HHS Public Access

Author manuscript

Ann Appl Stat. Author manuscript; available in PMC 2022 December 02.

Published in final edited form as:

Ann Appl Stat. 2022 September ; 16(3): 1380–1399. doi:10.1214/21-aoas1546.

BAYESIAN SEMIPARAMETRIC LONG MEMORY MODELS FOR DISCRETIZED EVENT DATA

Antik Chakraborty[1,a], **Otso Ovaskainen**[3,4,5,c], **David B. Dunson**[2,b]

[1]Department of Statistics, Purdue University

[2]Department of Statistical Science, Duke University

[3]Department of Biological and Environmental Science, University of Jyväskylä

[4]Organismal and Evolutionary Biology Research Programme, Faculty of Biological and Environmental Sciences, University of Helsinki

[5]Department of Biology, Centre for Biodiversity Dynamics, Norwegian University of Science and Technology, Trondheim

Abstract

We introduce a new class of semiparametric latent variable models for long memory discretized event data. The proposed methodology is motivated by a study of bird vocalizations in the Amazon rain forest; the timings of vocalizations exhibit self-similarity and long range dependence. This rules out Poisson process based models where the rate function itself is not long range dependent. The proposed class of FRActional Probit (FRAP) models is based on thresholding, a latent process. This latent process is modeled by a smooth Gaussian process and a fractional Brownian motion by assuming an additive structure. We develop a Bayesian approach to inference using Markov chain Monte Carlo and show good performance in simulation studies. Applying the methods to the Amazon bird vocalization data, we find substantial evidence for self-similarity and non-Markovian/Poisson dynamics. To accommodate the bird vocalization data in which there are many different species of birds exhibiting their own vocalization dynamics, a hierarchical expansion of FRAP is provided in the Supplementary Material.

Key words and phrases.

Fractional Brownian motion; fractal; latent Gaussian process models; long range dependence; nonparametric Bayes; probit; time series

[a] antik015@purdue.edu . [b] dunson@duke.edu . [c] otso.t.ovaskainen@jyu.fi .

SUPPLEMENTARY MATERIAL

Supplementary materials (DOI: 10.1214/21-AOAS1546SUPPA; .pdf). The supplementary document Chakraborty, Ovaskainen and Dunson (2022a) contains an extension of the FRAP framework to a grade-of-membership model, related priors and computational details for joint inference on multiple species, technical results for proving Theorem 3.2, additional simulation results from Section 4.
Code (DOI: 10.1214/21-AOAS1546SUPPB; .zip). This zip folder contains two `Rscript` files to implement the proposed method. One could download and run the `git_demo_run.R` file as it is.

1. Introduction.

Event data are often obtained in a discretized form in environmental and ecological applications. Instead of recording exact times of event occurrence, one records whether or not at least one event occurred within each interval. Such data can potentially be treated as a discrete time series (Tiao, Phadke and Box (1976), Stern and Coe (1984)), ignoring the underlying continuous time process that generated the events. While this simplification may be more amenable to standard time series analysis, it is often desirable to provide a self-explanatory stochastic model that is capable of capturing the temporal dynamics of the underlying event generating process (Davison and Ramesh (2020)).

In Davison and Ramesh (2020) and Ramesh, Thayakaran and Onof (2013), the authors use a Markov modulated Poisson process (MMPP) (Fischer and Meier-Hellstern (1993)) for the discretized events. Event intensities of an MMPP are directed by the states of an independently evolving continuous time Markov process whose different states correspond to different rates of events. Davison and Ramesh (2020) derived expressions for the likelihood of the observed binary series for an MMPP using Chapman–Kolmogorov equations of a continuous time Markov chain. They proposed a maximum likelihood approach for inference on the model parameters which include the instantaneous transition rate matrix of the continuous time Markov chain and the Poisson rates corresponding to each state of the chain. They also show that the autocorrelation function of the binary time series generated by an MMPP exhibits a geometric decay. Fearnhead and Sherlock (2006) proposed a Gibbs sampling algorithm for Bayesian inference.

The geometric rate of decay in autocorrelations of an MMPP makes it inapplicable to model time series with slower decay in autocorrelations. This is true for time series where the dependence structure is non-Markovian; a special class of time series that has non-Markovian dependence and is a focus in this article is known as long range dependent series. Roughly speaking, a time series is long-range dependent if its autocovariance function decays like a power function. Long-range dependence has been encountered in time series data from a large variety of fields, including hydrology (Hurst (1951)), finance (Lo (1989)), network traffic (Willinger et al. (2003)), and climatology (Franzke et al. (2020)) among others. A natural extension of the MMPP to accommodate long-range dependence is the fractional Poisson process (Laskin (2003)). However, likelihood computation of discretized data, obtained from a fractional Poisson process, is not straightforward.

In seminal work, Mandelbrot and Van Ness (1968) introduced fractional Brownian motion, a generalization of standard Brownian motion, and showed that the increments of this process are stationary and exhibit long range dependence. The general definition of fractional Brownian motion is a stochastic integral with respect to a standard Brownian motion where the order of integration is defined by a parameter $H \in (0, 1)$. Mandelbrot and Van Ness (1968) referred to H as the Hurst parameter after the hydrologist Harold Hurst, who discovered long-range dependence in time series while studying storage capacities of dams on the Nile river. Mandelbrot and Van Ness (1968) also established that the fractional Brownian motion is a self-similar stochastic process with no characteristic time scale (Graves et al. (2017)). Intuitively, self-similar processes retain statistical properties

over different time scales. When the increments of a self-similar process are stationary, these increments exhibit long-range dependence.

For discretized events the intensity of the latent counting process determines the correlation structure of the binary time series. If the binary series is long-range dependent, then an inhomogeneous Poisson process with fixed intensity $\lambda(t)$ is insufficient to explain the observed data, as it implies that increments in disjoint time intervals are independent. Furthermore, Beran et al. (2013), Chapter 2, showed that a doubly stochastic Poisson process with random intensity $\lambda(t)$ is long-range dependent if and only if $\lambda(t)$ is long-range dependent; refer to Samorodnitsky (2006), Pipiras and Taqqu (2017) for reviews on long-range dependence and self-similarity.

In this article we propose a latent semiparametric framework to model long-range dependent discretized event data via a FRActional Probit (FRAP) model. The FRAP model assumes a latent stochastic process responsible for generating the events of interest. Positive values of the process within a time interval imply one or more event occurrences within that interval. By setting the latent process as the fractional Brownian motion parameterized by the Hurst coefficient, we show the FRAP model is able to capture long-range dependence of the discretized events. By varying the Hurst coefficient within (0, 1), the spectrum of the model encompasses antipersistence when $H \in (0, 1/2)$, independence for $H = 1/2$, and long-range dependence when $H \in (1/2, 1)$. Moreover, we also include a nonparametric trend component in our model to account for nonstationarity of event occurrences. The proposed framework accommodates testing of long-range dependence in the data by comparing H_0: $H = 0.5$ vs. H_1: $H > 0.5$. We define a Bayesian approach to inference using a Gaussian process prior for the nonparametric trend. A Markov chain Monte Carlo (MCMC) sampling algorithm is proposed relying on sampling the latent process.

The rest of the article is organized as follows. In Section 2 we introduce the motivating Amazon bird vocalization data, including exploratory analyses revealing possible long-range dependence. Section 3 is dedicated to the development and analysis of the FRAP model. Section 4 contains simulation experiments evaluating the proposed approach, and Section 5 analyzes the Amazon data. In the Supplementary Material (Chakraborty, Ovaskainen and Dunson (2022a)), we extend the FRAP model to allow multiple types of events through a grade-of-membership model and provide details on prior specification and posterior computation.

2. Amazon bird vocalization data.

Bird songs play a major role in mate selection and thus have a pronounced impact on their population dynamics (Slabbekoorn and Smith (2002)). Identifying birds based on their vocalizations is a widely used method for estimating bird population sizes and following population trends over time, and automated acoustic monitoring is increasingly used in both ecological studies and in conservation (Laiolo (2010)). Bird songs are well known to follow a circadian pattern in that they sing most intensely early in the morning and late in the day (Krebs and Kacelnik (1983)).

We are motivated by an Amazon bird vocalization data set containing observations from the years 2010 to 2014. Audio monitoring devices were placed at different locations throughout the Amazon rain forest. Using the methods of Ovaskainen, de Camargo and Somervuo (2018), these recordings were converted to discretized binary time series (de Camargo, Roslin and Ovaskainen (2019)) containing 0–1 indicators of which species vocalized at least once in one minute time intervals for a 180-minute period starting at sunrise. A visual depiction of the binary sequence of vocalizations for the bird species Automolus ochrolaemus is provided in Figure 1. Based on the audio recordings, it is not possible to reliably distinguish different individual birds of the same species or to infer the number of birds vocalizing. We focus on three locations which are similar in habitat and close in latitude and longitude. Our data consist of recordings for 15 relatively common bird species. For each species we have about five to 10 days of recordings during the months of June to September with recordings starting typically around 5:15 AM. On average, a given species vocalized in 25–30 out of the 180 intervals.

Our analysis focuses on three characteristics of the bird vocalization dynamics. First, we are interested in the distribution of duration of bird song activity and inactivity; in particular, our results indicate that the duration cannot be adequately modeled by the exponential distribution. In the context of event data, exponential inter-event times are routinely assumed for mathematical and computational simplicity. However, many naturally occurring events, such as earthquakes (Ogata and Abe (1991)), landscape evolution (Weymer et al. (2018)), and human brain activity (Tagliazucchi et al. (2013)) have been shown not to follow such patterns. We are also interested in identifying time periods when birds are more likely to sing and recovering groups of bird species that have similar singing patterns.

Define the marginal probability of vocalization for a given time interval of length Δt to be the probability of observing at least one vocalization when a time interval of this length is selected at random. In the left panel in Figure 2 we show the marginal probabilities of a vocalization during minute intervals of length $\Delta t = \{1, 2, 4, 9, 15, 30, 60, 90\}$ for 15 different bird species. On the right panel in Figure 2, we show the probabilities of vocalizations conditioned on the event that the bird vocalized in the previous interval of the same length. Quite naturally, the marginal probabilities show an increasing pattern with the length of intervals. In comparison, the conditional probabilities show substantially less variation with Δt; for most species the conditional probabilities vary between (0.4, 0.75). Such scaling of summary statistics is commonly encountered in self-similar stochastic processes (Pipiras and Taqqu (2017)). Additionally, the distance autocorrelations (Zhou (2012)) and the periodogram of the binary series for one day of recording for the species Corythopis torquata is displayed in Figure 3. The distance autocorrelation is a popular alternative to the standard autocorrelation function for investigating nonlinear dependence structures and thus is more suitable for the binary time series data presented here. The slow decay in the distance autocorrelation and the spikes in the spectrum for small frequencies indicate potential long-range dependence in the data.

We will use the notation $X(t)$ for the stochastic process $\{X_t\}_{t \in \mathbb{R}}$. A stochastic process $X(t)$ is said to be self-similar if, for any $c > 0$, we have $X(ct) \overset{d}{=} c^H X(t)$ so that the random

variables $X(t)$ and $X(ct)$ are equivalent in distribution up to scaling factors governed by the parameter H. This parameter $H \in (0, 1)$ is commonly known as the Hurst exponent. A self-similar process with stationary increments has nonsummable autocovariances (Pipiras and Taqqu (2017)) and is known as a long-range dependent (LRD) time series. In such series the degree of long-range dependence is controlled by H. For continuous time series data, many methods have been proposed to estimate H: the ReScaled range (RS) analysis (Hurst (1951), Mandelbrot and Wallis (1969)), detrended fluctuation analysis (Peng et al. (1994)), log periodogram regression (Geweke and Porter-Hudak (1983)), local Whittle approximation (Robinson (1995)) etc. Although these methods typically apply to continuous data, we use these estimators in our exploratory analyses, in particular the estimators due to Geweke and Porter-Hudak (1983) and Robinson (1995).

