The Astrophysical Journal Supplement Series, 259:32 (17pp), 2022 April https://doi.org/10.3847/1538-4365/ac49ec

Pitfalls of Periodograms: The Nonstationarity Bias in the Analysis of Quasiperiodic Oscillations

Moritz Hübner[1,2], Daniela Huppenkothen[3], Paul D. Lasky[1,2], and Andrew R. Inglis[4,5]
[1] School of Physics and Astronomy, Monash University, Clayton, VIC 3800, Australia; moritz.thomas.huebner@gmail.com
[2] OzGrav: The ARC Centre of Excellence for Gravitational Wave Discovery, Clayton, VIC 3800, Australia
[3] SRON Netherlands Institute for Space Research, 2333 CA Leiden, The Netherlands
[4] Physics Department, The Catholic University of America, Washington, DC 20064, USA
[5] Solar Physics Laboratory, NASA Goddard Space Flight Center, Greenbelt, MD 20771, USA
Received 2021 August 17; revised 2021 December 27; accepted 2022 January 10; published 2022 March 14

Abstract

Quasiperiodic oscillations (QPOs) are an important key to understand the dynamic behavior of astrophysical objects during transient events like gamma-ray bursts, solar flares, and magnetar flares. Searches for QPOs often use the periodogram of the time series and perform spectral density estimation using a Whittle likelihood function. However, the Whittle likelihood is only valid if the time series is stationary since the frequency bins are otherwise not statistically independent. We show that if time series are nonstationary, the significance of QPOs can be highly overestimated and estimates of the central frequencies and QPO widths can be overconstrained. The effect occurs if the QPO is only present for a fraction of the time series and the noise level is varying throughout the time series. This can occur, for example, if background noise from before or after the transient is included in the time series or if the low-frequency noise profile varies strongly over the time series. Thus, we highlight the importance of careful segment selection prior to the analysis. We confirm the presence of this bias in previously reported results from solar flare data and show that significance can be highly overstated. Finally, we provide some suggestions that help identify whether an analysis is affected by this bias.

Unified Astronomy Thesaurus concepts: Time series analysis (1916); Period search (1955); Fast Fourier transform (1958); Solar x-ray flares (1816); Solar flares (1496); X-ray transient sources (1852); Bayesian statistics (1900); Bayes factor (1919); Bayesian information criterion (1920); Astrostatistics tools (1887); Model selection (1912); Astrostatistics strategies (1885)

1. Introduction

Quasiperiodic oscillations (QPOs)[6] are a common astrophysical phenomenon that are regularly observed across a variety of sources. While there is ample discussion about how they emerge in their respective sources, it is worth reexamining existing techniques for their detection and characterization. Observations of QPOs are scant for some objects such as in magnetar flares (Israel et al. 2005; Strohmayer & Watts 2006, 2005; Watts & Strohmayer 2006; Huppenkothen et al. 2012, 2014a, 2014b; Miller et al. 2019), or contentious in others like gamma-ray bursts (GRBs; Cenko et al. 2010; De Luca et al. 2010; Iwakiri et al. 2010; Morris et al. 2010; Tarnopolski & Marchenko 2021), and false detections may lead to unjustified theoretical inferences. Moreover, even if there are ample detections of QPOs like in solar flares (Nakariakov & Melnikov 2009; Van Doorsselaere et al. 2016; Zimovets et al. 2021) or X-ray binaries (Ingram & Motta 2019), making accurate inferences about their parameters and significance is essential for understanding the mechanisms that produce them.

There are an array of methods being applied to time series data to find QPOs. The most common tests involve periodograms and are based on either outlier analyses or Bayesian tests (Vaughan 2005, 2009). Other methods include wavelet analyses (Foster 1996; Torrence & Compo 1998; Lachowicz & Done 2010) and those based on Gaussian processes (Rasmussen & Williams 2006; Foreman-Mackey et al. 2017; Covino et al. 2020; Zhu & Thrane 2020).

In this paper, we explore the effects that the nonstationary nature of a transient light curve and the potential nonstationarity of a QPO within it have on detection significance and characterization when methods are used that assume the underlying processes to be stationary. We show that we are likely to overestimate the significance of the quasiperiodic signal under these conditions. This bias arises because neighboring bins in periodograms are only statistically independent for stationary time series. We thus refer to this effect as the "nonstationarity bias," which to the best of our knowledge has not been reported in the astrophysical literature to date. We show that merely adding photon-counting noise before or after a transient leads one to greatly overestimate the significance of a QPO and overconstrain the QPO parameters. More critically, this effect also occurs if the whole transient event is selected for analysis but the QPO is only present for a fraction of it.

