesis tests, each capable of detecting one of three specific types of spectral differences between two groups and to identify the scales and times (e.g., Lead and Nematode datasets—Sections 1.1.1 and 1.1.3) or scales only (e.g., Ultradian dataset—Section 1.1.2) at which these difference arise, as appropriate.

Formally, recall that we model the observed nonstationary circadian rhythms as (Gaussian) LSW processes, using the framework of [Nason, von Sachs and Kroisandt \(2000\)](#) (see Section 2.1 and Appendix C in the Supplementary Material ([Hargreaves et al. \(2019\)](#)) for details). Within our motivating datasets, the data naturally shared the same starting point (see Appendix A in the Supplementary Material ([Hargreaves et al. \(2019\)](#))). As our methodological development is motivated by experimental data, we assume all signals are of a common length  $T$ .

Thus denote each individual profile by  $\{X_{t,T}^{(i),r_i}\}_{t=0}^{T-1}$  with  $i = 1, 2$  corresponding to one of two groups (e.g., control/ treatment) and potential replicates  $r_i = 1, \dots, N_i$  (i.e.,  $N_i$  circadian traces in the  $i$ th group). Note that when  $N_i = 1$  we drop the  $r_i$  index for simplicity. Assume the signals in group  $i$  are underpinned by a common wavelet spectrum and denote this by  $S_j^{(i)}(t/T)$  for each group  $i = 1, 2$  at scales  $j \in \overline{1, J}$  ( $J = \log_2 T$ ) and rescaled times  $z = t/T \in (0, 1)$ .

3.1. *Lead dataset: Hypothesis testing for spectral equality ('WST' and 'FT')*. Put simply, our soil pollutant example focussed on detecting whether the two plant groups, 'Control' and 'Lead', display significant differences in the evolution of their spectral structures and, if so, the particular scales and times at which such differences occur. Mathematically we formalise our hypotheses as

$$(3.1) \quad H_0 : S_j^{(1)}(z) = S_j^{(2)}(z) \quad \forall j, z$$

versus the alternative  $H_A : S_{j^*}^{(1)}(z^*) \neq S_{j^*}^{(2)}(z^*)$  for some scale  $j^*$  and rescaled time  $z^*$ . In the time domain, we visually note that differences in the circadian rhythms of the two groups appear towards the end of the experiment (see Figure 1).

3.1.1. *A naive wavelet spectrum test ('WST')*. Since in reality we do not know the group spectrum  $S_j^{(i)}(z)$ , we replace it with a well-behaved estimator, denoted  $\hat{S}_j^{(i)}(z)$ . Assuming independent replicates are available for each group, we use the group ( $i = 1, 2$ ) averaged spectral estimators

$$(3.2) \quad \hat{S}_j^{(i)}(k/T) = \frac{1}{N_i} \sum_{r_i=1}^{N_i} L_j^{(i),r_i}(k/T),$$

where  $L_j^{(i),r_i}(k/T)$  is the empirical wavelet spectrum of the  $r_i$ th series in group  $i$  at scale  $j$  and time  $k$ . Assuming independence across the replicates and a Gaussian distribution for the spectral estimates, because the LSW theory constructs asymptotically unbiased spectral estimators, it follows that under the null hypothesis  $\hat{S}_j^{(1)}(k/T) - \hat{S}_j^{(2)}(k/T)$  has an asymptotically normal distribution with mean zero. Hence, should our spectral estimators satisfy the classical assumptions for a  $t$ -test (which in our context amount to independence of the spectral estimates across replicates and a Gaussian distribution), we propose a naive *wavelet spectrum test* (WST), centred on a test statistic of the form

$$(3.3) \quad T_{j,k} = \frac{\hat{S}_j^{(1)}(k/T) - \hat{S}_j^{(2)}(k/T)}{((\hat{\sigma}_{j,k}^{(1)})^2/N_1 + (\hat{\sigma}_{j,k}^{(2)})^2/N_2)^{1/2}} \sim t_{df} \quad \text{under the null hypothesis,}$$

where  $(\hat{\sigma}_{j,k}^{(i)})^2$  is an estimate of the variance of  $\hat{S}_j^{(i)}(k/T)$  for  $i = 1, 2$  across the  $N_i$  observations in group  $i$ , obtained using the standard sum-of-squares sample variance formula (as in Krzemieniewska, Eckley and Fearnhead (2014)). Under the

null hypothesis of spectral equality,  $T_{j,k}$  (asymptotically) follows a  $t$ -distribution with the number of degrees of freedom ( $df$ ) directly related to the variance estimation procedure we employ. Each test statistic is then compared with a critical value derived from the  $t$ -distribution in the usual way.

When the variance of  $\hat{S}_j^{(i)}(k/T)$  is unknown but common to both  $i = 1, 2$  groups (denoted  $(\sigma_{j,k})^2 := (\sigma_{j,k}^{(1)})^2 = (\sigma_{j,k}^{(2)})^2$ ), it can be estimated using the pooled estimator:

$$(3.4) \quad \hat{\sigma}_{j,k}^2 = \frac{(N_1 - 1)(\hat{\sigma}_{j,k}^{(1)})^2 + (N_2 - 1)(\hat{\sigma}_{j,k}^{(2)})^2}{N_1 + N_2 - 2},$$

replacing  $(\hat{\sigma}_{j,k}^{(1)})^2$  and  $(\hat{\sigma}_{j,k}^{(2)})^2$  in equation (3.3). The number of degrees of freedom in the  $t$ -distribution of the test statistic is then  $df = N_1 + N_2 - 2$ .

If there is no reason to believe the group variances are equal, then use a  $t$ -distribution with degrees of freedom

$$df = \frac{((\hat{\sigma}_{j,k}^{(1)})^2/N_1 + (\hat{\sigma}_{j,k}^{(2)})^2/N_2)^2}{\frac{((\hat{\sigma}_{j,k}^{(1)})^2/N_1)^2}{N_1 - 1} + \frac{((\hat{\sigma}_{j,k}^{(2)})^2/N_2)^2}{N_2 - 1}}.$$

However, the test statistic does not exactly follow the  $t$ -distribution, since two standard deviations are estimated in the statistic. Conservative critical values may also be obtained by using the  $t$ -distribution with  $N$  degrees of freedom, where  $N$  represents the smaller of  $N_1$  and  $N_2$  (Moore (2007)).

In practice, the spectral estimators in equation (3.2) may breach the Gaussianity testing assumption, especially when only a low number of replicates are available. The assumption of approximate normality for individual replicate spectral estimates, cautiously used in Fryzlewicz and Ombao (2009), will be strengthened by the presence of a higher collection of group replicates ( $N_1, N_2$ ) (see Section 4 for a discussion of WST’s features and caveats).

3.1.2. *Raw periodogram F-Test ('FT').* We now construct a testing procedure that is not reliant on the Gaussianity assumption whose validity we challenged above. Formally, for each scale  $j \in \mathbb{N}$  and rescaled time  $z \in (0, 1)$ , the spectral equality  $S_j^{(1)}(z) = S_j^{(2)}(z)$  is equivalent to  $\beta_j^{(1)}(z) = \beta_j^{(2)}(z)$  as the autocorrelation wavelet inner product matrix  $A$  that links the two (see equation (2.3)) is invertible. We therefore replace our initial collection of multiple hypothesis tests with equivalent reframed versions

$$H_0 : \beta_j^{(1)}(z) = \beta_j^{(2)}(z) \quad \forall j, z$$

against the alternative ( $H_A$ ) that there exist a scale  $j^*$  and rescaled time  $z^*$  such that  $\beta_{j^*}^{(1)}(z^*) \neq \beta_{j^*}^{(2)}(z^*)$ . In order to construct our test statistic, we test for spectral equality by examining the  $\beta_j(z)$  quantities instead.