To estimate H, according to Geweke and Porter-Hudak (1983) and Robinson (1995), we use the `LongMemoryTS` package in R. Table 1 shows the estimates of the Hurst exponent, according to Geweke and Porter-Hudak (1983) ($\widehat{H}_{\mathrm{GPH}}$) and Robinson (1995) ($\widehat{H}_{\mathrm{W}}$), for the 15 bird species from Figure 2. Both $\widehat{H}_{\mathrm{GPH}}$ and $\widehat{H}_{\mathrm{W}}$ have a tuning parameter m which is the number of Fourier frequencies. In Table 1 we report the estimated Hurst coefficients for $\widehat{H}_{\mathrm{GPH}}$ and $\widehat{H}_{\mathrm{W}}$ for $m = n^{1/2}, n^{2/3}, n^{4/5}$. The estimates of H, as seen from $\widehat{H}_{\mathrm{GPH}}$ and $\widehat{H}_{\mathrm{W}}$ in Table 1, suggest long memory behaviour although they often do not satisfy the constraint $0 < H < 1$.

Time series models for discrete valued data with LRD structure are relatively sparse. Classical approaches for count/discrete valued times series, such as the integer autoregressive moving-average (McKenzie (1985, 1986, 1988)) and discrete autoregressive moving-average (Jacobs and Lewis (1978a, 1978b)), cannot account for LRD (Davis et al. (2016), Chapter 21). Cui and Lund (2009) developed a model for stationary Bernoulli sequences with LRD based on renewal sequences. Livsey et al. (2018) provide a recipe for multivariate count time series with Poisson marginals and a flexible autocovariance structure that can adequately handle LRD but fit a misspecified likelihood for inference on relevant parameters. Additionally, it is not entirely straightforward to accommodate covariates in their method. More recently, Jia et al. (2021) developed a method to construct count time series with prescribed marginals through suitable transformations of a latent Gaussian series. However, the joint distribution of counts thus obtained is not easily determined. Estimates of the Hurst exponent, obtained from the quasi-maximum likelihood method from Livsey et al. (2018), are also included in Table 1 under the column $\widehat{H}_{\mathrm{QMLE}}$. The remaining columns in Table 1 refer to model estimates which are discussed later in Section 5.

Our goal is not simply to estimate the Hurst coefficient; we would like to define a realistic generative probability model for these data that takes into account the data collection process and can be used as a useful baseline for future ecological analyses that include spatial dependence, environmental covariates, and other complications. The estimated Hurst coefficients for our proposed fractional probit model are provided in Table 1; see Section 3.1 below. Interestingly, the Hurst coefficients are significantly above 0.5 for all 15 bird species. This suggests long-range dependence, a new finding of ecological interest, which should be considered in future analyses of animal occurrence time series. One can theoretically use a

(doubly stochastic) Poisson process to model these data; however, one should allow flexible rate functions to accommodate long memory behaviour Beran et al. (2013), Chapter 2.

3. Discretized event data.

We begin this section by defining some notation. Suppose event recordings are discretized at time points $\{t_0, t_1, \ldots, t_n\}$ where the time points belong to some index set $\mathcal{T}$. In this article we assume that $t_{i+1} - t_i = \Delta$ for all $i = 0, 1, \ldots, n-1$. Corresponding to each time interval, we have the following binary event indicators:

$$Z(t_{i-1}, t_i) = \begin{cases} 1 & \text{if at least one event occurred in } (t_{i-1}, t_i], \\ 0 & \text{otherwise.} \end{cases} \tag{3.1}$$

We consider R replications of this binary time series $\mathbf{Z} = \{Z^{(1)}, Z^{(2)}, \ldots, Z^{(R)}\}$. In our particular setting the replications correspond to different days of recording at a fixed location and for a fixed bird species.

3.1. Fractional probit model.

Consider, for now, a single replication of the binary series Z. We assume a latent continuous time process $y(t)$, $t \in \mathcal{T}$, is responsible for instigating events of interest. Let $\rho_0(y(s), y(t))$ denote the covariance function of $y(\cdot)$ for $s, t \in \mathcal{T}$. We want to derive a discrete time series from $y(t)$ so that it reflects the autocovariance structure of the observed binary data. Of particular interest are time series that exhibit long-range dependence motivated by the bird vocalization data. A time series $\{X_t, t \in \mathbb{Z}\}$ is said to have long-range dependence if its autocovariance function $\rho_X(k)$ at lag $k \in \mathbb{Z}$ decays polynomially as $k \to \infty$,

$$\rho_X(k) = L(k)k^{2d-1} \quad \text{for } d \in (0, 1/2), \tag{3.2}$$

where $L(\cdot)$ is a slowly varying function at infinity, meaning it is positive on $[c, \infty)$ with $c \geq 0$ and, for any $a > 0$, $\lim_{u \to \infty} L(au)/L(u) = 1$. The parameter d is called the long-range dependence parameter, and the series is said to have *long memory*. A popular alternative characterization of long-range dependent series relies on properties in the frequency domain. If $s_X(\lambda)$ is the spectral density of the times series $\{X_t, t \in \mathbb{Z}\}$, then the series is long-range dependent if

$$s_X(\lambda) = L^*(\lambda)\lambda^{-2d} \quad \text{for } d \in (0, 1/2) \text{ and } 0 < \lambda \leq \pi, \tag{3.3}$$

for some slowly varying function $L^*(\cdot)$ at zero. This definition implies that spectral densities of long-range dependent series have an infinite spike in a neighborhood around zero.

The concept of long memory is intricately related to self-similarity of processes. Broadly speaking, self-similar processes are obtained as normalized limits of partial-sum processes of a long memory series (Pipiras and Taqqu (2017)). While there are several well-studied self-similar processes, one of the most fundamental and perhaps the most popular is the fractional Brownian motion (fBM). A Brownian motion $B(t)$ is a stationary Gaussian process

with covariance function $K_B(s, t) = \tau^2 \min(s, t)$, $\tau > 0$. The fBM generalizes this covariance structure to the form

$$K_H(s,t) = \frac{\tau^2}{2}\left(|t|^{2H} + |s|^{2H} - |t-s|^{2H}\right), \quad H \in (0,1). \tag{3.4}$$

The parameter H is known as the Hurst exponent of the fBM. Henceforth, we shall write $B_H(t)$ to denote an fBM with Hurst exponent H. In (3.4), $\tau^2 = \mathbb{E}\{B_H(1)\}^2$. For $H = 0.5$ the Brownian motion is recovered. The self-similarity of the process stems from the fact that $B_H(ct) \overset{d}{=} c^H B_H(t)$. Setting $\epsilon_i^H = B_H(i) - B_H(i-1)$, $i \in \mathbb{Z}$, we obtain a stationary discrete time series, known as fractional Gaussian noise (fGN), elements of which marginally follow N(0, τ^2). The autocovariance function $\rho_\epsilon(k)$, $k = 0, 1, 2, \ldots$ of $\{\epsilon_n^H\}$ is

$$\rho_\epsilon(k) = \frac{\tau^2}{2}\left(|k+1|^{2H} - 2|k|^{2H} + |k-1|^{2H}\right) \sim \tau^2 H(2H-1)k^{2H-2} \quad \text{as } k \to \infty, \tag{3.5}$$

where, for two sequences a_n and b_n, $a_n \sim b_n$ implies that $a_n/b_n = 1$ as $n \to \infty$. Hence, for $H \in (1/2, 1)$ the series is LRD in the sense of equation (3.2) with LRD parameter $d = H - 1/2$. Our proposed model relies heavily on the simple observation that if we define a series $Z_i^* = \mathbb{I}(\epsilon_i^* > 0)$, where ϵ_i^* is a stationary Gaussian series, then the autocovariance function of this binary series Z_i^* is

$$\rho_{Z^*}(k) = \frac{1}{2\pi}\arcsin \rho_{\epsilon^*}(k); \tag{3.6}$$

see Livsey et al. (2018), Lemma 4.1, for a proof of this property. In particular, if $\epsilon_i^* = \epsilon_i^H$, then the binary series inherits the LRD property. To see this, suppose $H \in (1/2, 1)$, then for large lags k, $\rho Z^*(k) \approx \rho_\epsilon(k)$ since $\sin x \approx x$ for small x, that is, the series Z_i^* is also long-range dependent with Hurst coefficient H. In the context of discretized event data as described in (3.1), we then have the following latent formulation:

$$Z(t_{i-1}, t_i) = \begin{cases} 1 \text{ if } \epsilon_i^H = B_H(t_i) - B_H(t_{i-1}) > 0, \\ 0 \text{ otherwise,} \end{cases} \tag{3.7}$$

for $i = 0, 1, \ldots$; see also Livsey et al. (2018), equation (4.4), for an equivalent formulation for any latent Gaussian series. The above formulation accounts for long memory in the observed binary series, with the autocorrelation decay mimicking that of an fGN. Moreover, as a consequence of the scaling property of an fBM, a scale-free property of conditional probabilities consistent with Figure 2 is established in the following lemma.

LEMMA 3.1. *Let $B_H(t)$ be an fBM with Hurst coefficient H with $\tau^2 = 1$. Suppose we observe $B_H(t)$ at $i \in \mathbb{N}$ and let $X_i \equiv B_H(i)$, $i \geq 1$, $X_0 \equiv B_H(0)$. Define the binary series of indicators at time scale m, $Z_i^{(m)} = \mathbb{1}\{X_{i2^m} - X_{(i-1)2^m} > 0\}$, $i \geq 1$ so that, for $m = 0$, the*

series $Z_1^{(0)}, Z_2^{(0)}, \ldots$ *is as in* (3.7). *Then, for any* $m = 0, 1, \ldots$, *the conditional probability* $\mathbb{P}\left(Z_{i+1}^{(m)} = 1 \mid Z_i^{(m)} = 1\right)$ *is independent of the time scale* m. *In particular,*

$$P\left(Z_{i+1}^{(m)} = 1 \mid Z_i^{(m)} = 1\right) = \frac{1}{2} + \frac{1}{\pi}\arcsin\left(2^{2H-1} - 1\right). \tag{3.8}$$

Proof. See Appendix A.1. □

Two remarks are in order. First, for the special case $H = 0.5$, the conditional probability in equation (3.8) becomes 1/2 so that, when the series of indicators are generated from an underlying white noise series, the conditional probability of $Z_{i+1} = 1|Z_i = 1$ and the marginal probability of $Z_i = 1$ are equal. Second, since the function arcsin(·) is increasing, the conditional probability of $Z_{i+1} = 1|Z_i = 1$ increases with H, covering the cases of antipersistence $H < 0.5$, independence $H = 0.5$, and LRD for $H > 0.5$. Figure 4 depicts the relationship between the Hurst coefficient H and the conditional probabilities.

Additionally, the spectral density of the series Z_n can be shown to have a pole at zero frequency when $H > 1/2$, a distinctive feature of LRD series. Let $s_Z(\lambda)$ and $s_\epsilon(\lambda)$ denote the spectral density of the series Z_n and ϵ_n, respectively, for $-\pi \le \lambda \le \pi$. Then we have, for $H > 1/2$,

$$s_Z(\lambda) = \sum_{k=-\infty}^{\infty} \rho_Z(k)\exp(ik\lambda) = \sum_{k=-\infty}^{\infty} \frac{1}{2\pi}\arcsin\rho_\epsilon(k)\exp(ik\lambda) \ge \sum_{k=-\infty}^{\infty} \frac{1}{2\pi}\rho_\epsilon(k)\exp(ik\lambda) = \frac{1}{2\pi}s_\epsilon(\lambda),$$

where we have used the Jordan inequality $\arcsin x - x \ge 0$ for $0 < x < 1$ (Mitrinović and Vasic (1970)). Combining this with the fact that $s_\epsilon(\lambda) \sim (\tau^2/C)\lambda^{1-2H}$, $C = C(H) > 0$ in a neighborhood of 0, we see $s_Z(\lambda)$ also has a pole at $\lambda = 0$ for $H > 1/2$ and hence is LRD, according to definition (3.3).

When considering the Amazon bird vocalization data and other real data applications, a clear limitation of model (3.7) is the restriction of the marginal probabilities being fixed at 0.5. To be realistic, we need to allow the marginal probabilities to be arbitrary and varying smoothly according to the time of the day. Moreover, Mikosch and Stărică (2004) and Chen, Härdle and Pigorsch (2010), among many others, noted that long memory behavior can often be an artifact of nonstationarities.