It is our intuition that astrophysicists would rather be conservative and include more of the time series that does not include the QPO, rather than cherry-picking the part of the light curve that appears most likely to contain it. The erroneous reasoning behind this might be the belief that adding more noise to the signal should not increase the significance of the QPO. However, choosing a longer time segment can both increase and decrease the significance depending on the circumstances.

[6] In solar physics the term "quasiperiodic pulsation" is preferred. We apply the term QPO generically throughout for simplicity.

We structure this paper as follows. In Section 2 we lay out the Bayesian methods and models that we use throughout. Thereafter, in Section 3, we explain how the nonstationarity bias arises both on a conceptual level and with detailed mathematical arguments. We show in Section 4, based on simulated data, that the nonstationarity bias exists empirically. In Section 5 we show how the nonstationarity bias affects the analysis of solar flare light curves. We conclude in Section 6 and provide some outlook on how alternative methods can potentially better handle nonstationary data sets.

2. Methods

In the following, we recapitulate the Fourier-based methods commonly used to analyze QPOs that are relevant to this study. A general comprehensive overview on the topic of spectral density estimation for time series analysis is given in many popular textbooks, such as chapter 7 in Chatfield & Xing (2019), as well as in van der Klis (1989) and Barret & Vaughan (2012) in the context of astrophysical light curves.

2.1. Whittle Likelihood

A periodogram is an estimate of the power spectral density (PSD) of the signal based on a given time series $x(t)$. We calculate the periodogram $I(f)$ as the absolute square of the discrete Fourier transform of the signal

$$I(f) = \left| \sum_{i=1}^{N} x(t_i) \exp(-i2\pi f t_i) \right|^2 . \quad (1)$$

It is prudent to convolve time series data with a window function in practice to avoid spectral leakage, i.e., artificial features in the periodogram that arise owing to the finite nature of the time series (Percival & Walden 1993). For the analyses throughout this paper, we use the Hann window function (Blackman & Tukey 1958a, 1958b). We also performed our analyses with a Tukey window but found no quantitative differences in the results.

Periodograms are established as a standard method in the search for QPOs in solar flares (Inglis et al. 2015, 2016; Hayes et al. 2020) and in astrophysical transients elsewhere (e.g., Huppenkothen et al. 2012, 2014a, 2014b, 2017; Miller et al. 2019). Since individual frequency bins $I(f_j) = I_j$ are calculated by taking the sum of the square of the normal-distributed real and imaginary parts of the Fourier series, it follows that they are χ^2_2-distributed around the PSD $S(f_j) = S_j$ (Whittle 1951; Chatfield & Xing 2019). This specific case of a χ^2_2-distribution is identical to an exponential distribution around S_j

$$p(I_j|S_j) = \frac{1}{S_j} \exp(-I_j/S_j), \quad (2)$$

where $p(I_j|S_j)$ is the conditional probability to observe the power I_j given an underlying PSD S_j. This relation is generically true for any individual frequency bin in any periodogram. Assuming that all bins are statistically independent from one another, we obtain the Whittle likelihood function by taking the product over all $N/2$ frequency bins of a periodogram corresponding to a time series with N points

$$L(I|S) = \prod_{j=1}^{N/2} \frac{1}{S_j} \exp(-I_j/S_j). \quad (3)$$

We emphasize here that this is only true in the stationary limit (for infinitely long time series with equally spaced measurements and homoscedastic Gaussian white noise), and this will lead to biased estimates of the PSD in general, as we show in Section 3. Nevertheless, Equation (3) is the standard likelihood that is used for spectral density estimation.