In reality we do not know  $\beta_j^{(i)}(z)$  for  $i = 1, 2$  so we replace it by an asymptotically unbiased estimator. As data are available consisting of multiple time series with known group memberships, we replace  $\beta_j^{(i)}(z)$  with an estimate across the group replicates. Specifically, if we have  $N_i$  independent time series replicates from group  $i$ , we define

$$(3.5) \quad N_i \bar{I}_{j,k}^{(i)} := \sum_{r_i=1}^{N_i} I_{j,k}^{(i),r_i} \sim \beta_j^{(i)}(k/T) \chi_{N_i}^2.$$

The distribution above follows as the raw wavelet periodogram coefficient of each  $r_i$ th periodogram replicate  $I_{j,k}^{(i),r_i}$  is approximately (scaled)  $\chi_1^2$  distributed (e.g., Nason and Stevens (2015)) and independent of all other raw wavelet periodogram coefficients across all other replicates from the same group (also see Fryzlewicz and Ombao (2009) and the discussion in Section 2.1). Under the further assumption of group independence,  $\bar{I}_{j,k}^{(1)}$  and  $\bar{I}_{j,k}^{(2)}$  are independent and distributed as detailed in equation (3.5). Hence we propose the test statistic

$$(3.6) \quad F_{j,k} = \frac{\bar{I}_{j,k}^{(1)}}{\bar{I}_{j,k}^{(2)}} \sim F_{N_1, N_2} \quad \text{under the null hypothesis.}$$

Each test statistic is then compared with a critical value derived from the  $F_{N_1, N_2}$ -distribution in the usual way.

*Discussion.* An advantage of the FT, particularly as opposed to the WST, is that its underlying distributional assumption is theoretically, as well as practically, more reliable. We would therefore expect the FT to outperform the WST in many applications, and this is indeed validated across a variety of simulation settings (see Section 4).

As we wish to test many hypotheses of the type  $H_0 : \beta_j^{(1)}(k/T) = \beta_j^{(2)}(k/T)$  for several values of  $j$  and  $k$ , we are in the field of multiple-hypothesis testing. For all tests we develop, we use Bonferroni correction and, for a less conservative approach, the false discovery rate (FDR) procedure introduced by Benjamini and Hochberg (1995). Our simulations in Section 4 show that both these methods work well. However, of course the tests themselves are related to one another, but just as in Nason (2013) we do not pursue this topic further in this work.

The WST and FT developed above both report the time-scale locations of the significant differences between the two group spectra. These can be visualised as a ‘barcode’ plot, where a significant difference is represented by a black line at the time-scale location of the rejection of the null hypothesis (see, e.g., Figure 4, right). Alternatively, for all our proposed tests, practitioners can also be informed by the number of rejections (as a dissimilarity measure), with larger values indicating a greater departure from the null hypothesis (as discussed in Das and Nason (2016) and in Section 4.2).

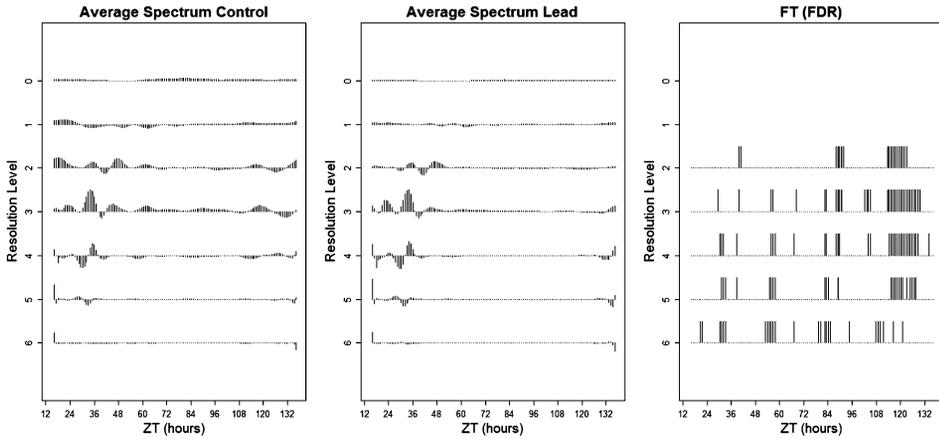


FIG. 4. Lead dataset. Left: Average estimated spectrum of the ‘Control’ group; Centre: Average estimated spectrum of the ‘Lead’ group; Right: ‘Barcode’ plot for FT (with FDR).

3.2. *Ultradian dataset: Hypothesis testing for spectral equality across scales* (‘HFT’). For certain biological applications, such as the Ultradian motivating example, it is more important to identify spectral differences between groups at scale-level and the time locations of spectral differences are of less interest. For such situations, we replace the spectral comparison  $H_0 : S_j^{(1)}(z) = S_j^{(2)}(z)$  of the previous section, in general equivalent to  $H_0 : \beta_j^{(1)}(z) = \beta_j^{(2)}(z)$ , by the comparison of the respective Haar–Fisz transforms, that is, test for

$$H_0 : \mathcal{F}\beta_j^{(1)}(z) = \mathcal{F}\beta_j^{(2)}(z) \quad \forall j, z.$$

Equivalently, in the notation established in Section 2.1 we test

$$(3.7) \quad H_0 : \mathcal{B}_j^{(1)}(z) = \mathcal{B}_j^{(2)}(z) \quad \forall j, z$$

versus the alternative ( $H_A$ ) that there exist some scale  $j^*$  and rescaled time  $z^*$  for which the equality does not hold. We shall refer to this test as the *Haar–Fisz test* (HFT). Intuitively, although the HFT identifies both scales and times at which the null hypothesis of spectral equality in the Haar–Fisz domain does not hold, as the Haar–Fisz transform essentially ‘averages’ within each scale of the raw wavelet periodogram, potential differences ‘spread’ throughout the scale. This property makes it ideal for identifying scale-level differences between group wavelet spectra (see, e.g., Figure 5, right).

As we do not know  $\mathcal{B}_j^{(i)}(z)$ , we replace it by its approximately unbiased estimator  $\mathcal{H}_{j,k}^{(i)}$  at scale  $j$  and time  $k$  (with  $z = k/T$ ) for group  $i = 1, 2$ . In applications which do not provide access to replicate data, we could adopt equation (3.3) with  $\hat{S}_j^{(i)}(k/T)$  replaced by  $\mathcal{H}_{j,k}^{(i)}$  and estimate the variance across each scale as

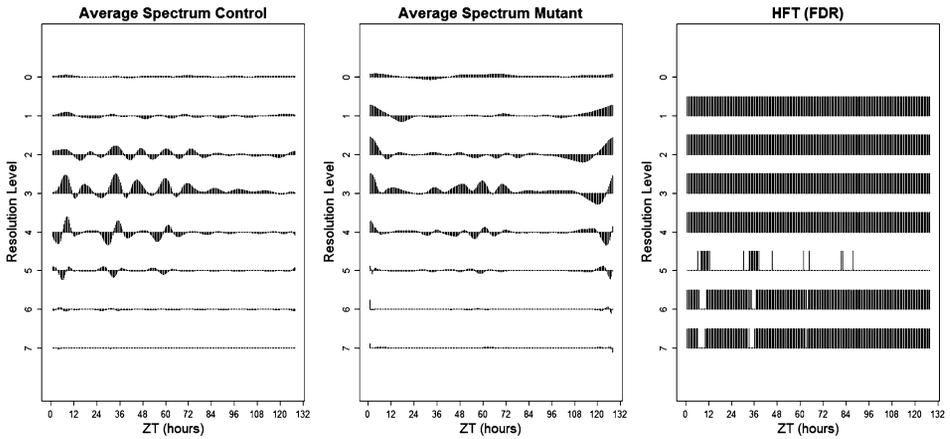


FIG. 5. Ultradian dataset. *Left*: Average estimated spectrum of the ‘Control’ group; *Centre*: Average estimated spectrum of the ‘Mutant’ group; *Right*: ‘Barcode’ plot for HFT (with FDR).

the Haar–Fisz transform stabilises variance (Nason and Stevens (2015)) (see Appendix D in the Supplementary Material (Hargreaves et al. (2019))). When replicates are available, we use equation (3.2) with  $\mathcal{H}_{j,k}^{(i)}$  to obtain group averaged estimators of  $\mathcal{B}_j^{(i)}(z)$ , denoted  $\hat{\mathcal{H}}_{j,k}^{(i)}$ , and propose a test statistic as in equation (3.3) with  $\hat{S}_j^{(i)}(k/T)$  replaced by  $\hat{\mathcal{H}}_{j,k}^{(i)}$ . The variance estimation techniques and subsequent test statistic distribution follow as detailed in Section 3.1 and the results of the HFT can also be visualised as a ‘barcode’ plot.