With this motivation we introduce a nonstationary component in the FRAP model by assuming that the latent process driving the events, say $y(t)$, admits an additive decomposition of the form $y(t) = f(t) + B_H(t)$ while letting

$$Z(t_{i-1}, t_i) = \begin{cases} 1 \text{ if } y(t_i) - y(t_{i-1}) = f(t_i) - f(t_{i-1}) + \epsilon_i^H > 0, \\ 0 \text{ otherwise,} \end{cases} \tag{3.9}$$

where we assume $f(\cdot)$ is continuously differentiable (see Section 4 for examples). The marginal probability of observing an event in interval $(t_{i-1}, t_i]$ is then

$P[Z(t_{i-1}, t_i) = 1] = P\left[f(t_i) - f(t_{i-1}) + \epsilon_i^H > 0\right] = \Phi\{f(t_i) - f(t_{i-1})\}$, where $\Phi(\cdot)$ is the cumulative density function of a standard Gaussian random variable. Hence, the variation in $f(\cdot)$ during $(t_{i-1}, t_i]$ determines the probability of observing an event during this time; a positive change increases the marginal probability, whereas a negative change decreases it. If $f(t_i) - f(t_{i-1}) = 0$, then the marginal probability is $P\left(\epsilon_i^H > 0\right) = 1/2$. To simplify notation, we write $Z_i = Z(t_{i-1}, t_i)$. The vector $\epsilon^H = \left(\epsilon_1^H, \ldots, \epsilon_n^H\right)$ follows an n-dimensional Gaussian distribution with mean 0 and covariance matrix Σ_H whose (i, j)th element is $\Sigma_H(i, j) = \tau^2\rho_\epsilon(|i - j|)$, defined in equation (3.5). The marginal probability of an event occurrence in the interval $(t_{i-1}, t_i]$ then becomes $P[Z(t_{i-1}, t_i) = 1] = \Phi[\{f(t_i) - f(t_{i-1})\}/\tau]$. In Figure 5 we show the variations in marginal probabilities when the nonstationary component $f(t)$ in model (3.9) is set to $f(t) = \sin(4\pi t)/90$ with $\tau = 1$.

Akin to probit models for longitudinal binary data with covariate information (Chib and Greenberg (1998)), we are interested in modeling the likelihood of the observed events $Z = (Z_1, \ldots, Z_n) \in \{0, 1\}^n$. However, in our context we have time series data with smooth trend $f(t)$ and temporal dependence captured through ϵ^H. Letting $\mathbf{f} = \{f(t_0), \ldots, f(t_n)\}$ and putting the pieces together, we get the following probit-type model:

$$P(Z \in E \mid \mathbf{f}, H) = P(W \in E_W \mid \mathbf{f}, H), \quad W \sim \mathrm{N}\left(A\mathbf{f}, \tau^2\Sigma_H\right), E \subset \{0, 1\}^n, \tag{3.10}$$

where E_W is the intersection of half-planes $E_W = \cap_{i:\, Z_i = 1}(W_i > 0) \cap_{i:\, Z_i = 0}(W_i \leq 0)$ and the matrix $A \in \mathbb{R}^{n \times n}$ is such that $A_{ii} = 1$, $A_{i,i-1} = -1$ and $A_{ij} = 0$ for $j \neq i, i - 1$. For identifiability we impose the restriction that $f(0) = 0$. Then, under model $f(\cdot)/\tau$ is identifiable. To accommodate this restriction, we let $A_{11} = 1$, $A_{1,j} = 0$, $j = 2, \ldots, n$; the other rows of A remain unchanged.

Model (3.10) is quite flexible in incorporating a smooth trend $f(t)$ and autocorrelated errors. In the special case in which $H = 0.5$, the error term becomes uncorrelated so that $f(t)$ is assumed to characterize the pattern over time in the data. In contrast, when $H > 0.5$, we obtain long range dependence. The model provides a useful basis for testing of long-range dependence via comparing H_0: $H = 0.5$ to H_1: $H > 0.5$ in the presence of potential nonstationarity.

3.2. Priors and posterior computation.

Without loss of generality, we assume that the time points $\{t_0, \ldots, t_n\} \in \mathcal{T} = [0, T]$. Let $\Theta = \{(f, \beta, \tau): f \in \mathcal{F}, \beta \in \mathbb{R}, \tau \in \mathbb{R}^+\}$ be the parameter space in model (3.10), where we let $\mathcal{F}$ be the space of continuously differentiable functions on $\mathcal{T}$ and $\beta = \log\{H/1 - H\}$. Let Π_β denote the prior on β and Π_τ denote the prior on τ. We choose $\Pi_\beta \equiv \mathrm{N}(0, 1)$ and $\Pi_\tau \equiv$ Inverse-Gamma(a_τ, b_τ) for positive constants a_τ, b_τ. For the nonparametric component we let $f \sim \Pi_f$, where Π_f is an appropriate prior for an unknown smooth function. In particular, we choose a zero mean Gaussian process (GP) with a squared exponential covariance kernel (Rasmussen and Williams (2006)) scaled by the precision parameter τ^2 of the latent process, defined as

$$C(s,t) = \tau^2\sigma^2 \exp\left\{-\frac{(s-t)^2}{2\phi^2}\right\}, \quad \sigma, \phi > 0, \tag{3.11}$$

for $s, t \in \mathcal{T}$. For numerical stability we follow the standard practice of adding a small positive quantity ν to the diagonal elements of the GP covariance matrix so that $C(s,t) = \tau^2\sigma^2\exp - \left\{(s-t)^2/2\phi^2\right\} + \upsilon\mathbb{1}(s=t)$. Consequently, the induced prior on $\mathbf{g} = A\mathbf{f}$ is again a multivariate Gaussian distribution with covariance matrix $C_{\mathbf{g}} = \tau^2 ACA'$, where C is an $n \times n$ matrix with $C_{ij} = C(t_i, t_j)$. To learn the hyperparameters (σ, ϕ) from the data, we transform them to the logarithmic scale and augment the parameter space Θ to $\Theta_* = \Theta \times \eta$, where $\eta = \{(\log\sigma, \log\phi): \sigma, \phi > 0\}$. We place independent standard Gaussian priors on each component of η. Thus, $\Pi\eta \equiv N(0, 1) \times N(0, 1)$. The prior specification is completed by setting $\Pi = \Pi_f \times \Pi_\beta \times \Pi_\tau \times \Pi_\eta$.

For the Amazon bird vocalization data, we have replications $\{Z^{(1)}, \ldots, Z^{(R)}\}$ of Z over different days which have minimal empirical correlations. We assume these replicates are conditionally independent involving the same $f(t)$ but with different realizations of the latent residual term leading to different realizations $W^{(r)}$, for $r = 1, \ldots, R$, of W in equation (3.10). Including also the priors, this leads to the following hierarchy:

$$\begin{aligned} &P\left(Z^{(r)} \in E_r \mid \mathbf{f}, \beta, \tau\right) = P\left(W^{(r)} \in E_{W^{(r)}} \mid \mathbf{f}, \rho\right), \quad \beta = \log\left\{H/(1-H)\right\}, \\ &W^{(r)} \mid \rho, \mathbf{f}, \eta \sim N\left(A\mathbf{f}, \tau^2\Sigma_H\right), \\ &\mathbf{f} \mid \tau^2, \eta \sim \Pi_f, \\ &\beta \sim \Pi_\beta, \tau \sim \Pi_\tau, \eta \sim \Pi_\eta, \end{aligned} \tag{3.12}$$

for any $E_r \subset \{0, 1\}^n$ and $E_{W^{(r)}}$, as defined after equation (3.10).

Posterior computation under the hierarchical FRAP model (3.12) is potentially challenging. We initially considered an integrated nested Laplace approximation (INLA) which was developed for approximate Bayesian inference in latent Gaussian models by Rue, Martino and Chopin (2009). However, the non-Markovian structure of the FRAP model renders the INLA paradigm nonapplicable (Rue and Held (2005)). In a recent article, Sørbye and Rue (2018) applied the INLA framework to a fGN model where the authors approximate the fGN by a mixture of first-order autoregressive processes. This approximation technique works quite well when the observed time series is quite long $n \sim 500$ and the number of replications available is also very high $R \sim 1000$. For the Amazon bird vocalization data, both the length and the replications are quite small compared to these numbers.

Algorithm 1: MCMC algorithm to draw samples from the posterior under model (3.12)

Data: $\mathbf{Z} = \{Z^{(1)}, \ldots, Z^{(R)}\}$, L = number of MCMC samples
Result: L posterior samples from $\Pi(\Theta_* \mid \mathbf{Z}) : \{\hat{\Theta}_*^{(l)}\}_{l=1}^{L}$
Initialize $\beta = 0$, $\mathbf{g} = A\mathbf{f} = 0$, $\log\sigma = 0$ and $\log\phi = 0$;
for $l = 1 : L$ **do**

- **Update** $W_r \mid - \overset{\text{ind}}{\sim} \mathrm{N}(\mathbf{g}, \tau^2\Sigma_H)\mathbb{1}_{B_{W_r}}, r = 1, \ldots R$
- **Update** $\mathbf{g} \mid - \sim \mathrm{N}\{\frac{R}{\tau^2}\Phi^{-1}\Sigma_H^{-1}\overline{W}, \Phi^{-1}\}$, where $\overline{W} = \frac{1}{R}\sum_{r=1}^{R} W_r$ and $\Phi = \frac{R}{\tau^2}\Sigma_H^{-1} + \frac{1}{\tau^2}C_{\mathbf{g}}^{-1}$.
- **Update** $\beta \mid -$ using a Metropolis random walk with proposal density $\mathrm{N}(\beta_{l-1}, s_1^2)$.
- **Update** $\tau \mid - \sim$ Inverse-Gamma$(\frac{nR}{2} + a_\tau, S + b_\tau)$, where $S = \frac{1}{2}[\mathrm{trace}\{(W - G)'\Sigma_H^{-1}(W - G)\} + \mathbf{g}'C_g^{-1}\mathbf{g}]$ and G is a $n \times R$ matrix with all columns equal to $\mathbf{g}$.
- **Update** $\eta \mid -$ jointly via a Metropolis random walk with proposal density $\mathrm{N}(\eta_{l-1}^{(1)}, s_2^2) \times \mathrm{N}(\eta_{l-1}^{(2)}, s_2^2)$, where $\eta^{(1)} = \log\sigma$ and $\eta^{(2)} = \log\phi$.

end

We instead focus on Markov chain Monte Carlo (MCMC), developing a practical algorithm that exploits the structure of the model, as detailed in Algorithm 1. We use $\theta \mid -$ to denote the full conditional distribution of a parameter θ, given other parameters and the data in Algorithm 1. The Metropolis random walk steps to update the Hurst exponent and the Gaussian process kernel hyperparameters are implemented following the adaptive Metropolis algorithm (Roberts and Rosenthal (2001)). Adaptive Metropolis modifies the classical version of the algorithm by varying the covariance of the noise in the random walk targeting the optimal acceptance rate (Roberts and Rosenthal (2001)). Suppose s_1 and s_2 are the noise variance of the random walk updates of β and η, respectively. We start with $s_1 = 0.1$ and $s_2 = 0.2$ and update them at MCMC iteration l by increasing or decreasing by a factor of $\exp(l^{-0.5})$ whenever l is divisible by 50. Adaptation targets an acceptance probability of ~ 0.3. Values of $f(\cdot)$ at a set of test points can also be evaluated by accommodating a further step in Algorithm 1 following Rasmussen and Williams (2006), equations 2.22–2.24.

The main computational bottleneck of Algorithm 1 involves simulating the truncated Gaussian random variables for updating the latent variables W_r. This is done using R package `tmvtnorm`. Unfortunately, we found the popular circulant embedding algorithm (Pipiras and Taqqu (2017), Chapter 2.11) to simulate Gaussian long-range dependent sequences to be quite slow when these constraints are imposed. To accelerate computation, the R copies of the latent variables are generated in parallel. The R code to implement the FRAP model, given R copies of discretized events, is available at https://github.com/antik015/Fractional-Probit-Model; the code is also available in Chakraborty, Ovaskainen and Dunson (2022b).

3.3. Asymptotics.

Here, we consider infill asymptotics, so we assume we can make measurements at finer time points $\{t_0, \ldots, t_n\}$ as $n \to \infty$ within the interval $[0, T]$. We assume the noise variance $\tau = 1$. Also, we set the number of replications $R = 1$ since the proof does not depend on a specific value of R. Let the true trend function be $f_0 \in \mathcal{F}$ and the true Hurst coefficient be H_0, satisfying $0 < a < H_0 < b < 1$ for some $a, b \in (0, 1)$. Define $\theta_0 = (f_0, H_0)$ and P_0 to be the true data generating probability measure, and consider any weak neighborhood U of θ_0. By showing that the joint prior $\Pi \equiv \Pi_\beta \times \Pi_f$ has positive Kullback–Leibler support we have the following consistency result.