2.2. Models of Power Spectra

Many astrophysical transients show excess power at low frequencies, and it is often assumed that this can be modeled using a red-noise process (see, e.g., Huppenkothen et al. 2012; Inglis et al. 2015, 2016; Broomhall et al. 2019; Ingram & Motta 2019; Miller et al. 2019). One basic noise model that is commonly used is a combination of a red-noise power law with amplitude A and spectral index α and a white-noise amplitude C (Inglis et al. 2015, 2016), i.e.,

$$\begin{aligned} S^{\mathrm{RW}}(f) &= S^{\mathrm{R}}(f) + S^{\mathrm{W}} \\ &= Af^{-\alpha} + C, \end{aligned} \quad (4)$$

where we use the R and W superscripts as short-hand for "red noise" and "white noise," respectively. This spectral shape usually emerges as a combination from the overall structure of the transient and additional variability on smaller timescales. White noise, specifically, can arise as a result of either uncorrelated noise fluctuation of the source or Poisson counting noise that we often encounter in astrophysics. The shape is thus not due to a stationary process but rather due to a combination of unknown deterministic and nonstationary stochastic processes adding up to mimic a red-noise spectrum (Huppenkothen et al. 2012).

A common way to model a QPO is to add a Gaussian or Lorentzian enhancement to S^{RW}, e.g.,

$$\begin{aligned} S^{\mathrm{RWQ}}(f) &= S^{\mathrm{RW}}(f) + S^{\mathrm{Q}}(f) \\ &= S^{\mathrm{RW}}(f) + \frac{B}{\pi\sigma} \frac{\sigma^2}{(f - f_0)^2 + \sigma^2}, \end{aligned} \quad (5)$$

where B is the QPO amplitude, f_0 is its mean frequency, and σ is the half-width at half-maximum scale parameter. Explicitly modeling the red-noise component is important because the QPO is likely to overlap with the red-noise-dominated part of the PSD. Not considering the red-noise-dominated part of the PSD may thus lead to false positives.

There are extensions to the noise model, such as the broken or bent power law (B superscript), that can fit more structured red noise and are often a better fit to the data

$$S^{\mathrm{BW}}(f) = Af^{-\alpha_1} \left(1 + \left(\frac{f}{\delta} \right)^{\alpha_2 - \alpha_1} \right) + C, \quad (6)$$

where $\alpha_{1,2}$ are the power-law indices present before and after the break frequency δ where the power law changes, and we also enforce $\alpha_2 < \alpha_1$ to avoid degeneracies. There are also various different formulations of bent or broken power laws with various degrees of smoothness.

2.3. Model Selection and Parameter Estimation

There are several ways to assert the significance of a QPO when using spectral density estimation. A widely used frequentist way to detect QPOs is to use outlier statistics. As a first step, one fits the PSD using a model that does not contain a QPO, e.g., Equation (4) or Equation (6). Thereafter, every frequency bin is checked if its amplitude exceeds a set probability threshold based on the χ^2_2-distribution in Equation (2). One also has to account for the number of trials, i.e., the number of frequency bins tested by applying a Bonferroni correction (Bonferroni 1936). Additionally, it is possible to rebin the periodogram into a smaller number of bins. For QPOs with widths larger than the width of a single frequency bin, rebinning may be useful because it increases the significance of the QPO relative to the noise continuum.

Alternative to the outlier statistics, one can cast this as a model Bayesian selection problem where we find the preferred model to fit our data, in our case Equations (4), (5), and (6). There are multiple ways to quantify model preference in Bayesian statistics. One approach that has been used in solar physics (Ireland et al. 2015; Inglis et al. 2015, 2016) is to calculate the Bayesian information criterion (BIC),

$$\mathrm{BIC} = k\ln(n) - 2\ln(L_{\max}), \tag{7}$$

where k is the number of free parameters (e.g., three for S^{RW}, six for S^{RWQ}, and five for S^{BW}), $n = N/2$ is the number of data points, and $L_{\max}$ is the maximum likelihood value. This method is relatively cheap computationally, as the calculation of the maximum likelihood can be obtained with relatively few likelihood evaluations using a maximization algorithm. A lower BIC indicates a better fit to the data; thus, the difference in BIC for S^{RW} and S^{RWQ} is a measure of statistical significance of the QPO,

$$\begin{aligned}\Delta\mathrm{BIC} &= \mathrm{BIC}(S^{\mathrm{RWQ}}) - \mathrm{BIC}(S^{\mathrm{RW}}) \\ &= (k_{\mathrm{RWQ}} - k_{\mathrm{RW}})\ln(N/2) - 2\ln\frac{L_{\max}(I\mid S^{\mathrm{RWQ}})}{L_{\max}(I\mid S^{\mathrm{RW}})},\end{aligned} \tag{8}$$