The rationale of this approach is also to bring the data (in this context, the Haar–Fisz transform of the raw wavelet periodogram) closer to Gaussianity and to break the dependencies across time. Consequently, the assumptions behind the  $t$ -test are closely adhered to and the dependencies between the multiple tests we perform are weak. In practice, due to its scale averaging construction, the HFT unsurprisingly results in many more time-localised rejections than the actual number of differing coefficients in the original spectra, and does sometimes have difficulty discriminating between spectra which differ by a small number of coefficients; however, the HFT does correctly identify scale-level spectral differences (see Section 4 for further investigations).

3.3. *Nematode dataset: Hypothesis testing for ‘same shape’ spectra (‘HT’)*. In applications such as the Nematode example, the focus may be on identifying whether groups evolve according to spectra that have the same shape at each scale, thus indicating that the same patterns are identified in the data, albeit with potentially different magnitudes.

Mathematically, for a scale-dependent (nonzero) constant denoted by  $C_j$ , we formalise our hypotheses as

$$(3.8) \quad H_0 : S_j^{(1)}(z) = S_j^{(2)}(z) + C_j \quad \forall j, z$$

versus the alternative  $H_A : S_{j^*}^{(1)}(z^*) \neq S_{j^*}^{(2)}(z^*) + C_{j^*}$  for some scale  $j^*$  and time  $z^*$ .

Denoting by  $\underline{C}$ , the  $J \times 1$  vector that holds  $C_j$  as its  $j$ th component and recalling equation (2.3), we can equivalently reframe the problem into testing whether

$$H_0 : \beta_j^{(1)}(z) = \beta_j^{(2)}(z) + c_j \quad \text{or equivalently} \quad H_0 : \beta_j^{(D)}(z) = c_j \quad \forall j, z,$$

where  $c_j$  is the  $j$ th entry of the vector  $\underline{c} = A\underline{C}$  and  $\beta_j^{(D)}(z) := \beta_j^{(1)}(z) - \beta_j^{(2)}(z)$ .

In the spirit of the tests developed in Fan and Lin (1998), and as undertaken by von Sachs and Neumann (2000) and Nason (2013), at each scale  $j$  we assess the constancy through time of  $\beta_j^{(D)}(z)$  by examining its associated Haar wavelet coefficients. Although, in principle, any wavelet system could be adopted, von Sachs and Neumann (2000) note that the Haar wavelet coefficients are ideal for testing the constancy of a function. Hence we employ these wavelets and refer to the test developed in this section as the *Haar Test* (HT).

The underlying principle behind these tests is that the wavelet transform of a constant function is zero, hence under  $H_0$  above, the wavelet coefficients of  $\beta_j^{(D)}(z)$  are

$$v_{\ell,p}^j = \int_0^1 \beta_j^{(D)}(z) \psi_{\ell,p}^H(z) dz = c_j \int_0^1 \psi_{\ell,p}^H(z) dz = 0,$$

where  $\{\psi_{\ell,p}^H(z)\}_{\ell,p}$  denote the usual Haar wavelets at scale  $\ell$  and location  $p$ .

This suggests performing multiple hypothesis testing on the collection of hypotheses

$$H_0 : v_{\ell,p}^j = 0 \quad \forall j, \ell \text{ and } p$$

against the alternative ( $H_A$ ) that there exist  $j^*$ ,  $\ell^*$  and  $p^*$  such that  $v_{\ell^*,p^*}^{j^*} \neq 0$ .

As the spectral and related quantities are unknown, and since the wavelet transform is linear, we estimate each  $v_{\ell,p}^j$  by  $\hat{v}_{\ell,p}^j = \hat{v}_{\ell,p}^{j,(1)} - \hat{v}_{\ell,p}^{j,(2)}$ , with the Haar wavelet coefficients corresponding to each group  $i = 1, 2$  estimated in the spirit of Nason (2013) as

$$(3.9) \quad \hat{v}_{\ell,p}^{j,(i)} = 2^{-\ell/2} \left( \sum_{r=0}^{2^{\ell-1}-1} I_{j,2^\ell p-r}^{(i)} - \sum_{q=2^{\ell-1}}^{2^\ell-1} I_{j,2^\ell p-q}^{(i)} \right),$$

at each (original) scale  $j$  and Haar scale  $\ell$  and locations  $p, q$ .

With the availability of independent replicates within each group, we estimate the group  $i$  Haar wavelet coefficients as

$$(3.10) \quad \hat{v}_{\ell,p}^{j,(i)} = \frac{1}{N_i} \sum_{r_i=1}^{N_i} \hat{v}_{\ell,p}^{j,(i),r_i},$$

where each  $\hat{v}_{\ell,p}^{j,(i),r_i}$  is obtained as in equation (3.9) for the  $r_i$ th replicate.

Under a specific set of assumptions, [Nason \(2013\)](#) shows the asymptotic normality of the Haar wavelet coefficient estimator of the wavelet periodogram at scale  $j$ . Thus, in our setting, each  $\hat{v}_{\ell,p}^{j,(i),r_i}$  for  $i = 1, 2$  is asymptotically normal with mean  $v_{\ell,p}^{j,(i),r_i}$  and variance  $(\sigma_{\ell,p}^{j,(i)})^2$ . Using the replicate independence, we have that  $\hat{v}_{\ell,p}^{j,(i)}$  is asymptotically normally distributed with mean  $v_{\ell,p}^{j,(i)}$  and variance  $(\sigma_{\ell,p}^{j,(i)})^2/N_i$  and note that its distributional closeness to the normal increases via a central limit theorem argument with the increasing number of replicates.

The group independence assumption then leads to an asymptotically joint normal distribution for  $(\hat{v}_{\ell,p}^{j,(1)}, \hat{v}_{\ell,p}^{j,(2)})$ . Following the continuous mapping theorem, we obtain that  $\hat{v}_{\ell,p}^j = \hat{v}_{\ell,p}^{j,(1)} - \hat{v}_{\ell,p}^{j,(2)}$  has an asymptotic normal distribution with mean  $v_{\ell,p}^{j,(1)} - v_{\ell,p}^{j,(2)}$  and variance  $((\sigma_{\ell,p}^{j,(1)})^2/N_1 + (\sigma_{\ell,p}^{j,(2)})^2/N_2)$ .

In the presence of replicates, we propose a test statistic of the form discussed in equation (3.3)

$$(3.11) \quad T_{\ell,p}^j = \frac{\hat{v}_{\ell,p}^j}{((\hat{\sigma}_{\ell,p}^{j,(1)})^2/N_1 + (\hat{\sigma}_{\ell,p}^{j,(2)})^2/N_2)^{1/2}} \sim t_{df} \quad \text{under the null hypothesis,}$$

where  $(\hat{\sigma}_{\ell,p}^{j,(i)})^2$  is an estimate of the variance of  $\hat{v}_{\ell,p}^{j,(i)}$  for  $i = 1, 2$  across the  $N_i$  observations in group  $i$ , obtained using the standard sum-of-squares sample variance formula and  $df$  denotes the degrees of freedom associated with the variance estimation procedure (see Section 3.1.1). Each test statistic is then compared with a critical value derived from the  $t$ -distribution in the usual way.

In order to control the asymptotic bias derivation, one of the assumptions under which the distributional theory is derived consists of limiting the scales of the Haar wavelet coefficients  $v_{\ell,p}^j$  to be sufficiently coarse,  $\ell = 0, \dots, (J - \lceil J/2 \rceil - 2)$ . Furthermore, as in [Nason \(2013\)](#), we only consider the wavelet coefficients of the periodogram at levels  $j \geq 3$  in order to avoid the effects of a region similar to the ‘cone of influence’ described by [Torrence and Compo \(1998\)](#).