THEOREM 3.2. *Suppose* $f_0 \in \mathcal{F}$ *and* $0 < a < H_0 < b < 1$ *for some* $a, b \in (0, 1)$. *Write* $\theta_0 = (f_0, H_0)$, *and consider any weak neighborhood* U *of* θ_0. *Then, the posterior probability of the set* U^c *given the series of indicators* $\Pi(U^c \mid Z_1, \ldots, Z_n) \to P_0$*-probability as* $n \to \infty$.

A proof of Theorem 3.2 is provided in the Appendix.

4. Simulation experiments.

We report the results of a detailed simulation study for different choices of the latent trend function $f(\cdot)$ in equation (3.9) while varying the number of replications R. We assume discretized observations are available for a period of $n = 90$ time units, and the number of replications R considered is {10, 25, 50}. The following choices of the trend function are considered:

1. $f_1(t) = \sin\frac{4\pi t}{90}$;
2. $f_2(t) = 5[1 + \exp\{-2.5(t-45)/15\}]^{-1}$;
3. $f_3(t) = -2\{(t-45)/45\}^2 + 2$;
4. $f_4(t) = -1.2\left\{(t-45)/45\right\} + 0.5\cos\left(\frac{3\pi t}{90}\right) - 1.7$;
5. $f_5(t) = 0.1 f_1(t)\log\{f_2(t)\}$.

We note here that $f_2(\cdot)$ slightly violates the assumption that the nonstationary component in model (3.9) at $t = 0$ is 0. We define the squared empirical ℓ_2 norm of a function $g(\cdot)$, evaluated on the points $\{t_1, \ldots, t_n\}$, as $\| g \|_{2,n} = n^{-1}\sum_{i=1}^{n}\{g(t_i)\}^2$. Given an estimator $\tilde{f}(\cdot)$ of $\underline{f}(\cdot) := f(\cdot)/\tau$ in model (3.12), we evaluate the performance of Algorithm 1 by computing the *re*lative mean square error (ReMSE), defined as $\text{ReMSE} = \|e\|_{2,n}/\|\underline{f}\|_{2,n}$, where $e(\cdot) = \underline{f}(\cdot) - \tilde{f}(\cdot)$. The latent trends $f(\cdot)$ are chosen from the aforementioned list, and $\tilde{f}(\cdot)$ is set to be the pointwise posterior mean of $f(\cdot)/\tau$ at $\{t_1, \ldots, t_{90}\}$, obtained under the hierarchy (3.12). We considered three choices for the Hurst exponent, namely, {0.5, 0.75, 0.9}, ranging from independent increments for $H = 0.5$ to highly correlated increments for $H = 0.9$. We generated the binary data by first evaluating $y(t) = f(t) + B_H(t)$ at $\{t_0, t_1, \ldots, t_n\}$; to simulate the noise vector, we sampled $\epsilon^H \sim \text{N}(0, \tau^2 \Sigma_H)$ with $\tau^2 = 0.05^2, 0.1^2, 0.15^2$. Representing each positive increment of $y(\cdot)$ by 1, the discretized series Z is obtained, and the sampling is repeated R times to complete the data generation process. For each combination of $f(\cdot)$, H, τ, and R, we performed 30 independent evaluations of the proposed framework, and in Table 2 we report the average ReMSE and the average estimated Hurst exponent for $\tau = 0.1$ with the value of ν fixed at 0.001; results for $\tau = 0.05, 0.15$ are provided in the Supplementary Material (Chakraborty, Ovaskainen and Dunson (2022a)).

Estimates of the Hurst exponent are quite accurate across all the combinations of R, H, and $f(\cdot)$. This is important in the context of the Amazon bird vocalization data for which we have, on average, 10 days of data. Naturally, the ReMSE in Table 2 is inversely proportional to the number of replications R, decreasing by a factor of two when the number of replications is doubled. Interestingly, the degree of LRD also controls the ReMSE. For all

the choices of $f(\cdot)$, the average ReMSE increases with H. Large H implies strong dependence in the data which makes the problem of recovering $f(\cdot)$ harder. This was investigated formally in Hall and Hart (1990) who observed that the rates of recovering $f(\cdot)$ decrease with H. Set $p(t_1, t_2) = \Phi[\{f(t_2) - f(t_1)\}/\tau]$ as the true marginal probability under model (3.9) with trend function $f(\cdot)$ and let $\hat{p}(t_1, t_2) = \Phi\left[\left\{\hat{f}(t_2) - \hat{f}(t_1)\right\}/\hat{\tau}\right]$ denote samples from the posterior distribution of f and τ obtained fitting Algorithm 1. The black line in Figure 6 is the posterior mean of the marginal probabilities $\hat{p}(t_1, t_2)$, and the red line plots $p(t_1, t_2)$ for the case $H = 0.75$ and $R = 50$ and five choices of $f(\cdot)$ in consideration here. We also show the pointwise 95% credible bands of $\hat{p}(t_1, t_2)$. The best result is obtained for $f_1(t)$. The credible bands mostly provide accurate uncertainty quantification for all the cases. However, when the number of replications R is smaller the problem of accurately estimating the marginal probabilities becomes much harder, especially for high values of H. Posterior samples of the Hurst exponent for one case are also included in the figure.

To further investigate the behavior of the posterior distribution of the Hurst exponent, we carried out an independent simulation experiment focusing on the coverage probability of the credible intervals. We fix the number of replicates at $R = 5$ and vary the Hurst exponent together with the latent trends as above. For each such combination, we generated 100 data sets and applied model (3.12). Our findings for 95% credible intervals are summarized in Table 3. The coverage probabilities (CP) for all the cases considered are close to the nominal level. The average lengths (l) of the intervals vary substantially for different choices of H along with the standard deviation. For example, the average length of the intervals are maximum for the case $H = 0.5$ with very little variation, but, when $H = 0.9$, the intervals become shorter on average although their variability increases by almost a factor of 3.

5. Application to Amazon bird vocalization data.

5.1 Analysis and results.

We applied the FRAP model to the 15 bird species mentioned in Section 2. For each of these species, we have 180 minutes of recordings available for multiple days. The estimated Hurst exponents for these 15 species are reported in Table 1 with the posterior mean, lower, and upper end of the 95% credible intervals under columns $\widehat{H}_{\text{FRAP}}$, $\widehat{H}_{\text{LR}}$, and $\widehat{H}_{\text{UR}}$, respectively. All the species show high long-range dependence in their temporal vocalization patterns. The posterior mean estimate of the Hurst exponent for the birds range from a minimum of 0.83 up to 0.94. The variation in the Hurst exponent across species is very small with an overall mean of 0.88 and standard deviation 0.04. The high value of the Hurst exponents is consistent with the data in the sense that birds either vocalize or remain silent over long periods of time. We note that this is a combination of two factors which are occurrence and vocalization activity. First, due to their movement activity, a bird individual may be in the vicinity of the recorder for some time and then move to another location. Second, conditional on the bird being present, it may sustain its vocalization activity over some time and remain silent over another time.

Figure 7 shows posterior means and 95% pointwise intervals for the species-specific marginal probabilities of vocalizations occurring in each of the 180 time intervals between

5.15–8.15 a.m. for all 15 species listed in Table 1. Due to data sparsity and the high Hurst exponent, the raw posterior samples exhibited spiky patterns over time, and, hence, we (mildly) smoothed the samples prior to calculating the posterior summaries in Figure 7. While these trends should not be overinterpreted, we do see some general patterns appearing. For example, for Cercomarca cinerascens, Frederickena viridis, Grallaria varia, Micrastur mirandollei, Myrmeciza ferruginea, Percnostola ruffifrons, Pipra erythrocephala, Pithys albifrons, and Ramphastos vitellinus we see an increase in vocalization activity after 7 a.m., whereas Automolus ochrolaemus, Corythopis torquata, Hylexetastes perrotii, and Ibycter americanus more or less maintain a uniform activity level during this time. Micrastur gilvicollis and Hylophilus muscicapinus show more activity during the early hours of the day. Since groups of birds show similar vocalization patterns, in Chakraborty, Ovaskainen and Dunson (2022a), we extend the FRAP framework to a hierarchical setting that shares information across different species.

5.2 FRAP vs. MMPP.

We compare the fit of the proposed FRAP model with the MMPP model (Davison and Ramesh (2020)) for discretized event data via summary statistics derived from the posterior distribution and maximum likelihood estimates, respectively. The particular summary statistics in which we are interested are the conditional probabilities in Figure 2. In the context of the FRAP model, the distribution of the binary indicators Z is completely characterized by the latent variables W. The posterior predictive distribution of W_{R+1}, given the observed binary indicators $Z_1, \ldots, Z_R$, is $p(W_{R+1} \mid Z_1, \ldots, Z_R) = \int p(W_{R+1} \mid \theta_*) p(\theta_* \mid Z_1, \ldots, Z_R)$, where $\theta_* = (\mathbf{f}, \beta, \tau, \sigma, \phi)^{\mathrm{T}}$ and $p(W_{R+1} \mid \theta_*) \sim \mathrm{N}(A\mathbf{f}, \tau^2 \Sigma_H)$, $H = \log\{\beta/(1-\beta)\}$. To sample the latent variable W_{R+1}, we use the MCMC samples of θ_* obtained from Algorithm 1, that is, given $\theta_*^{(l)}$, the lth MCMC sample from $p(\theta_* \mid Z_1, \ldots, Z_{\mathrm{R}})$, we draw $W_{R+1}^{(l)} \sim \mathrm{N}\left(A\mathbf{f}^{(\mathbf{l})}, \tau^{2(l)} \Sigma_H^{(l)}\right)$. Then, equation (3.9) is used to obtain the corresponding binary series $Z_{R+1}^{(l)}$.

The MMPP assumes event occurrence is governed by specific states of an unobserved continuous time Markov chain, hereafter referred to as CTMC, $X(t)$ with finite state space $\{1, 2, \ldots, K\}$ and instantaneous transition probability matrix $\mathrm{G} \in \mathbb{R}^{K \times K}$. Given the chain is in state $k \in \{1, \ldots, K\}$ at time t, events occur following a Poisson process with rate λ_k. The event generating process is then parameterized by the G and $\lambda = \{\lambda_1, \ldots, \lambda_k\}$. The likelihood of a discretized series of events under the MMPP model has been derived in Davison and Ramesh (2020). Let $\hat{G}$ and $\hat{L}$ denote the maximum likelihood estimates of G and L, respectively, using R replicates of binary event indicators $Z_1, \ldots, Z_R$. For the Amazon bird vocalization data we generate a series of binary event indicators Z_{R+1} using the plug-in estimates $\hat{G}$ and $\hat{L}$ with $k = 2$.

Having generated event indicators Z_{R+1} from the two models for each of the 15 species in Table 1, we compute the conditional probability of occurrence of a vocalization, given a vocalization in the previous interval for time scales $\Delta t = \{1, 2, 4, 9, 15, 30, 60, 90\}$; for the FRAP model we compute the conditional probabilities for each MCMC sample $Z_{R+1}^{(l)}$

and consider the average. In the left panel of Figure 8, we plot these probabilities using the estimates obtained from the MMPP model, and in the right panel we plot the average conditional probability for different time scales across MCMC samples. The proposed FRAP model captures the scaling of the conditional probabilities seen in the observed data (Figure 2) while the MMPP does not. We also fitted the MMPP with $K = 3$ states, but the results were very similar.

5.3. Model diagnostics.

We also carried out typical model diagnostics for count time series data, discussed in Czado, Gneiting and Held (2009) and Kolassa (2016). Specifically, we use marginal calibration plots to assess model fit. We first draw samples from the predictive distribution of $Z_{180} | Z_1, \ldots, Z_{179}$ for a particular species of bird. We then compute $P(Z_{180} = 1 \mid Z_1, \ldots, Z_{179})$ using the Monte Carlo average. This predictive probability is then matched with the observed probability $P(Z_{180} = 1)$ which is computed as $R^{-1} \sum_{r=1}^{R} Z_{180}^{(r)}$. In Figure 9 we plot the differences in the predicted and observed probabilities for the 15 different species. For some birds the difference is very small, whereas for other birds this difference goes up to 0.25, especially when the number of replicates available is small. Overall, the model performs adequately; prediction of vocalizations can potentially be improved by including covariates, such as weather and habitat conditions at the sampling site.