where $\Delta\mathrm{BIC} < 0$ would indicate that S^{RWQ} is preferred and vice versa. We can also perform model selection via calculation of Bayes factors BF. To understand Bayes factors, we start with Bayes's theorem

$$p(\theta|d, S) = \frac{\pi(\theta|S)L(d|\theta, S)}{Z(d|S)}, \tag{9}$$

where θ are the model parameters; d are the data, i.e., the periodogram in our case; p is the posterior probability of the parameters; π is the prior probability of the parameters; L is the likelihood of the data given the parameters; and Z is the evidence, or fully marginalized likelihood. All these probabilities are conditioned on a model we want to evaluate, e.g., a PSD S. The Bayes factor comparing two models is the ratio of their evidences. For example, the Bayes factor comparing S^{RWQ} and S^{RW} is

$$\mathrm{BF} = \frac{Z(d|S^{\mathrm{RWQ}})}{Z(d|S^{\mathrm{RW}})}. \tag{10}$$

It thereby measures the relative odds of the underlying data to have been produced by either model, though it does not measure whether the model itself is a good fit to the data, similar to the BIC. The evidence is calculated by rearranging and integrating Bayes's theorem

$$Z(d|S) = \int \pi(\theta)L(d|\theta, S)d\theta. \tag{11}$$

Evaluating this integral is much more computationally challenging than calculating the BIC but can be achieved thanks to improvements in algorithms such as nested sampling (Skilling 2006) and accessible software implementations such as BILBY (Ashton et al. 2019; Speagle 2020; Romero-Shaw et al. 2020). The Bayes factor obtained via evidence calculation is seen as the superior standard for model selection in Bayesian statistics since it involves prior beliefs about the distribution of parameters (Weakliem 1999). The penalty factor $k\ln(n)$ for the BIC also increases with the size of the data set and can thus favor overly simplistic models (Weakliem 1999; Gelman et al. 2013). Since the BIC relies on a point estimate of the likelihood, it is also a less reliable measure if the likelihood has multiple modes. Additionally, nested sampling provides posterior distributions on the model parameters. These distributions often have astrophysical implications.

Bayes factors and BICs work best if the underlying models are truly discrete, but S^{RW} is a special case of both S^{RWQ} and S^{BW}. There are reasonable criticisms on the use of Bayes factors or the BIC for cases like this, as S^{RW} is a subset of both S^{RWQ} and S^{BW} and thus not a truly discrete model (Gelman et al. 2013). In this case, the Bayes factor depends on prior choices of the parameters specific to S^{RWQ} and S^{BW}. Alternatively, we could use tests that measure whether the posterior of the QPO amplitude B is inconsistent with being zero and provides a better fit to the data. Nevertheless, we use Bayes factors as model selection criteria since prior choices will yield a constant difference in our Bayes factors and will not distort the overall trend when we investigate the nonstationarity bias. We also consider the effects on the ΔBICs where it is instructive.

For all the Bayes factor calculations and posterior samples in this publication we use DYNESTY via the BILBY interface (Ashton et al. 2019; Speagle 2020; Romero-Shaw et al. 2020).

3. Pitfalls of Periodograms

Assume a discrete time series of length T with N data points containing a QPO and noise. The underlying stochastic process is characterized in the frequency domain as a PSD $S(f)$, which can be estimated by calculating a periodogram $I(f)$. We recall that the value in any specific one of the $N/2$ frequency bins in the periodogram is proportional to the absolute square of the Fourier amplitude.

In the top panel of Figure 1 we extend a stationary time series by appending zeros on either end until we reach a length xT with xN data points, where we call x the *extension factor*. As we display in the bottom panel, this increases the number of frequency bins in the periodogram to $xN/2$, while decreasing the periodogram to $I(f)/x$ if we do not apply a normalization. Physically, this means that the QPO and noise appear to be persistent over the extended time period, boosting the QPO's significance.