To aid the visualisation of the WST, FT and HFT results, we use a ‘barcode’ plot that indicates the time- and scale- locations where significant differences are present. The HT can also indicate where the significant differences are located

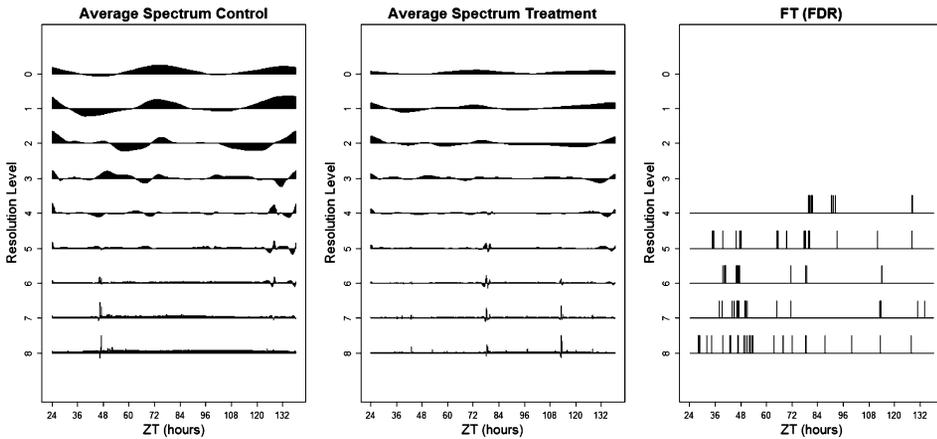


FIG. 6. Nematode dataset. *Left*: Average estimated spectrum of the ‘Control’ group; *Centre*: Average estimated spectrum of the ‘Treatment’ group; *Right*: ‘Barcode’ plot for FT (with FDR).

in the series and can plot the results in a manner similar to the wavelet test of stationarity (see [Nason \(2013\)](#)). However, due to its construction, these locations are more difficult to interpret than for the WST, FT and HFT (see [Figure 6](#)).

**4. Simulation studies.** The goals of the simulation studies were: (1) to evaluate the empirical power and size of our new tests; (2) to consider the effect of sample size on the accuracy of the tests; (3) to investigate two approaches to multiple-hypothesis testing: Bonferroni correction (denoted ‘Bon.’) and the false discovery rate procedure (‘FDR’); (4) to investigate the performance of our proposed tests when certain modelling assumptions are broken and (5) to evaluate the empirical power and size of our new tests in comparison with the adaptive Neyman Test (ANT) of [Fan and Lin \(1998\)](#) (see [Section 2.2](#)). This benchmark method performs well in practice when the assumption that the data can be modelled as a functional time series is valid.

In this section we briefly outline the basic structure of each simulated experiment (a comprehensive description of the simulation studies can be found in [Appendix D](#) in the [Supplementary Material \(Hargreaves et al. \(2019\)\)](#)). In each case, we assumed that the signal was a realisation from one of  $i = 1, 2$  possible groups. For each group, we generated a set of  $N_1 = N_2 = 1, 10, 25, 50$  signal realisations of common length  $T = 256$ , the equivalent of a free-running period of four days. For each realisation, we obtained the raw and corrected wavelet periodograms using (unless otherwise stated) the Haar wavelet (from the `locits` software package for R—available from the CRAN package repository) although, in principle, any wavelet system can be used (see [Section 4.3](#)). The Haar-transformed and Haar-Fisz transformed raw wavelet periodogram were subsequently obtained and the spectral testing procedures carried out as described in [Section 3](#). The results

are compared with the known group memberships, and the procedure is then repeated 1000 times to obtain empirical size and power estimates as outlined in the following sections.

4.1. *Power comparisons.* To explore statistical power we simulate a set of  $N_1 = N_2 = 1, 10, 25, 50$  signal realisations from each group where the individual group spectra are defined such that there exists a scale  $j^*$  and time  $t^*$  such that  $S_{j^*}^{(1)}(t^*/T) \neq S_{j^*}^{(2)}(t^*/T)$ . The empirical power estimates are obtained by counting the number of times our tests reject the null hypothesis of spectral equality. The models we will use are denoted P1–P12 respectively and are briefly described below (details can be found in Appendix D in the Supplementary Material (Hargreaves et al. (2019))).

1. *P1: Fixed Spectra.* We follow Krzemieska, Eckley and Fearnhead (2014) and design the spectra of the two groups to differ at the finest level (resolution level 7) by 100 coefficients.

2. *P2: Fixed Spectra-Fine Difference.* We modify the model P1 by fixing ‘Group 1’ but defining the spectrum of ‘Group 2’ such that the spectra of the two groups now differ by only 6 coefficients.

3. *P3: Fixed Spectra-Plus Constant.* Modify the model P1 by fixing ‘Group 1’ but defining the spectrum of ‘Group 2’ such that the spectra of the two groups differ by a constant in the finest resolution level.

4. *P4/P5: Gradual Period Change.* This study replicates a typical circadian experiment with changes that cannot be captured by standard analyses assuming stationarity and only reporting an average period value. We thus define 3 possible groups, where each group represents a signal that gradually changes period from 24 to: 25 (Group 1), 26 (Group 2) and 27 (Group 3) over (approximately) two days, before continuing with the relevant period for a further two days (also see Hargreaves et al. (2018)). To determine which changes can be discriminated by the methods, we perform two studies within this setting: simulations from Groups 1 and 2 (P4) and simulations from Groups 1 and 3 (P5).

5. *P6/P7: AR Processes with time-varying coefficients.* We simulate from an important class of nonstationary processes—AR(2) processes with: abruptly (P6) and slowly (P7) changing parameters (as in Fryzlewicz and Ombao (2009)).

6. *P8–P12: Functional Time Series (Constant Period).* This study follows Zielinski et al. (2014) and generates each time series using an underlying cosine curve with additive noise, which also coincides with the theoretical assumptions of the ANT. We define time series as realisations from one of six possible groups, each with a different (constant) period, relevant to our circadian setting. To determine which period changes can be discriminated by the methods, we perform five studies within this setting: simulations from a group with a period of 24 hours versus a group with a period of 21, 22, 23, 23.5 and 23.75 hours (models P8–P12 respectively).

TABLE 1  
*Simulated power estimates (%) for models P1–P7 with nominal size of 5% with  $N_1 = N_2 = 25$  realisations from each group. Highest empirical power estimates are highlighted in bold*

Model	WST (Bon.)	WST (FDR)	FT (Bon.)	FT (FDR)	HFT (Bon.)	HFT (FDR)	HT (Bon.)	HT (FDR)
P1	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
P2	39.3	48.0	<b>100.0</b>	<b>100.0</b>	29.1	31.8	86.2	86.4
P3	100.0	100.0	100.0	100.0	100.0	100.0	4.3	4.4
P4	1.0	2.7	45.5	54.5	33.2	36.5	<b>100.0</b>	<b>100.0</b>
P5	5.9	14.6	97.0	99.9	<b>100.0</b>	<b>100.0</b>	<b>100.0</b>	<b>100.0</b>
P6	<b>100.0</b>	<b>100.0</b>	87.5	92.6	44.8	89.1	66.5	67.7
P7	<b>100.0</b>	<b>100.0</b>	54.3	64.5	97.4	99.9	<b>100.0</b>	<b>100.0</b>

4.1.1. *Discussion of findings.* The empirical power values for  $N_1 = N_2 = 25$  (this is the typical number of available replicates in circadian studies; see Appendix A in the Supplementary Material (Hargreaves et al. (2019))) for models P1–P7 are reported in Table 1. We found that all tests perform well when the spectra differ by a large number of coefficients (model P1). The FT (and, to a lesser extent, the HT) are able to discriminate between spectra that differ by a small number of coefficients (model P2) whereas the HFT has lower empirical power. By construction, the HT cannot differentiate between spectra that differ by a constant at a particular resolution level (model P3), but we found that the HT performs well in our synthetic circadian example of gradual small period change across many time-scale locations (models P4 and P5). Due to the higher distributional reliability of the FT, it unsurprisingly outperforms the WST when the time series are generated from a defined spectrum (models P1–P5). However, distributional properties of the time-varying AR process ensure that the WST performs best when data are generated using models P6 and P7, with the HT and HFT also performing well for model P7.

*Effect of Sample Size.* The number of replicates in each group ( $N_1, N_2$ ) is also an important factor in achieved power. The results for the HFT with  $N_1 = N_2 = 1$  are shown in Table S6 (Appendix D.2 in the Supplementary Material (Hargreaves et al. (2019))), since we recall that the HFT is the only proposed test which can be applied when replicate data is not available—see Section 3.2. The results for all tests with  $N_1 = N_2 = 10$  and 50 replicates are shown in Table S7 (Appendix D.2 in the Supplementary Material (Hargreaves et al. (2019))). Increasing the number of replicates should, and indeed does, increase the empirical power of all tests (with the exception of the HT for model P3). For example, note the increase in empirical power (particularly for models P2 and P4) as the number of replicates increases from 10 to 25.