6. Discussion.

In this article we proposed a novel class of models for characterizing long-range dependence in discretized event data, along with a Bayesian approach to inference under these models. We are particularly motivated by bird vocalization studies and, indeed, are involved in ongoing collaborations collecting many such datasets across the globe in order to obtain new insights into biodiversity, interactions among species, and the role of biotic and abiotic factors. The proposed class of FRAP models provides an important starting point for building realistic models for these emerging datasets as well as related datasets from precipitation and storm event modeling. Immediate next directions are to add complexity to the models in order to more realistically characterize structure in the data, ranging from spatial dependence to covariate effects. Such extensions are conceptually quite straightforward.

There are several other important directions that are potentially less trivial. The first is to broaden the class of models from a latent fractional Brownian motion to a broader class of stochastic processes with long-range dependence. This may include long-range modifications to usual Gaussian process covariance kernels (e.g., Matern) as well as non-Gaussian cases; for example, Levy processes, alpha-stable processes, etc. The second critical direction is developing much faster computational algorithms. There is an immense literature on algorithms for accelerating computation in Gaussian process models but, to our knowledge, very little consideration of the case in which there is long-range dependence. In our motivating applications we are faced with immense datasets containing automated recordings over time at many different locations around the world. To scale up to such

datasets, we plan to consider divide-and-conquer algorithms and variational approximations, among other directions.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments.

The authors acknowledge support from the United States Office of Naval Research (ONR) and the European Research Council (ERC). We also thank the Editor and two anonymous referees for their constructive suggestions.

APPENDIX SECTION

A.1. Proof of Lemma 3.1.

Since the fBM is a Gaussian process, from Corollary 2.6.3 of Pipiras and Taqqu (2017), we get $\mathbb{E}\{B_H(i)\} = 0$ and $\mathbb{E}\{B_H(i)\}^2 = i^{2H}$ for any $i \in \mathbb{N}$. Hence, $B_H(i) \sim \mathrm{N}(0, i^{2H})$. By stationarity of the incremental process of fBM, it is enough to show (3.8) holds for $i = 1$. Define $Y_1^{(m)} = X_{2^m}$ and $Y_2^{(m)} = X_{2^m+1} - X_{2^m}$. Then, $Y_1^{(m)} \sim \mathrm{N}\left(0, 2^{2Hm}\right)$. Also, $\mathbb{E}\left(Y_2^{(m)}\right)^2 = \mathbb{E}(X_{2^m+1})^2 + \mathbb{E}(X_{2^m})^2 - 2\,\mathrm{Cov}(X_{2^m+1}, X_{2^m})$. From (3.4) we get $\mathrm{Cov}(X_{2^m+1}, X_{2^m}) = 2^{2H(m+1)}$. Thus, we have $Y_2^{(m)} \sim \mathrm{N}\left(0, 2^{2Hm}\right)$. Finally, $\mathrm{Cov}(Y_1, Y_2) = \mathrm{Cov}(X_{2^m}, X_{2^m+1}) - \mathbb{E}(X_{2^m})^2$, which, after applying (3.4) again, we obtain $\mathrm{Cov}(Y_1, Y_2) = 2^{2Hm}(2^{2H-1} - 1)$. Setting $\lambda^2 = 2^{2Hm}$,

$$
\begin{aligned}
&P\left(Z_2^{(m)} = 1 \mid Z_1^{(m)} = 1\right) = \frac{P(Y_1 > 0, Y_2 > 0)}{P(Y_1 > 0)} \\
&= 2P(Y_2/\lambda > 0, Y_2/\lambda > 0) \\
&= 2\left[\frac{1}{4} + \frac{1}{2\pi}\arcsin\left\{\frac{1}{\lambda^2}\mathrm{Cov}(Y_1, Y_2)\right\}\right] \\
&= \frac{1}{2} + \frac{1}{\pi}\arcsin\left(2^{2H-1} - 1\right).
\end{aligned}
$$

A.2. Mixing of MCMC chain in Algorithm 1.

We briefly comment on the mixing of the MCMC chain obtained via Algorithm 1. With L MCMC samples we calculate the effective sample sizes (ESS) for the parameters $f(\cdot)//\tau$ H as

$$
\mathrm{ESS} = \frac{L}{1 + 2\sum_{j=1}^{J}\rho(k)}, \tag{A.1}
$$

where $\rho(j)$ is the autocorrelation at lag j. We set $J = 30$ as the maximum lag and $L = 10{,}000$. For the 180 parameters $f(t)/\tau$, where $t = 1, \ldots, 180$, the average effective sample size for the 15 species were 2012.21 and that for the Hurst coefficient H averaged over all the species is 1941.44.

A.3. Proof of Theorem 3.2.

For any set $V \in \Theta$, the posterior probability $\Pi(V \mid Z_1, \ldots, Z_n) = \int \Pi(V \mid W, Z_1, \ldots, Z_n)\, \Pi(W \mid Z_1, \ldots, Z_n)\, dW$. Now, fix any weak neighborhood U of θ_0. Weak consistency conditional on the latent variables is proved in Section S6 of Chakraborty, Ovaskainen and Dunson (2022a). Thus, the random variable $\Pi(U^c \mid W, Z_1, \ldots, Z_n)$ converges to 0 in P_0-probability. We now extend the proof for the marginal probability $\Pi(U^c \mid W, Z_1, \ldots, Z_n)$. Fix any $\delta > 0$. Then we have

$$
\begin{aligned}
& E_{P_0} \Pi\left(U^C \mid Z_1, \ldots, Z_n\right) \\
& = E_{P_0} \int \Pi\left(U^c \mid W, Z_1, \ldots, Z_n\right) \Pi(W \mid Z_1, \ldots, Z_n) dW \\
& = E_{P_0} \int_{\Pi\left(U^c \mid W, Z_1, \ldots, Z_n\right) \leq \delta} \Pi\left(U^c \mid W, Z_1, \ldots, Z_n\right) \Pi(W \mid Z_1, \ldots, Z_n) dW \\
& + E_{P_0} \int_{\Pi\left(U^c \mid W, Z_1, \ldots, Z_n\right) > \delta} \Pi\left(U^c \mid W, Z_1, \ldots, Z_n\right) \Pi(W \mid Z_1, \ldots, Z_n) dW \\
& \leq \delta + P_0\left\{ \Pi\left(U^C \mid W, Z_1, \ldots, Z_n\right) > \delta \right\},
\end{aligned}
$$

where we use the fact that $\Pi(U^c \mid W, Z_1, \ldots, Z_n) \leq 1$.

REFERENCES

Beran J, Feng Y, Ghosh S and Kulik R (2013). Long-Memory Processes: Probabilistic Properties and Statistical Methods. Springer, Heidelberg. 10.1007/978-3-642-35512-7

Chakraborty A, Ovaskainen O and Dunson DB (2022a). Supplement to “Bayesian semiparametric long memory models for discretized event data.” 10.1214/21-AOAS1546SUPPA

Chakraborty A, Ovaskainen O and Dunson DB (2022b). Code to implement methods in “Bayesian semiparametric long memory models for discretized event data.” 10.1214/21-AOAS1546SUPPB

Chen Y, Härdle WK and Pigorsch U (2010). Localized realized volatility modeling. J. Amer. Statist. Assoc 105 1376–1393. 10.1198/jasa.2010.ap09039

Chib S and Greenberg E (1998). Analysis of multivariate probit models. Biometrika 85 347–361.

Cui Y and Lund R (2009). A new look at time series of counts. Biometrika 96 781–792. 10.1093/biomet/asp057

Czado C, Gneiting T and Held L (2009). Predictive model assessment for count data. Biometrics 65 1254–1261. 10.1111/j.1541-0420.2009.01191.x [PubMed: 19432783]

Davis RA, Holan SH, Lund R and Ravishanker N (2016). Handbook of Discrete-Valued Time Series. CRC Press.

Davison A and Ramesh N (1996). Some models for discretized series of events. J. Amer. Statist. Assoc 91 601–609.

De Camargo U, Roslin T and Ovaskainen O (2019). Spatio-temporal scaling of biodiversity in acoustic tropical bird communities. Ecography 42 1936–1947.

Fearnhead P and Sherlock C (2006). An exact Gibbs sampler for the Markov-modulated Poisson process. J. R. Stat. Soc. Ser. B. Stat. Methodol 68 767–784. 10.1111/j.1467-9868.2006.00566.x

Fischer W and Meier-Hellstern K (1993). The Markov-modulated Poisson process (MMPP) cookbook. Perform. Eval 18 149–171. 10.1016/0166-5316(93)90035-S

Franzke CL, Barbosa S, Blender R, Fredriksen H-B, Laepple T, Lambert F, Nilsen T, Rypdal K, Rypdal M et al. (2020). The structure of climate variability across scales. Reviews of Geophysics 58 e2019RG000657.

Geweke J and Porter-Hudak S (1983). The estimation and application of long memory time series models. J. Time Series Anal 4 221–238. 10.1111/j.1467-9892.1983.tb00371.x

Graves T, Gramacy R, Watkins N and Franzke C (2017). A brief history of long memory: Hurst, Mandelbrot and the road to ARFIMA, 1951–1980. Entropy 19 437.

Hall P and Hart JD (1990). Nonparametric regression with long-range dependence. Stochastic Process. Appl 36 339–351. 10.1016/0304-4149(90)90100-7

Hurst HE (1951). Long-term storage capacity of reservoirs. Trans. Amer. Soc. Civ. Eng 116 770–799.

Jacobs PA and Lewis PAW (1978a). Discrete time series generated by mixtures. I. Correlational and runs properties. J. Roy. Statist. Soc. Ser. B 40 94–105.

Jacobs PA and Lewis PAW (1978b). Discrete time series generated by mixtures. II. Asymptotic properties. J. Roy. Statist. Soc. Ser. B 40 222–228.

Jia Y, Kechagias S, Livsey J, Lund R and Pipiras V (2021). Latent Gaussian count time series. J. Amer. Statist. Assoc 1–28.

Kolassa S (2016). Evaluating predictive count data distributions in retail sales forecasting. Int. J. Forecast 32 788–803.

Krebs JR and Kacelnik A (1983). The dawn chorus in the great tit (Parus major): Proximate and ultimate causes. Behaviour 83 287–308.

Laiolo P (2010). The emerging significance of bioacoustics in animal species conservation. Biol. Conserv 143 1635–1645.

Laskin N (2003). Fractional Poisson process Commun. Nonlinear Sci. Numer. Simul 8 201–213. 10.1016/S1007-5704(03)00037-6

Livsey J, Lund R, Kechagias S and Pipiras V (2018). Multivariate integer-valued time series with flexible autocovariances and their application to major hurricane counts. Ann. Appl. Stat 12 408–431. 10.1214/17-AOAS1098

Lo AW (1989). Long-term memory in stock market prices. Technical Report, National Bureau of Economic Research.

Mandelbrot BB and Van Ness JW (1968). Fractional Brownian motions, fractional noises and applications. SIAM Rev 10 422–437. 10.1137/1010093

Mandelbrot BB and Wallis JR (1969). Some long-run properties of geophysical records. Water Resour. Res 5 321–340.

Mckenzie E (1985). Some simple models for discrete variate time series 1. J. Am. Water Resour. Assoc 21 645–650.

Mckenzie E (1986). Autoregressive moving-average processes with negative-binomial and geometric marginal distributions. Adv. in Appl. Probab 18 679–705. 10.2307/1427183

Mckenzie E (1988). Some ARMA models for dependent sequences of Poisson counts. Adv. in Appl. Probab 20 822–835. 10.2307/1427362

Mikosch T and Stărică C (2004). Nonstationarities in financial time series, the long-range dependence, and the IGARCH effects. Rev. Econ. Stat 86 378–390.

Mitrinovic DS and Vasic PM (1970). Analytic Inequalities 1. Springer.

Ogata Y and Abe K (1991). Some statistical features of the long-term variation of the global and regional seismic activity. International Statistical Review/Revue Internationale de Statistique 139–161.