The example displayed in Figure 1 is not realistic in practice but serves to illustrate the effect. A more realistic scenario occurs if the stationary time series is instead extended with

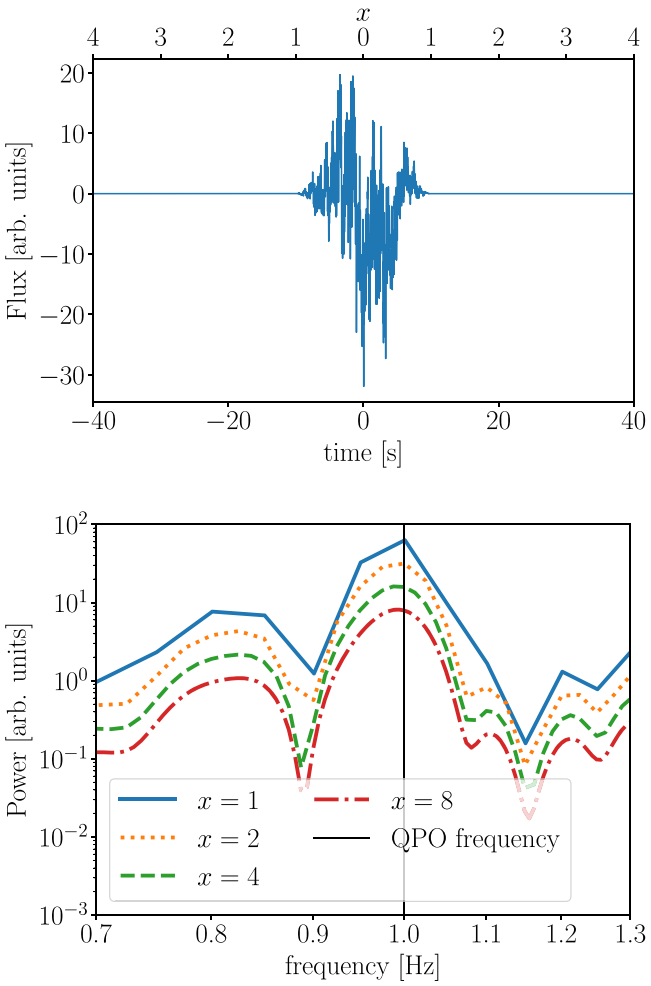

Figure 1. Effects of extending a time series with zeros. In the top panel, we show a stationary time series between −10 s and 10 s that we multiply with a Hann window to ensure a smooth turn-on from zero. We extended this time series with zeros on either end up to a total duration of 200 s. The upper x-axis shows how the extension factor x maps onto the time axis. Taking only the stationary data from −10 s to 10 s corresponds to $x = 1$, i.e., no extension has been applied. The extension factor in this case implies that we select data from $-10x$ s to $10x$ s. In the bottom panel, we show the periodogram given different extension factors, with the blue curve representing the stationary data ($x = 1$). The other colors (orange, green, red) show the effect of extending the time series with zeros on either end. In these curves, neighboring bins are not statistically independent, which manifests itself in the emerging smooth structure of the periodogram. The bins are also decreasing inversely proportional to x since we do not apply a normalization.

low-amplitude white noise and the QPO is within the red-noise-dominated part of the PSD. This raises the white-noise level relative to red noise and the QPO. Yet it still increases the significance of the QPO, which is determined by its amplitude relative to the dominant red-noise level. Such a scenario might occur if we overselect data surrounding a transient.

Another scenario is that the overall time series is nonstationary, and instead the QPO and the overall noise processes vary, or switch on and off, during a selected time window. For example, the QPOs reported in the 2004 SGR 1806−20 hyperflare are directly associated with specific rotational phases of the magnetar (Strohmayer & Watts 2005; Israel et al. 2005; Strohmayer & Watts 2006; Watts & Strohmayer 2006; Huppenkothen et al. 2014b; Miller et al. 2019). Parts of the time series might only contribute low levels of noise and thereby increase the number of bins without suppressing the QPO in the periodogram. These parts, therefore, increase the significance of the QPO without having delivered any actual information about it. In some circumstances, spectrograms may provide a better insight into the exact behavior of the time series, but this comes with its limitations in many cases, as we lay out in Appendix C.

We now turn toward a more mathematical explanation of the nonstationarity bias. Let us consider how we can calculate the signal-to-noise ratio (S/N) of a QPO when red noise and white noise are present.[7] Given a single frequency bin f_j, this is

$$\rho(f_j) = \frac{S^{\mathrm{Q}}(f_j)}{S^{\mathrm{RW}}(f_j)}, \tag{12}$$

where we denote the S/N as ρ. Assuming that all frequency bins in a periodogram are statistically independent, the S/N adds in quadrature

$$\rho_{\mathrm{tot}} = \sqrt{\sum_{j=1}^{N/2} \rho(f_j)^2}\,, \tag{13}$$

where we have summed over the $N/2$ bins in the periodogram and ρ_{tot} denotes the total S/N of the QPO.