*Approach to Multiple-hypothesis Testing.* These studies show that the Bonferroni correction provides a more conservative approach. The false discovery rate

TABLE 2

Performance Comparison: *Simulated power estimates (%) for models P8–P12 with nominal size of 5% with  $N_1 = N_2 = 25$  realisations from each group and using the false discovery rate procedure (FDR). Note: Control group period is 24 hours in each model*

Model	Test Group Period	WST (FDR)	FT (FDR)	HFT (FDR)	HT (FDR)	ANT
P8	21	100.0	100.0	100.0	100.0	100.0
P9	22	100.0	100.0	100.0	100.0	100.0
P10	23	100.0	100.0	92.0	100.0	100.0
P11	23.5	100.0	100.0	31.8	100.0	100.0
P12	23.75	100.0	97.9	9.1	98.3	100.0

gives an empirical power greater than (or equal to) that of the Bonferroni correction (see, e.g., model P6 in Table 1).

*Performance Comparison.* We also report that the empirical power of the ANT for model P5 (gradual period change, 25 replicates) was 10.7%, which is below the results in Table 1 for our proposed tests. This is to be expected as the underlying assumptions of the ANT are no longer met. (Similar results are obtained for models P1–P7, hence we do not provide these here.)

Table 2 presents a selection of the performance comparison results for models P8–P12 when  $N_1 = N_2 = 25$ . (The results for all tests with  $N_1 = N_2 = 10$  replicates are also shown in Table S8, Appendix D.2 in the Supplementary Material (Hargreaves et al. (2019)).) As expected, the ANT performs extremely well in all these studies since the underlying assumptions of the methodology are adhered to. Nevertheless, it is encouraging that the WST, FT and HT also all have an empirical power over 95% (25 replicates) showing that our methodology can also be successfully applied to functional time series as designed for the ANT. However, the HFT had difficulty discriminating between groups when the period difference was less than two hours. This was no surprise as the HFT was constructed to detect differences in scale only and, due to the lower frequency resolution of the wavelet spectrum, the total power within each scale of the wavelet spectrum will be very similar for both groups.

4.1.2. *Power comparisons: Conclusions.* In practice, the suitability of the testing procedures is determined by a combination of factors, such as the practical problem posed by scientists, the degree to which the data adheres to the underlying theoretical assumptions and the number of available replicates. For example, models P1–P3 all stem from a simulated LSW structure and thus would be subject to a test for time-scale equality departure, carried out through an ‘FT’, as its theoretical assumptions are closely adhered to. Recall that the ‘WST’ was proposed as a ‘naive’ variant and is heavily reliant on the number of replicates in order to achieve the appropriate distributional properties; thus, its best results are obtained

for models that have been simulated from time-varying AR processes. Meanwhile, for data following models that exhibit a gradual period change (such as P4–P5) one might be interested in identifying scale-dependent patterns or discrepancies, carried out through the ‘HT’ or ‘HFT’.

4.2. *Size comparisons.* To explore statistical size, we simulate data from a number of models and we assess how often our hypothesis tests reject the null hypothesis of spectral equality (i.e., the time series are generated in the same way for both test groups). The models we will use are denoted M1–M5 respectively and are briefly described below (details can be found in Appendix D in the Supplementary Material (Hargreaves et al. (2019))).

1. *M1: Fixed Spectra.* We simulate all data from the wavelet spectrum associated with Group 1 in models P1, P2 and P3, which we define as  $\{S_j^{(1)}(z)\}_{j=1}^J$  in equation (D.1).

2. *M2: Gradual Period Change.* We simulate all data from the wavelet spectrum which corresponds to a time series that gradually changes period from 24 to 25 hours (over approximately two days), before continuing with period 25 hours for a further two days (i.e., Group 1 from models P4/P5).

3. *M3: AR Processes With Abruptly Changing Parameters.* Each time series is generated from the process defined by equation (D.5) with the abruptly changing parameters as defined for group  $i = 1$  in Table S4 (i.e., Group 1 from model P6).

4. *M4: AR Processes With Slowly Changing Parameters.* Each time series is generated from the process defined by equation (D.6) with the slowly changing parameters as defined for group  $i = 1$  in Table S5 (i.e., Group 1 from model P7).

5. *M5: Functional Time Series (Constant Period).* All data are simulated (using equation (D.7) from the model that corresponds to a time series with a constant period of 24 hours (i.e., Group 1 from models P8–P12).

4.2.1. *Discussion of findings.* The empirical size values for models M1–M4 with  $N_1 = N_2 = 25$  (this is the typical number of available replicates in circadian experiments; see Appendix A in the Supplementary Material (Hargreaves et al. (2019))) are reported in Table 3. The results for the HFT with  $N_1 = N_2 = 1$  are shown in Table S6, Appendix D.2 in the Supplementary Material (Hargreaves et al. (2019)). (Recall: the HFT is the only proposed test which can be applied when replicate data is not available—see Section 3.2). The results for all tests with  $N_1 = N_2 = 10$  and 50 replicates are shown in Table S9 (Appendix D.2 in the Supplementary Material (Hargreaves et al. (2019))).

These studies show that the empirical size corresponding to all proposed tests (apart from the FT for model M4 with  $N_1 = N_2 = 10$  and 25) are less than the nominal size of 5%. A close inspection of rejections for the FT for model M4 with  $N_1 = N_2 = 10$  and 25 and both multiple-hypothesis testing methods (Table S10 in Appendix D.2 in the Supplementary Material (Hargreaves et al. (2019))) reveals

TABLE 3

*Simulated size estimates (%) for models M1–M4 with nominal size of 5% and  $N_1 = N_2 = 25$  realisations from each group. Empirical size estimates over the nominal size of 5% are highlighted in bold*

Model	WST (Bon.)	WST (FDR)	FT (Bon.)	FT (FDR)	HFT (Bon.)	HFT (FDR)	HT (Bon.)	HT (FDR)
M1	0.6	1.3	2.5	3.1	0.1	2.0	2.3	2.7
M2	0.3	0.6	3.0	3.9	0.4	3.3	2.5	2.7
M3	0.2	1.5	3.6	3.9	0.0	1.6	3.5	3.8
M4	0.4	0.9	4.6	<b>5.2</b>	1.0	2.4	3.4	3.8

that, for this particular example, the number of rejections is often 1. If we disregard such situations, the empirical size of the FT also falls below the nominal size of 5% for all sample sizes and multiple-hypothesis testing procedures. In practice, circadian scientists are mostly interested in the numbers of rejections and their locations and often choose to disregard situations where very few coefficients are significantly different. Indeed, this is also our approach in Section 5.

*Effect of Sample Size.* Note that the tests scale well with increasing sample size, with the nominal size acting as an upper bound, a behaviour also present in other related empirical size investigations; see, for example, [Cho \(2016\)](#).

*Approach to Multiple-hypothesis Testing.* These studies show that the Bonferroni correction provides a more conservative approach, whereas the false discovery rate (using the correction outlined above) is closer to the nominal size.

*Performance Comparison.* The results for model M5 with  $N_1 = N_2 = 10$  and 25 are shown in Table S8 (Appendix D.2 in the Supplementary Material ([Hargreaves et al. \(2019\)](#))). Note that the empirical size estimates for our proposed tests are all lower than the nominal size of 5%, whereas for 10 replicates the empirical size of the ANT is 7.9%.

**4.2.2. Size comparisons: Conclusions.** These studies show that the empirical size corresponding to all proposed tests is less than the nominal size of 5% (apart from the FT for model M4 with  $N_1 = N_2 = 10$  and 25—where, in most cases, the number of significant coefficients was less than 5). We thus recommend using the less conservative FDR procedure (ignoring situations with very small numbers of rejections). Note this also yields better results for empirical power (see Section 4.1.1) whilst also remaining below the nominal size.

**4.3. Sensitivity analysis.** In this section we investigate the sensitivity of our proposed tests to certain modelling assumptions. We investigate: (1) departures from the normality assumption and (2) impact of the choice of wavelet family used within the spectral estimation procedures of each of our proposed tests. Throughout this section, we use  $N_1 = N_2 = 25$ , since this is the typical number of available

replicates in circadian experiments (see Appendix A in the Supplementary Material (Hargreaves et al. (2019))).