Ovaskainen O, DE Camargo UM and Somervuo P (2018). Animal sound identifier (ASI): Software for automated identification of vocal animals. Ecol. Lett 21 1244–1254. 10.1111/ele.13092 [PubMed: 29938881]

Peng C-K, Buldyrev SV, Havlin S, Simons M, Stanley HE and Goldberger AL (1994). Mosaic organization of DNA nucleotides. Phys. Rev. E 49 1685.

Pipiras V and Taqqu MS (2017). Long-Range Dependence and Self-Similarity. Cambridge Series in Statistical and Probabilistic Mathematics 45. Cambridge Univ. Press, Cambridge.

Ramesh N, Thayakaran R and Onof C (2013). Multi-site doubly stochastic Poisson process models for fine-scale rainfall. Stoch. Environ. Res. Risk Assess 27 1383–1396.

Rasmussen CE and Williams CKI (2006). Gaussian Processes for Machine Learning. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA.

Roberts GO and Rosenthal JS (2001). Optimal scaling for various Metropolis–Hastings algorithms. Statist. Sci 16 351–367. 10.1214/ss/1015346320

Robinson PM (1995). Gaussian semiparametric estimation of long range dependence. Ann. Statist 23 1630–1661. 10.1214/aos/1176324317

Rue H and Held L (2005). Gaussian Markov Random Fields: Theory and Applications. Monographs on Statistics and Applied Probability 104. CRC Press/CRC, Boca Raton, FL. 10.1201/9780203492024

Rue H, Martino S and Chopin N (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. J. R. Stat. Soc. Ser. B. Stat. Methodol 71 319–392. 10.1111/j.1467-9868.2008.00700.x

Samorodnitsky G (2006). Long range dependence. Found. Trends Stoch. Syst 1 163–257. 10.1561/0900000004

Slabbekoorn H and Smith TB (2002). Bird song, ecology and speciation. Philos. Trans. R. Soc. Lond. B, Biol. Sci 357 493–503. 10.1098/rstb.2001.1056 [PubMed: 12028787]

Sørbye SH and RUE H (2018). Fractional Gaussian noise: Prior specification and model comparison. Environmetrics 29 e2457. 10.1002/env.2457

Stern R and Coe R (1984). A model fitting analysis of daily rainfall data. J. R. Stat. Soc., A 147 1–18.

Tagliazucchi E, Von Wegner F, Morzelewski A, Brodbeck V, Jahnke K and Laufs H (2013). Breakdown of long-range temporal dependence in default mode and attention networks during deep sleep. Proc. Natl. Acad. Sci. USA 110 15419–15424. [PubMed: 24003146]

Tiao GC, Phadke M and Box GE (1976). Some empirical models for the Los Angeles photochemical smog data. J. Air Pollut. Control Assoc 26 485–490. [PubMed: 1262603]

Weymer BA, Wernette P, Everett ME and Houser C (2018). Statistical modeling of the long-range dependent structure of barrier island framework geology and surface geomorphology. Earth Surf. Dyn 6 431–450.

Willinger W, Paxson V, Riedi RH and Taqqu MS (2003). Long-range dependence and data network traffic. In Theory and Applications of Long-Range Dependence 373–407. Birkhäuser, Boston, MA.

Zhou Z (2012). Measuring nonlinear dependence in time-series, a distance correlation approach. J. Time Series Anal 33 438–457. 10.1111/j.1467-9892.2011.00780.x

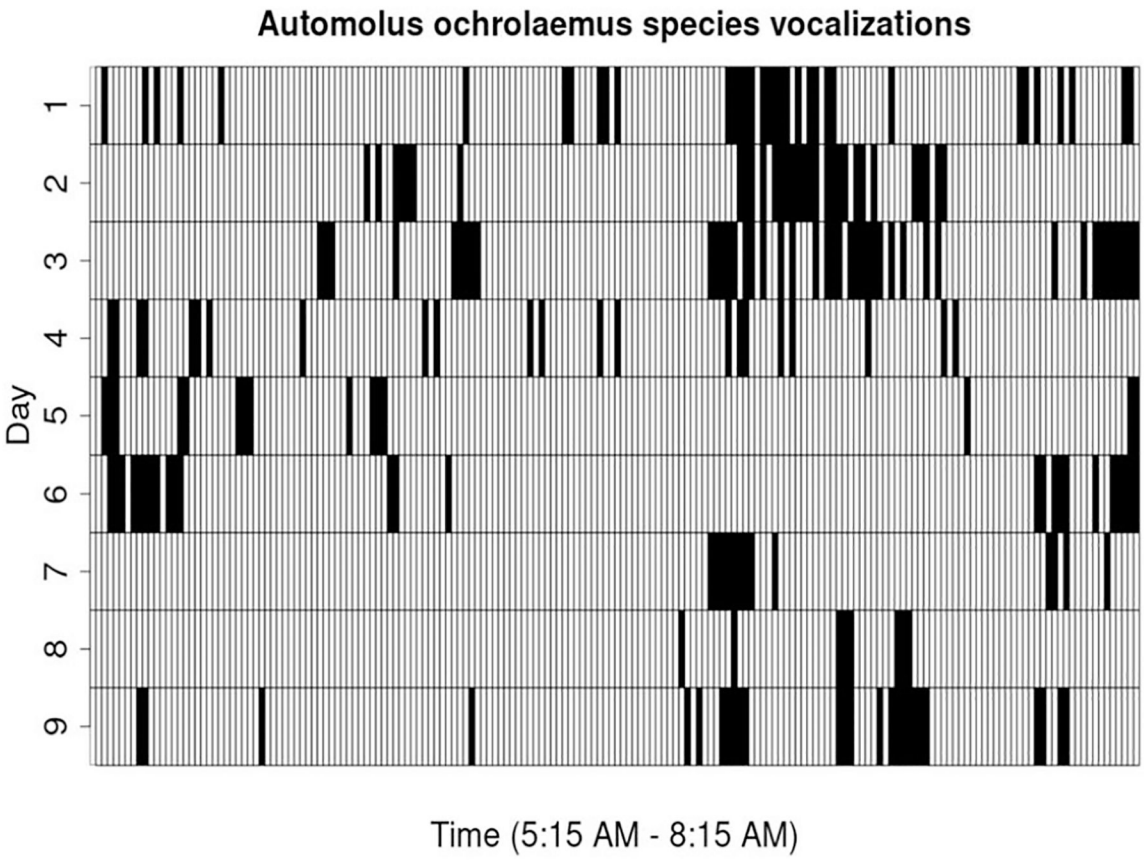


Fig. 1.

Binary sequence of all vocalizations of birds from the Automolus ochrolaemus species, during nine days (not necessarily consecutive) of recording. White and black grids represent absence or presence of vocalizations, respectively.

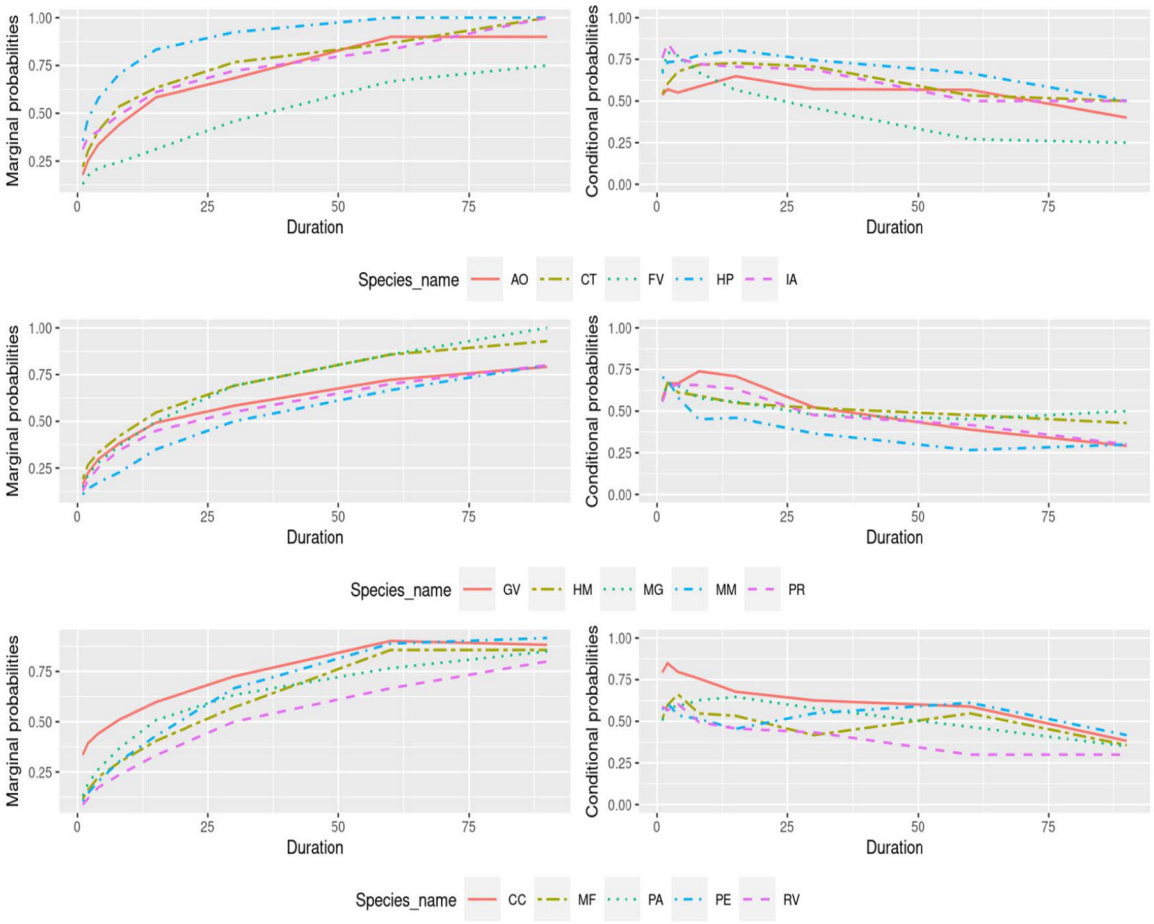


Fig. 2.

Marginal (left panel) and conditional (right panel) probabilities of bird vocalizations for 15 different species at different time scales $\Delta t = \{1, 2, 4, 9. 15, 30, 60, 90\}$. The names of the species from Table 1 have been abbreviated using the first letter of their genus name (first word) and first letter of their specific epithet (second word), due to space constraints.

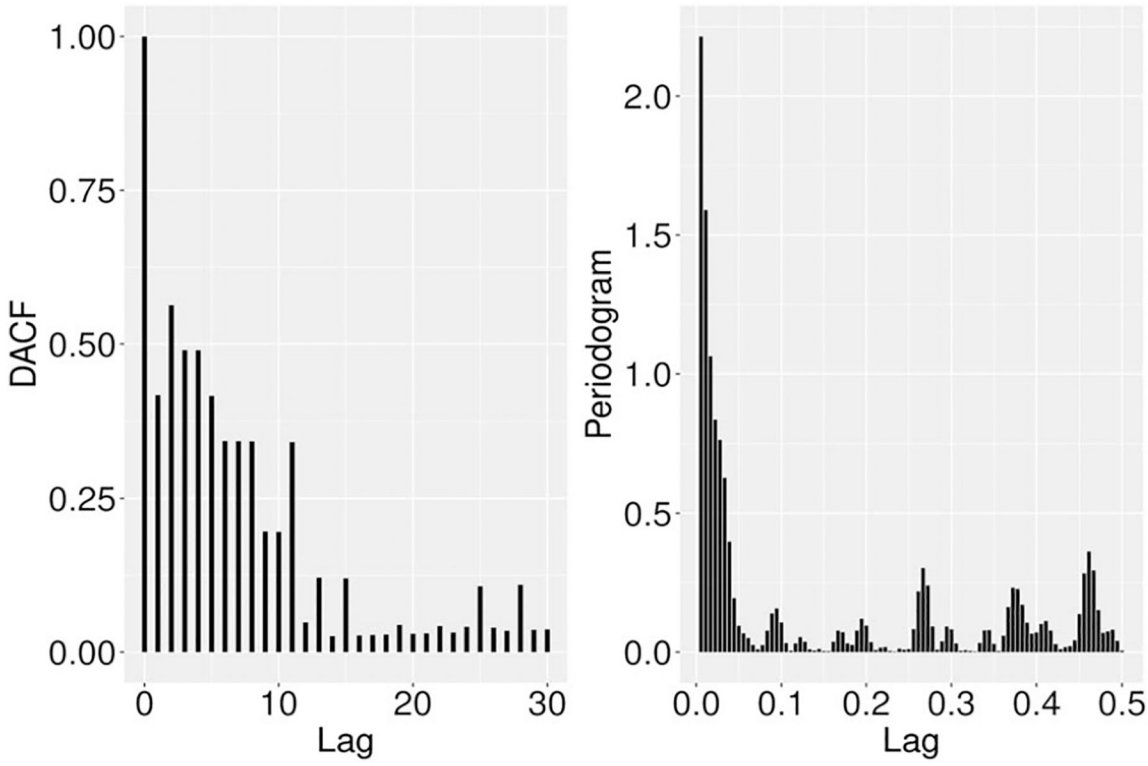


Fig. 3.