3.1. QPOs in Stationary Noise

We begin by considering stationary noise, as this most closely matches the requirements for the Whittle likelihood to be valid. We show that even in this limit a nonstationary QPO can create biased estimates.

As the simplest scenario, we consider a stationary time series with N elements made up of a QPO and some arbitrary noise. We denote the S/N of this time series as ρ^N_{tot}. Suppose we look at an extended version of this time series with xN elements, which we create with the same QPO and noise process. We intuitively expect the S/N to increase proportionally to $\sqrt{x}$, and indeed we calculate

$$\begin{aligned}\rho^{xN}_{\mathrm{tot}} &= \sqrt{\sum_{j=1}^{xN/2} \rho^x(f_j)^2} \\ &= \sqrt{x\sum_{j=1}^{N/2} \rho(f_j)^2} \\ &= \sqrt{x}\,\rho^N_{\mathrm{tot}},\end{aligned} \tag{14}$$

where we write the S/N at a given frequency in the extended time series as $\rho^x(f_j)$, which is the same as in the original time series. We also use the N and xN superscripts for the S/Ns to indicate whether we are looking at the original or extended time series. Note that we have implicitly changed how the f_j are indexed between the steps. Additionally, we assume in the first step that the PSD is only slowly varying between frequency bins, and we can therefore approximate the newly added frequency bins by their closest neighbors from the original PSD.

Let us now consider what happens if we extend the time series with zeros on either end, which is illustrated in Figure 1. This scenario is not truly physical, as we always expect some background noise, but it is an instructive example to consider.

[7] The argument here holds for arbitrary noise.

It is also a limit of many practical observations such as magnetar flares or prompt GRB emissions that are surrounded by low levels of photon-counting noise. Extending with zeros implies both $S^{\mathrm{RWQ}} \to S^{\mathrm{RWQ}}/x$ and $S^{\mathrm{RW}} \to S^{\mathrm{RW}}/x$, so $\rho^x(f) = \rho(f)$, and thus, following the same steps as before,

$$\rho^{xN}_{\mathrm{tot}} = \sqrt{x}\,\rho^N_{\mathrm{tot}}, \tag{15}$$

which is identical to Equation (14). This is a curious result, as extending a time series by zeros clearly should not increase the S/N or the significance of a signal. The resolution to this seeming paradox is that it is invalid to assume that S/N adds in quadrature as we did in Equation (14). By extending the time series with zeros, we added bins to the periodogram that are not statistically independent. A practical example of this is X-ray data from solar flares in the 1–8 Å band, which has almost no white-noise component owing to the very low measurement uncertainties and mostly low-frequency components that can occur below the QPO frequency. We discuss this in Section 5.

Alternatively, let us consider what happens when we consider a nonstationary extended time series in which the noise is present over xN elements but the QPO is only present for N elements. In that case, S^{RW} is the same as in the original time series, whereas S^{Q} gets reduced to S^{Q}/x as we decrease the average QPO power in the time series, similarly to how the periodograms decrease in amplitude with increasing x in the bottom panel of Figure 1. We calculate Equation (13) for the extended time series again,

$$\begin{aligned}\rho^{xN}_{\mathrm{tot}} &= \sqrt{\sum_{j=1}^{xN/2} \rho^x(f_j)^2} \\ &= \sqrt{x\sum_{j=1}^{N/2} \rho^x(f_j)^2} \\ &= \sqrt{x\sum_{j=1}^{N/2} \rho(f_j)^2/x^2} \\ &= \frac{1}{\sqrt{x}}\rho^N_{\mathrm{tot}},\end{aligned} \tag{16}$$

where we used the same first step as in Equation (14) and used the fact that the QPO is suppressed by the extension in the second step. Note that the S/N decreases with $1/\sqrt{x}$ instead of $1/x$ as we might expect intuitively since the nonstationarity effect still boosts the S/N with $\sqrt{x}$.