4.3.1. *Departures from normality.* Recall the proposed statistical testing methodology assumes the innovations  $\{\xi_{j,k}\}$  to be normally distributed. To investigate the impact of this assumption, we computationally assess the power and size of the proposed tests within the settings outlined in Section 4 for models P1–P5 and M1–M2 but simulated using non-Gaussian innovations (specifically following a  $t$ -distribution with 5, and subsequently 3, degrees of freedom). The results can be found in Table S11 (Appendix D.2 in the Supplementary Material (Hargreaves et al. (2019))). Unsurprisingly, when the normality assumption is broken, the empirical power of all tests is less than (or equal to) the empirical power when the innovations follow a standard normal distribution. The increasing distributional departure from normality appears to be of little relevant influence when testing data simulated from models P1 and P3 (across all tests), while the empirical power drops for the HT corresponding to models P2 and P4/P5. The testing procedures break for models P4/P5 with  $t_3$ -distributed innovations as, intuitively, the presence of heavier innovations make the gradual period change structure of models P4/P5 very difficult to discriminate. We also note that the HT is heavily reliant on the distributional assumptions (see Section 3.3) which explains its sensitivity. Due to its construction (see Section 3.1.2), the FT appears to more readily reject the null hypothesis, increasing the empirical size of the test. However, if we disregard situations where there are a very low number of rejections (see Section 4.2.1) the empirical size of the FT falls below the nominal size of 5% for both multiple-hypothesis testing procedures and all studies (other than M1 with FDR). We report here that the empirical power of the ANT for model P1 (fixed spectra) with  $t$ -distributions with 5 degrees of freedom was 6.8%, which is below the results in Table S11 for all our proposed tests (which are all over 99.9%). This is to be expected since, as in Section 4.1, the underlying assumptions of the ANT are not valid. (Similar results are obtained for models P2–P7, hence we do not provide these here.)

We also investigated the power and size for models P8–P12 and M5 (see Section 4) simulated using non-Gaussian errors (specifically following  $t$ -distributions with 5, and subsequently 3, degrees of freedom). The results can be found in Table S12 (Appendix D.2 in the Supplementary Material (Hargreaves et al. (2019))). The WST, FT and HT appear to share a good degree of robustness as they all have an empirical power over 99% for models P8–P11, showing that our methodology can also be successfully applied to functional time series (as designed for the ANT) with non-Gaussian error. Akin to the previous results for the gradual period change models P4/P5, the distribution of the noise term does appear to have an adverse effect in model P12, where the difference between the periods of the two underlying signals is only 15 minutes. Across this study, the HFT was most affected. A possible explanation is that the HFT was constructed to detect differences

in scale only and, due to the lower frequency resolution of the wavelet spectrum, the total power within each scale of the wavelet spectrum will be very similar for both groups. This issue will have been compounded by the heavier tailed distribution of the noise term. We also report here that, in the settings of this study, the performance of ANT was sustained as its underlying assumptions are adhered to.

*4.3.2. Choice of wavelet.* The wavelet system gives a representation for non-stationary time series under which we estimate the wavelet spectrum and subsequently perform hypothesis testing. We investigated the sensitivity of our methods to the wavelet choice. For models P1–P5, the Haar wavelet was used for spectral estimation, but different, potentially mismatched wavelets were used to generate the processes from the spectrum: Haar wavelets, Daubechies' least-asymmetric wavelets with four vanishing moments and Daubechies' extremal phase wavelets with ten vanishing moments. Models P6–P12 were not generated from LSW spectra (see Section 4), hence we report the results when using a selection of wavelets for the empirical wavelet spectrum.

The results in Tables S13 and S14 (Appendix D.2 in the Supplementary Material (Hargreaves et al. (2019))) show that our methodology is fairly robust to the wavelet choice. The empirical size estimates all fall below the nominal size. The results indeed support the intuition that, as the scope of our work is to devise tests that locally identify dissimilarities between pairs of spectra, the short support overlaps of Haar wavelets counterbalance their otherwise reduced capacity of representing smooth signals.

*4.4. Summary of findings.* A summary of the hypothesis tests developed in this manuscript detailing the test name, its acronym, strengths and weaknesses can be found in Table S15 (Appendix E in the Supplementary Material (Hargreaves et al. (2019))).

**5. Real data analysis: Back to the motivating circadian datasets.** We now use our proposed methodology to analyse the motivating examples (Section 1). Prior to analysis, we investigate whether the normality assumption is tenable for each of our motivating datasets. The results (Appendix B.2 in the Supplementary Material (Hargreaves et al. (2019))) show that, for each of our motivating datasets, the normality assumption is appropriate. We then model each circadian trace as a (Gaussian) LSW process, estimate its corresponding group wavelet spectral representation and consequently construct the appropriate test statistic that aims to identify whether a departure towards a specific type of spectral difference is present or not (as described in Section 3). For each dataset, the corresponding number of rejections can be found in Table S3 (Appendix B.1 in the Supplementary Material (Hargreaves et al. (2019))), with corresponding representative 'barcode' plots in Figures 4, 5 and 6.

We also note here that the data naturally shared the same starting point and had the same length (see Appendix A in the Supplementary Material (Hargreaves et al. (2019))). Therefore, instances where these conditions are not satisfied are not the focus of this paper and we leave these issues for future research.

**5.1. Lead dataset.** Section 1.1.1 outlined the scientific aims to determine whether lead nitrate affects the circadian clock and, if so, to detect the times and scales at which any significant differences arise between the ‘Control’ and ‘Lead’ exposure groups. Therefore, we are particularly interested in the results of the FT. Table S3 shows the results for the FT and includes both the more conservative Bonferroni correction and FDR. In order to visualise the areas of null hypothesis rejection of spectral equality between the control and lead-exposure groups, both group average estimated spectra as well as the ‘barcode’ plot for the FT (with FDR) appear in Figure 4. Figure 4 indicates that the differences between the two spectra lie in resolution levels 2–4, directly corresponding to a circadian rhythm, with the number of rejections increasing with exposure time. We conclude that there is evidence that exposure to lead does affect the circadian clock of *A. thaliana*, and this change manifests itself after approximately three days of free-running conditions.

**5.2. Ultradian dataset.** Section 1.1.2 introduced this experiment and highlighted the need to detect whether any differences appear in the circadian and ultradian components of the ‘Control’ and ‘Mutant’ groups. Hence we are interested in the results of the HFT, specifically developed to identify the scales, rather than the times, at which potential differences arise. Table S3 shows the results for the HFT, including both the Bonferroni correction and FDR. The results indicate rejections of the null hypothesis of spectral equality between the control and mutant plants across a range of scales. The group average estimated spectra and ‘barcode’ plot for the HFT (with FDR) can be found in Figure 5. Note that the differences between the two spectra lie in the coarsest resolution levels 1–4, associated with circadian rhythms, and higher-frequency levels 6 and 7, corresponding to an ultradian rhythm. We conclude that there is evidence that the mutant plants have altered circadian and ultradian rhythms within *A. thaliana*.

**5.3. Nematode dataset.** The experiment in Section 1.1.3 aimed to elucidate the effect of a pharmacological treatment on the *C. elegans* clock. The average estimated spectra of the ‘Control’ and ‘Treatment’ groups in Figure 6 share a common profile but with differences in magnitude, indicating that the HT would be appropriate in this context. Table S3 shows that the HT found no significant difference between the shapes of the two spectra, but when tested for equality, the FT (with FDR) found multiple rejections of the null hypothesis of spectral equality between the ‘Control’ and ‘Treatment’ groups (refer to the ‘barcode’ plot in Figure 6). This provides evidence that the two spectra have the same profile within

each scale up to an additive nonzero constant. We thus conclude that there is evidence that the treatment significantly affects the intensity of the spectral behaviour, but not its pattern. The spectral differences are present at the highest frequencies (resolution levels 6–8) as an early response to the onset of treatment (prior to time  $T = 48$ ); see Figure 6.

*5.4. Discussion of results.* Overall, we recall that, for each of our motivating datasets, the established Fourier-based tests currently adopted within the circadian community found no significant difference between the groups (see Table S1 in Appendix B.1 in the Supplementary Material (Hargreaves et al. (2019))), even though qualitative differences are easily noted (see Section 1.1). This methodology assumes data stationarity, but for our motivating datasets we have shown that this assumption is not appropriate (see Table S2 in Appendix B.1 in the Supplementary Material (Hargreaves et al. (2019))). Our proposed methodology was able to detect the visually apparent differences between the motivating datasets when the current methodology could not (see Tables S3 and S1 in Appendix B.1 in the Supplementary Material (Hargreaves et al. (2019))). Due to the nonstationary character of the proposed approach, it also additionally indicates precise times and/or scales at which differences become manifest.