Distance autocorrelation (left panel) at different lags for the binary indicators obtained from one day of recording of vocalizations for the species Corythopis torquata. On the right panel the periodogram for the same time series is shown.

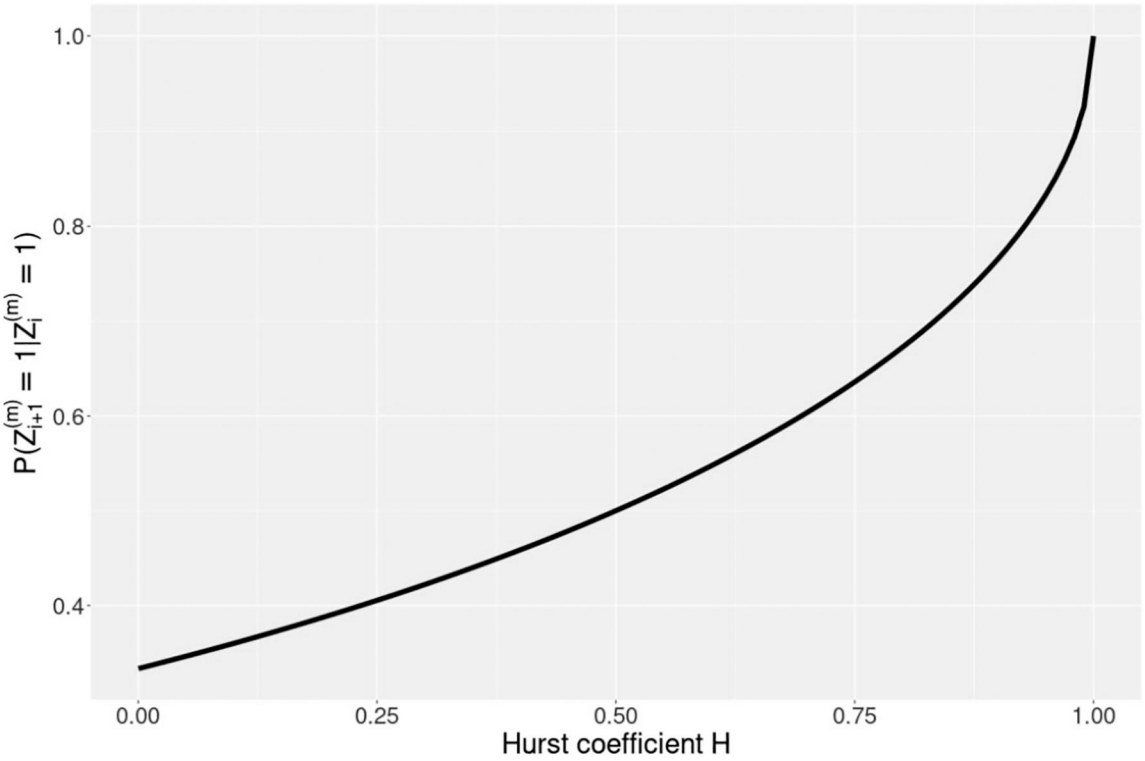


Fig. 4.

Relation between the Hurst coefficient H and the conditional probabilities obtained from equation (3.8).

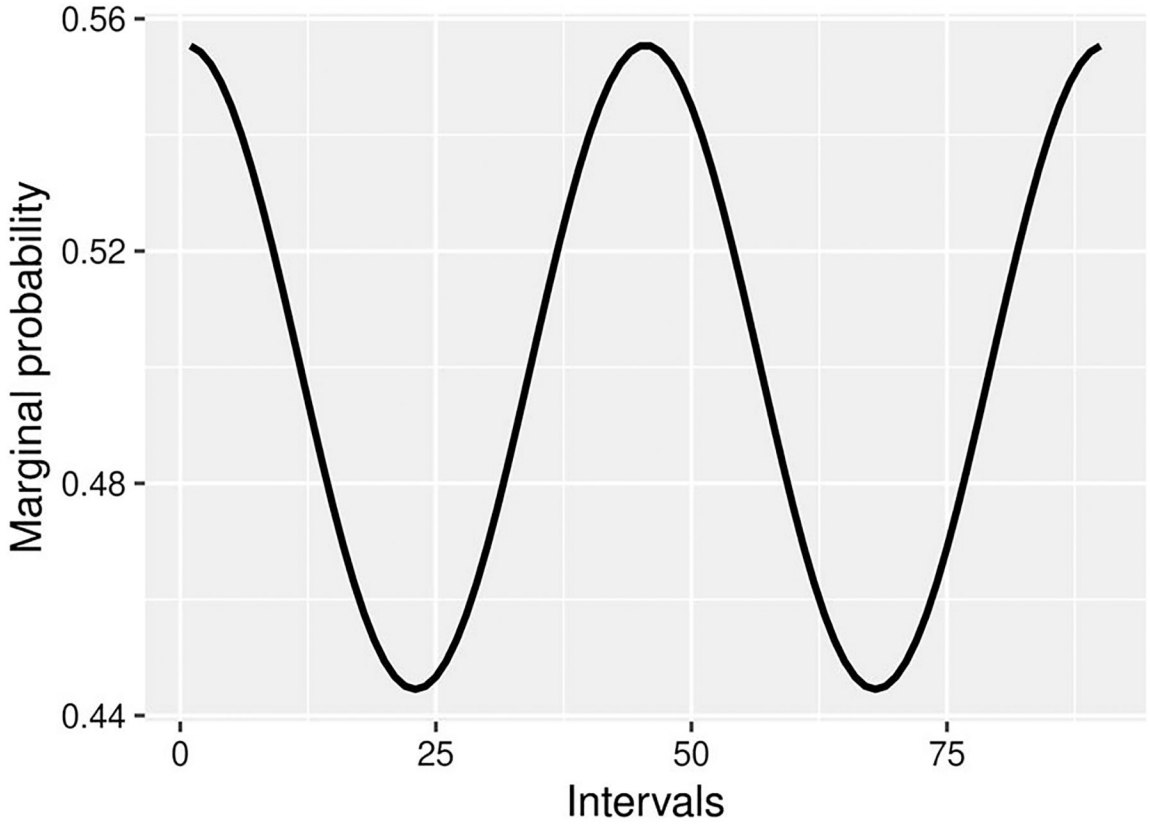
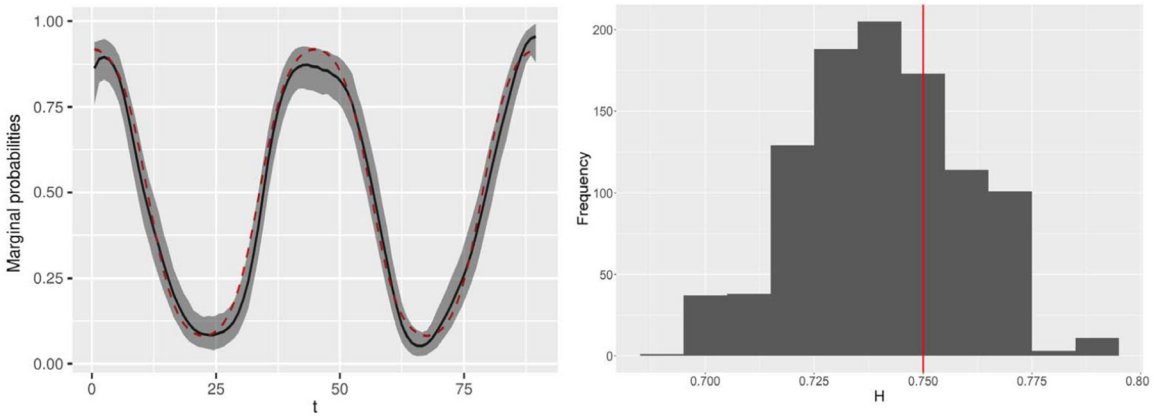


Fig. 5.

Variation in marginal probabilities of observing a vocalization or an event when $f(t) = \sin(4\pi t)/90$ for time intervals (0, 1], (1, 2], …, (89, 90]. Here, $\tau = 1$, and the marginal probabilities are calculated as $\Phi\{f(i + 1) - f(i)\}$ for i = 0, …, 89.



(a) $y(t) = \sin \frac{4\pi t}{90} + B_H(t)$ (b) Corresponding posterior samples of H

Fig. 6.

Figure (a) shows the posterior mean and 95% credible bands for marginal probabilities in one minute intervals when $f(t) = \sin\frac{4\pi t}{90}$. The values of the Hurst coefficient and the number of replications were H = 0.75 and R = 50, respectively. Red dashed and black solid lines correspond to the true values and the posterior mean, respectively. Gray shaded regions are credible bands. Corresponding posterior samples of H are shown in (b). A red line is added at the true value H = 0.75.

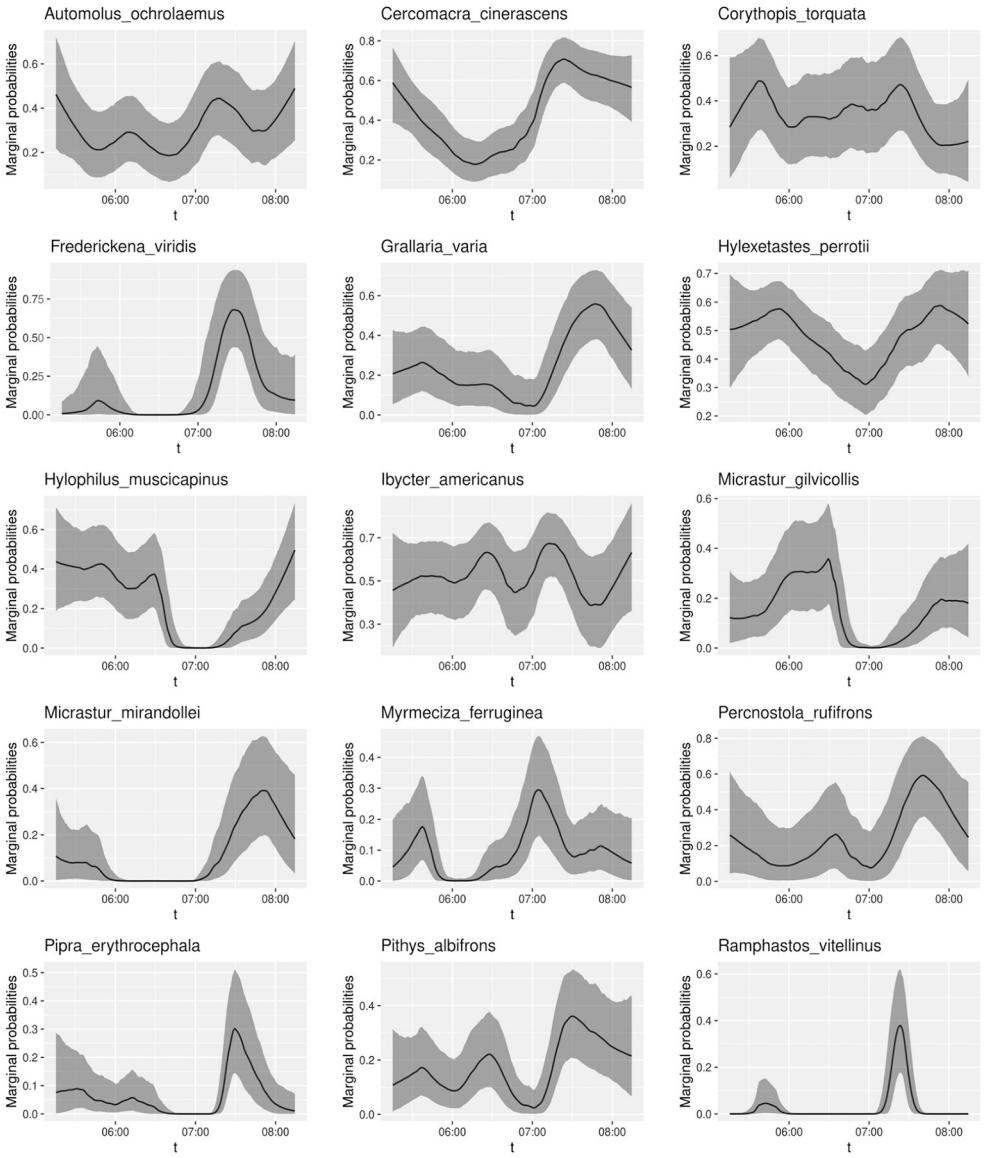


Fig. 7.