**6. Conclusions and further work.** This work was stimulated by a variety of challenging applications faced by the circadian–biology community, which is becoming increasingly aware of the nonstationary characteristics present in much of their data (Hargreaves et al. (2018), Zielinski et al. (2014), Leise et al. (2013)). Our methodology fills the gap in the current literature by developing and testing much needed tools for the formal spectral comparison of nonstationary data. Our methods are developed as testing procedures, analogous to the period analysis techniques currently adopted within the circadian community. Motivated by three complementary applications in circadian biology, our new methodology allows the identification of three specific types of spectral difference. Table S15 in Appendix E in the Supplementary Material (Hargreaves et al. (2019)), provides a summary of the hypothesis tests developed in this manuscript detailing their strengths and weaknesses.

The competitive performance of our methods was comparatively assessed in an extensive simulation study (Section 4). Additionally, when compared to existing methods currently adopted within the circadian community, our proposed tests were able to discriminate between real data sets (Table S3) where the current methodology could not (Table S1).

In the applications provided, we illustrated the important implications in further understanding the mechanisms behind the plant and nematode circadian clocks, and the environmental implications associated with soil pollution. However, we note that our methodology can readily be applied to other circadian datasets, as well as to data originating in other fields, as long as the data share the same dyadic

length ( $T$ ). This assumption is easily achievable for most experimental data, but for other setups might necessitate further specific treatments depending on the discrepancy between the number of observations.

In all of our proposed hypothesis tests, we wish to test many hypotheses of the type  $H_0 : S_j^{(1)}(k/T) = S_j^{(2)}(k/T)$  for several values of  $j$  and  $k$ . In this manuscript we tested the Bonferroni correction and, for a less conservative approach, the false discovery rate (FDR) procedure. We recommend the use of the FDR procedure, as this gave a higher empirical power and was closer to the nominal size in the simulation studies (see Section 4). However, the multiple-hypothesis testing methods we use do not account for the dependence of the spectral coefficients. The hypothesis tests developed in Sections 3.2 and 3.3 alleviate this problem by transforming the data to produce coefficients that are approximately uncorrelated but, as neither method fully decorrelates the data, multiple-hypothesis testing methods that take the dependence of the (transformed) spectral coefficients into account are an interesting avenue of further work.

#### SUPPLEMENTARY MATERIAL

**Appendix for “Wavelet spectral testing: Application to nonstationary circadian rhythms.”** (DOI: [10.1214/19-AOAS1246SUPP](https://doi.org/10.1214/19-AOAS1246SUPP); .pdf). The supplementary material contains Appendices A–E which provide additional details throughout this manuscript. In particular, Appendix A outlines the experimental details that led to the datasets introduced in Section 1.1 and subsequently analysed in Sections 5.1, 5.2 and 5.3. Appendix B contains: a summary of the output of the analysis of the motivating datasets in BRASS; the results of the Priestley–Subba Rao test of stationarity (for each time series) and the number of rejections for the relevant proposed hypothesis testing procedure, for each motivating example dataset. Appendix C provides the technical details that underline the LSW process model introduced by [Nason, von Sachs and Kroisandt \(2000\)](#). Appendix D gives a more detailed description of the simulation studies outlined in Section 4. Appendix E provides a summary of the hypothesis tests developed in this manuscript detailing the test name, its acronym, strengths and weaknesses for each of the proposed tests.

#### REFERENCES

- ATKINSON, A. D., HILL, R. R., PIGNATIELLO, J. J. JR., VINING, G. G., WHITE, E. D. and CHICKEN, E. (2017). Wavelet ANOVA approach to model validation. *Simul. Model. Pract. Theory* **78** 18–27.
- BELL-PEDERSEN, D., CASSONE, V. M., EARNEST, D. J., GOLDEN, S. S., HARDIN, P. E., THOMAS, T. L. and ZORAN, M. J. (2005). Circadian rhythms from multiple oscillators: Lessons from diverse organisms. *Nat. Rev. Genet.* **6** 544–556.
- BENJAMINI, Y. and HOCHBERG, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B* **57** 289–300. [MR1325392](#)

- BUJDOSO, N. and DAVIS, S. J. (2013). Mathematical modeling of an oscillating gene circuit to unravel the circadian clock network of *Arabidopsis thaliana*. *Front. Plant Sci.* **4** 3.
- CHO, H. (2016). A test for second-order stationarity of time series based on unsystematic subsamples. *Stat* **5** 262–277. [Paging previously given as 1–16]. [MR3584714](#)
- COSTA, M. J., FINKENSTÄDT, B., GOULD, P. D., FOREMAN, J., HALLIDAY, K. J., HALL, A. J. W. and RAND, D. A. (2011). Estimating periodicity of oscillatory time series through resampling techniques. Centre for Research in Statistical Methodology, Univ. Warwick.
- DAS, S. and NASON, G. P. (2016). Measuring the degree of non-stationarity of a time series. *Stat* **5** 295–305. [MR3589268](#)
- DUSIK, V., SENTHILAN, P. R., MENTZEL, B., HARTLIEB, H., WÜLBECK, C., YOSHII, T., RAABE, T. and HELFRICH-FÖRSTER, C. (2014). The MAP kinase p38 is part of *Drosophila melanogaster*'s circadian clock. *PLoS Genet.* **10** e1004565.
- EDWARDS, K. D., AKMAN, O. E., KNOX, K., LUMSDEN, P. J., THOMSON, A. W., BROWN, P. E., POKHILKO, A., KOZMA-BOGNAR, L., NAGY, F. et al. (2010). Quantitative analysis of regulatory flexibility under changing environmental conditions. *Mol. Syst. Biol.* **6** 424.
- FAN, J. and LIN, S.-K. (1998). Test of significance when data are curves. *J. Amer. Statist. Assoc.* **93** 1007–1021. [MR1649196](#)
- FOLEY, J. A., DEFRIES, R., ASNER, G. P., BARFORD, C., BONAN, G., CARPENTER, S. R., CHAPIN, F. S., COE, M. T., DAILY, G. C. et al. (2005). Global consequences of land use. *Science* **309** 570–574.
- FRYZLEWICZ, P. (2005). Modelling and forecasting financial log-returns as locally stationary wavelet processes. *J. Appl. Stat.* **32** 503–528. [MR2112368](#)
- FRYZLEWICZ, P. and NASON, G. P. (2006). Haar–Fisz estimation of evolutionary wavelet spectra. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **68** 611–634. [MR2301011](#)
- FRYZLEWICZ, P. and Ombao, H. (2009). Consistent classification of nonstationary time series using stochastic wavelet representations. *J. Amer. Statist. Assoc.* **104** 299–312. [MR2504379](#)
- FRYZLEWICZ, P., VAN BELLEGEM, S. and VON SACHS, R. (2003). Forecasting non-stationary time series by wavelet process modelling. *Ann. Inst. Statist. Math.* **55** 737–764. [MR2028615](#)
- GUO, W., DAI, M., Ombao, H. C. and VON SACHS, R. (2003). Smoothing spline ANOVA for time-dependent spectral analysis. *J. Amer. Statist. Assoc.* **98** 643–652. [MR2011677](#)
- HARANG, R., BONNET, G. and PETZOLD, L. R. (2006). WAVOS: A MATLAB toolkit for wavelet analysis and visualization of oscillatory systems. *BMC Research Notes* **5** 163.
- HARGREAVES, J. K., KNIGHT, M. I., PITCHFORD, J. W., OAKENFULL, R. J. and DAVIS, S. J. (2018). Clustering nonstationary circadian rhythms using locally stationary wavelet representations. *Multiscale Model. Simul.* **16** 184–214. [MR3749378](#)
- HARGREAVES, J. K., KNIGHT, M. I., PITCHFORD, J. W., OAKENFULL, R. J., CHAWLA, S., MUNNS, J. and DAVIS, S. J. (2019). Supplement to “Wavelet spectral testing: application to nonstationary circadian rhythms.” DOI:[10.1214/19-AOAS1246SUPP](#).
- HOLAN, S. H., WIKLE, C. K., SULLIVAN-BECKERS, L. E. and COCROFT, R. B. (2010). Modeling complex phenotypes: Generalized linear models using spectrogram predictors of animal communication signals. *Biometrics* **66** 914–924. [MR2758228](#)
- KILLICK, R., ECKLEY, I. A. and JONATHAN, P. (2013). A wavelet-based approach for detecting changes in second order structure within nonstationary time series. *Electron. J. Stat.* **7** 1167–1183. [MR3056071](#)
- KON, N., SUGIYAMA, Y., YOSHITANE, H., KAMESHITA, I. and FUKADA, Y. (2015). Cell-based inhibitor screening identifies multiple protein kinases important for circadian clock oscillations. *Commun. Integr. Biol.* **8** e982405.
- KRZEMIENIEWSKA, K., ECKLEY, I. A. and FEARNHEAD, P. (2014). Classification of non-stationary time series. *Stat* **3** 144–157.
- LEISE, T. L., INDIC, P., PAUL, M. J. and SCHWARTZ, W. J. (2013). Wavelet meets actogram. *J. Biol. Rhythms* **28** 62–68.

- MARTINEZ, J. G., BOHN, K. M., CARROLL, R. J. and MORRIS, J. S. (2013). A study of Mexican free-tailed bat chirp syllables: Bayesian functional mixed models for nonstationary acoustic time series. *J. Amer. Statist. Assoc.* **108** 514–526. [MR3174638](#)
- MCCLUNG, C. R. (2006). Plant circadian rhythms. *Plant Cell* **18** 792–803.
- MCKAY, J. L., WELCH, T. D. J., VIDAKOVIC, B. and TING, L. H. (2012). Statistically significant contrasts between EMG waveforms revealed using wavelet-based functional ANOVA. *J. Neurophysiol.* **109** 591–602.
- MILLAR, A. J., CARRINGTON, J. T., TEE, W. V. and HODGE, S. K. (2015). Changing planetary rotation rescues the biological clock mutant *lhy cca1* of *Arabidopsis thaliana*. *BioRxiv* at Cold Spring Harbor Laboratory.
- MINORS, D. S. and WATERHOUSE, J. M. (2013). *Circadian Rhythms and the Human*. Butterworth-Heinemann, Oxford.
- MOORE, D. S. (2007). *The Basic Practice of Statistics* **2**. WH Freeman, New York.
- MOORE, A., ZIELINSKI, T. and MILLAR, A. J. (2014). Online period estimation and determination of rhythmicity in circadian data, using the BioDare data infrastructure. *Methods Mol. Biol.* **1158** 13–44.
- MORRIS, J. S. (2015). Functional regression. *Ann. Rev. Stat. Appl.* **2** 321–359.
- MORRIS, J. S., BALADANDAYUTHAPANI, V., HERRICK, R. C., SANNA, P. and GUTSTEIN, H. (2011). Automated analysis of quantitative image data using isomorphic functional mixed models, with application to proteomics data. *Ann. Appl. Stat.* **5** 894–923. [MR2840180](#)
- NASON, G. P. (2008). *Wavelet Methods in Statistics with R. Use R!* Springer, New York. [MR2445580](#)
- NASON, G. (2013). A test for second-order stationarity and approximate confidence intervals for localized autocovariances for locally stationary time series. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **75** 879–904. [MR3124795](#)
- NASON, G. P. and SAVCHEV, D. (2014). White noise testing using wavelets. *Stat* **3** 351–362.
- NASON, G. P. and STEVENS, K. (2015). Bayesian wavelet shrinkage of the Haar–Fisz transformed wavelet periodogram. *PLoS ONE* **10** e0137662.
- NASON, G. P., VON SACHS, R. and KROISANDT, G. (2000). Wavelet processes and adaptive estimation of the evolutionary wavelet spectrum. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **62** 271–292. [MR1749539](#)
- NICHOLSON, F. A., SMITH, S. R., ALLOWAY, B. J., CARLTON-SMITH, C. and CHAMBERS, B. J. (2003). An inventory of heavy metals inputs to agricultural soils in England and Wales. *Sci. Total Environ.* **311** 205–219.
- OGDEN, T. R. (1997). On preconditioning the data for the wavelet transform when the sample size is not a power of two. *Comm. Statist. Simulation Comput.* **26** 467–486.
- OH, H.-S., AMMANN, C. M., NAVEAU, P., NYCHKA, D. and OTTO-BLIESNER, B. L. (2003). Multi-resolution time series analysis applied to solar irradiance and climate reconstructions. *J. Atmos. Sol.-Terr. Phys.* **65** 191–201.
- PEREA-GARCÍA, A., ANDRÉS-BORDERÍA, A., DE ANDRÉS, S. M., SANZ, A., DAVIS, A. M., DAVIS, S. J., HUIJSER, P. and PEÑARRUBIA, L. (2015). Modulation of copper deficiency responses by diurnal and circadian rhythms in *arabidopsis thaliana*. *J. Exp. Bot.* **67** 391–403.
- PRICE, T. S., BAGGS, J. E., CURTIS, A. M., FITZGERALD, G. A. and HOGENESCH, J. B. (2008). WAVECLOCK: Wavelet analysis of circadian oscillation. *Bioinformatics* **24** 2794–2795.
- PRIESTLEY, M. B. (1981). *Spectral Analysis and Time Series. Vol. 1. Univariate Series, Probability and Mathematical Statistics*. Academic Press, New York. [MR0628735](#)
- PRIESTLEY, M. B. and SUBBA RAO, T. (1969). A test for non-stationarity of time-series. *J. Roy. Statist. Soc. Ser. B* **31** 140–149. [MR0269062](#)
- RAMSAY, J. O. and SILVERMAN, B. W. (2005). *Functional Data Analysis*, 2nd ed. *Springer Series in Statistics*. Springer, New York. [MR2168993](#)

- SENESIL, G. S., BALDASSARRE, G., SENESI, N. and RADINA, B. (1998). Trace element inputs into soils by anthropogenic activities and implications for human health. *Chemosphere* **39** 343–377.
- SHUMWAY, R. H. (1988). *Applied Statistical Time Series Analysis*. Prentice-Hall, Englewood Cliffs, NJ.
- TAVAKOLI, S. and PANARETOS, V. M. (2016). Detecting and localizing differences in functional time series dynamics: A case study in molecular biophysics. *J. Amer. Statist. Assoc.* **111** 1020–1035. MR3561926
- TORRENCE, C. and COMPO, G. P. (1998). A practical guide to wavelet analysis. *Bull. Am. Meteorol. Soc.* **79** 61–78.
- VAN BELLEGEM, S. and VON SACHS, R. (2008). Locally adaptive estimation of evolutionary wavelet spectra. *Ann. Statist.* **36** 1879–1924. MR2435459
- VIDAKOVIC, B. (2001). Wavelet-based functional data analysis: Theory, applications and ramifications. In *Proceedings of the 3rd Pacific Symposium on Flow Visualization and Image Processing*, Maui, HI.
- VITATERNA, M. H., TAKAHASHI, J. S. and TUREK, F. W. (2001). Overview of circadian rhythms. *Alcohol Res. Health* **25** 85–93.
- VON SACHS, R. and NEUMANN, M. H. (2000). A wavelet-based test for stationarity. *J. Time Series Anal.* **21** 597–613. MR1794489
- ZIELINSKI, T., MOORE, A. M., TROUP, E., HALLIDAY, K. J. and MILLAR, A. J. (2014). Strengths and limitations of period estimation methods for circadian data. *PLoS ONE* **9** 96462.

J. K. HARGREAVES  
 M. I. KNIGHT  
 DEPARTMENT OF MATHEMATICS  
 UNIVERSITY OF YORK  
 YORK YO10 5DD  
 UNITED KINGDOM  
 E-MAIL: [jkh516@york.ac.uk](mailto:jkh516@york.ac.uk)  
[marina.knight@york.ac.uk](mailto:marina.knight@york.ac.uk)

J. W. PITCHFORD  
 DEPARTMENTS OF MATHEMATICS AND BIOLOGY  
 UNIVERSITY OF YORK  
 YORK YO10 5DD  
 UNITED KINGDOM  
 E-MAIL: [jon.pitchford@york.ac.uk](mailto:jon.pitchford@york.ac.uk)

R. J. OAKENFULL  
 S. CHAWLA  
 J. MUNNS  
 S. J. DAVIS  
 DEPARTMENT OF BIOLOGY  
 UNIVERSITY OF YORK  
 YORK YO10 5DD  
 UNITED KINGDOM  
 E-MAIL: [rachael.oakenfull@york.ac.uk](mailto:rachael.oakenfull@york.ac.uk)  
[sangeeta.chawla@york.ac.uk](mailto:sangeeta.chawla@york.ac.uk)  
[jack.munns@york.ac.uk](mailto:jack.munns@york.ac.uk)  
[seth.davis@york.ac.uk](mailto:seth.davis@york.ac.uk)