Smoothed marginal probabilities of vocalization obtained by fitting model (3.9) for the 15 species listed in Table 1 for 180 test intervals of duration one minute from 5.15–8.15 a.m. Shaded regions are 95% credible intervals and black lines are posterior means.

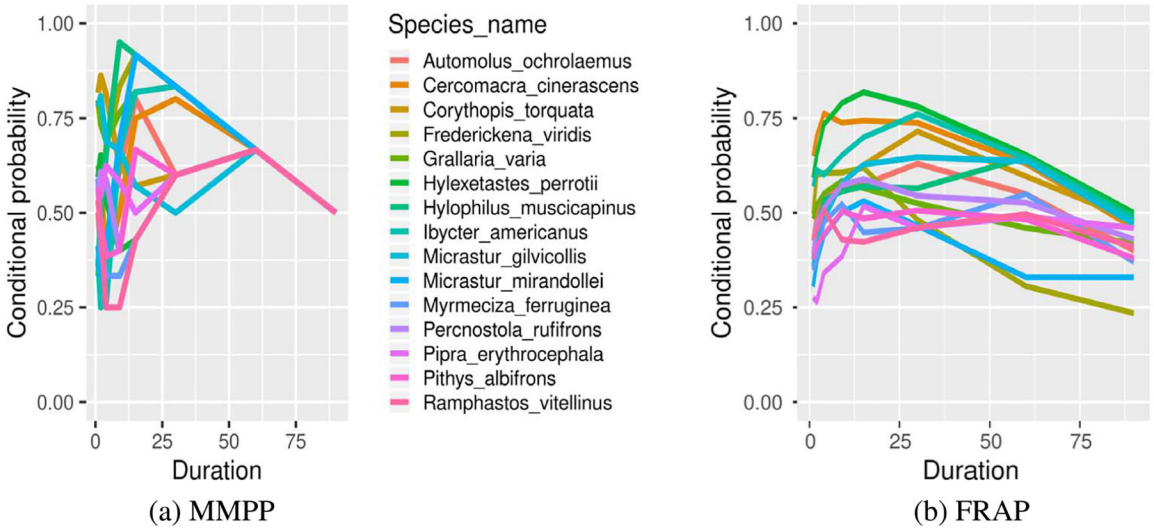


Fig. 8.

Conditional probabilities of vocalizations for the 15 different species at different time scales $\Delta t = \{1, 2, 4, 9, 15, 30, 60, 90\}$ obtained from fitted model for the MMPP (left) and samples from posterior predictive for the FRAP model (right).

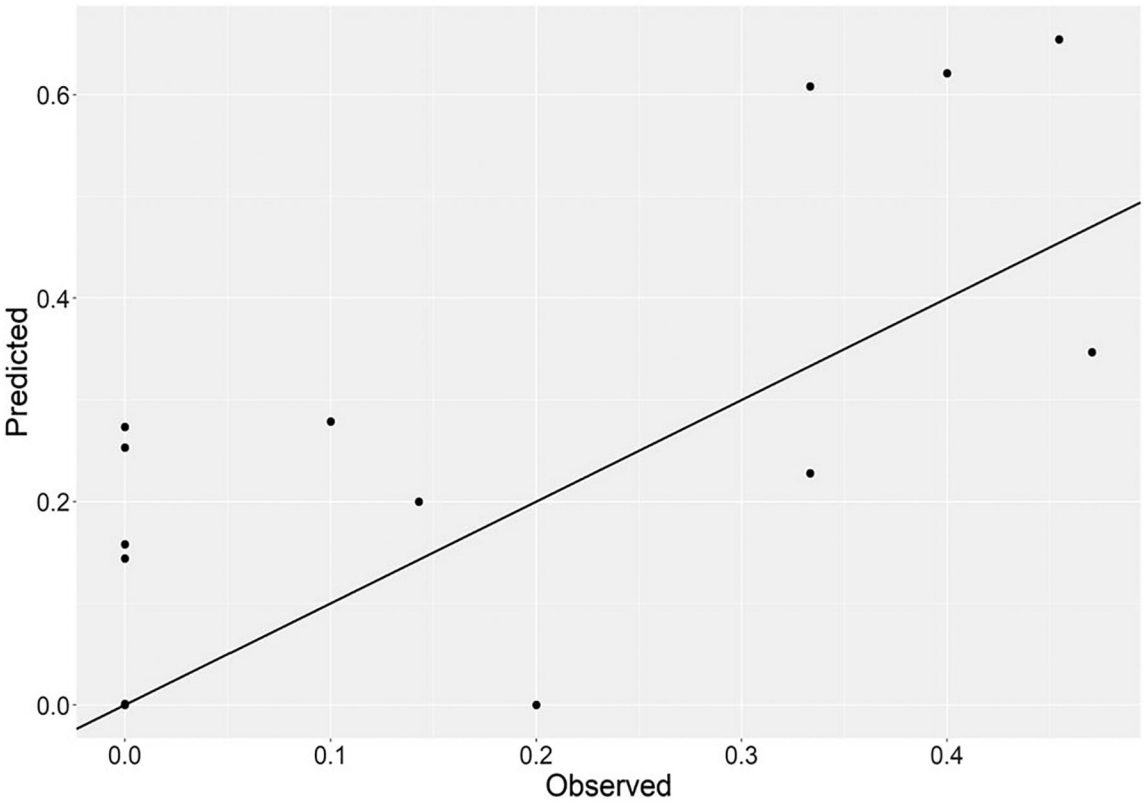


Fig. 9.

Difference between one-step ahead prediction probabilities for $Z_{180} = 1$ and observed probabilities for the 15 species of bird analyzed here.

Table 1

Estimated Hurst exponents for the 15 bird species using ($\widehat{H}_{\mathrm{DFA}}$), Geweke and Porter-Hudak (1983) ($\widehat{H}_{\mathrm{GPH}}$), Robinson (1995) ($\widehat{H}_{\mathrm{W}}$), Livsey et al. (2018) ($\widehat{H}_{\mathrm{QMLE}}$), and the FRAP model. For the FRAP model we include the posterior mean ($\widehat{H}_{\mathrm{FRAP}}$) along with the 95% credible intervals ($\widehat{H}_{\mathrm{LR}}$, $\widehat{H}_{\mathrm{UR}}$)

	$\widehat{H}_{\mathrm{GPH}}$			$\widehat{H}_{\mathrm{W}}$						
Species name	$m = n^{1/2}$	$m = n^{2/3}$	$m = n^{4/5}$	$m = n^{1/2}$	$m = n^{2/3}$	$m = n^{4/5}$	$\widehat{H}_{\mathrm{QMLE}}$	$\widehat{H}_{\mathrm{LR}}$	$\widehat{H}_{\mathrm{FRAP}}$	$\widehat{H}_{\mathrm{UR}}$
Automolus ochrolaemus	0.72	0.82	0.89	0.65	0.86	0.86	0.67	0.85	0.89	0.95
Cercomacra cinerascens	1.04	1.06	1.17	0.99	1.13	1.04	0.83	0.90	0.92	0.94
Corythopis torquata	0.91	0.94	0.89	0.78	0.98	0.85	0.72	0.80	0.84	0.88
Frederickena viridis	1.25	1.03	1.07	1.20	1.10	1.02	0.63	0.79	0.86	0.93
Grallaria varia	1.02	0.91	0.91	0.95	0.98	0.91	0.66	0.84	0.89	0.93
Hylexetastes perrotii	0.79	0.94	0.92	0.71	0.98	0.89	0.85	0.89	0.93	0.95
Hylophilus muscicapinus	0.80	0.99	0.96	0.68	0.99	0.92	0.69	0.80	0.87	0.93
Ibycter americanus	0.96	1.08	1.10	0.88	1.21	0.99	0.81	0.90	0.94	0.96
Micrastur gilvicollis	0.84	0.90	0.98	0.73	0.95	0.91	0.64	0.80	0.84	0.89
Micrastur mirandollei	0.83	0.96	1.05	0.76	1.02	1.02	0.61	0.83	0.88	0.93
Myrmeciza ferruginea	0.87	1.10	0.99	0.78	0.92	0.88	0.61	0.81	0.85	0.88
Percnostola rufifrons	0.81	0.96	0.89	0.78	0.99	0.90	0.63	0.87	0.92	0.96
Pipra erythrocephala	0.63	0.86	0.96	0.60	0.95	0.93	0.60	0.79	0.84	0.88
Pithys albifrons	0.77	0.86	0.90	0.71	0.89	0.86	0.63	0.83	0.87	0.90
Ramphastos vitellinus	0.84	0.99	1.01	0.76	1.02	0.96	0.59	0.77	0.83	0.89

Table 2

Relative mean square error (ReMSE) for different choices of the latent trend function f(t) for the model (3.10) under hierarchy (3.12). For each f(t), three values of the Hurst exponent are considered: {0.5, 0.75, 0.9} together with {10, 25, 50} replications. The results reported are averages of 30 independent simulation experiments for each combination

		$f_1(t)$		$f_2(t)$		$f_3(t)$		$f_4(t)$		$f_5(t)$	
Hurst exponent (H)	**Replications (R)**	**MSE**	$\widehat{H}$	**MSE**	$\widehat{H}$	**MSE**	$\widehat{H}$	**MSE**	$\widehat{H}$	**MSE**	$\widehat{H}$
0.5	10	1.26	0.55	1.02	0.49	0.12	0.52	0.09	0.52	1.38	0.50
	25	0.58	0.48	0.40	0.48	0.01	0.50	0.01	0.51	0.96	0.51
	50	0.40	0.54	0.17	0.50	0.007	0.48	0.005	0.50	0.08	0.50
0.75	10	2.13	0.76	1.88	0.76	0.14	0.74	0.28	0.76	4.81	0.74
	25	1.37	0.75	1.20	0.77	0.06	0.76	0.03	0.75	1.46	0.75
	50	0.84	0.74	0.24	0.74	0.04	0.75	0.02	0.75	0.55	0.74
0.9	10	4.18	0.88	6.52	0.87	0.70	0.88	0.20	0.90	14.96	0.87
	25	3.08	0.89	2.61	0.89	0.29	0.89	0.18	0.89	5.87	0.93
	50	1.11	0.89	0.99	0.88	0.08	0.87	0.07	0.89	3.34	0.88

Table 3

Coverage probability (CP) of 95% credible intervals for the Hurst exponent under the hierarchy (3.12). Also included are the average length (l) of the credible intervals with corresponding standard deviation inside parenthesis. The number of replicates in each case is R = 5

	$f_1(\cdot)$		$f_2(\cdot)$		$f_3(\cdot)$		$f_4(\cdot)$		$f_5(\cdot)$	
H	**CP**	**l**	**CP**	**l**	**CP**	**l**	**CP**	**l**	**CP**	**l**
0.5	0.97	0.16 (0.05)	0.94	0.19 (0.04)	0.92	0.20 (0.02)	0.92	0.21 (0.03)	0.98	0.20 (0.02)
0.75	0.91	0.15 (0.03)	0.92	0.14 (0.14)	0.92	0.14 (0.02)	0.90	0.14 (0.02)	0.93	0.13 (0.01)
0.9	0.90	0.13 (0.10)	0.89	0.14 (0.11)	0.91	0.12 (0.10)	0.90	0.12 (0.09)	0.94	0.14 (0.10)