A further consequence is that the product in the Whittle likelihood becomes invalid. Due to the extended segment length, both the PSD and the periodogram need to be divided by a factor of x. Therefore, we can rewrite the likelihood as

$$\begin{aligned}L^{xN}(I|S(f|\theta)) &= \prod_{j=1}^{xN/2} \frac{x}{S_j(\theta)}\exp(-I_j/S_j(\theta)) \\ &\approx x^{xN/2}\left(\prod_{j=1}^{N/2} \frac{1}{S_j(\theta)}\exp(-I_j/S_j(\theta))\right)^x,\end{aligned} \tag{17}$$

where θ is the set of parameters in the PSD and we use the N and xN superscripts on the likelihoods to indicate whether we are looking at the original or extended time series. Again, we assume that the PSD is slowly varying with frequency and thus statistically nonindependent bins are reasonably approximated by the closest frequency bins obtained from the stationary time series. The factor $x^{xN/2}$ that we have introduced does not matter since it is a constant. Hence, we can write

$$\ln L^{xN}(I|S(\theta)) \propto x \ln L^N(I|S(\theta)). \tag{18}$$

This means $\ln L^{xN}(I|S(\theta)) \approx x \ln L^N(I|S(\theta))$, i.e., we have steepened the log likelihood function by a factor of x,

$$\frac{\partial \ln L^{xN}(I|S(\theta))}{\partial\theta} \approx x\frac{\partial \ln L^N(I|S(\theta))}{\partial\theta}. \tag{19}$$

This steepening means that inferred posterior distributions are generally tighter, which leads to overconstrained parameter estimates. We calculate how this changes the BIC:

$$\begin{aligned}\Delta\mathrm{BIC}^{xN} &= (k_{\mathrm{RWQ}} - k_{\mathrm{RW}})\ln(xN/2) \\ &\quad - 2x\ln\frac{L^N_{\max}(I\mid S^{\mathrm{Q}})}{L^N_{\max}(I\mid S^{\mathrm{RW}})} \\ &= x\Delta\mathrm{BIC} \\ &\quad + (k_{\mathrm{RWQ}} - k_{\mathrm{RW}})(\ln x - (x-1)\ln(N/2)),\end{aligned} \tag{20}$$

where $k_{\mathrm{RWQ}} = 6$ and $k_{\mathrm{RW}} = 3$ are the number of free parameters for the respective models and xN is used as a superscript. Thus, ΔBIC is approximately proportional to x or N since the impact of the $\ln x$ and constant terms is minimal. The impact this has on the Bayes factor is not straightforward to calculate in the general case. Since likelihood ratios and ΔBICs can be understood as a related quantity of the ln$B$$F$, we expect a similar approximately linear behavior, which we find empirically in Section 4.

3.2. QPOs in Nonstationary Red Noise

Stationary noise assumptions often do not hold in reality. Instead, red noise and QPOs often occur simultaneously during a transient and are preceded and followed by white noise.

Assume a stationary time series with red noise, white noise, and a QPO that sits in the red-noise-dominated part of the PSD. If we extend this time series with zeros, the argument in Section 3.1 holds and the S/N calculated with Equation (13) grows proportional to $\sqrt{x}$. Extending with white noise instead of zeros will also lead to the S/N growing with $\sqrt{x}$ as long as the QPO remains in the red-noise-dominated part of the PSD. To show this, we define f_{break} as the breaking frequency where white noise and red noise are of equal magnitude,

$$f_{\mathrm{break}} = \left(\frac{C}{A}\right)^{-1/\alpha}, \tag{21}$$

which means that for $f \ll f_{\mathrm{break}}$ red noise dominates, whereas white noise dominates for $f \gg f_{\mathrm{break}}$. If our QPO has $f_0 < f_{\mathrm{break}}$, it mainly competes with red noise in terms of S/N; thus, a rising level of white noise will not meaningfully reduce the S/N until $f_0 \approx f_{\mathrm{break}}$. We can write this as

$$\rho_{\mathrm{tot}} \approx \rho_{\mathrm{tot}}(f < f_{\mathrm{break}}), \tag{22}$$

where $\rho_{\mathrm{tot}}(f < f_{\mathrm{break}})$ is the total S/N just based on frequencies less than the breaking frequency. If we extend the time series with white noise or zeros by a factor of x, we suppress the red-noise term by $1/x$ and shift the breaking frequency

$$f^x_{\mathrm{break}} = \left(\frac{Cx}{A}\right)^{-1/\alpha} = x^{-1/\alpha} f_{\mathrm{break}}, \tag{23}$$

where the superscript x indicates that we are looking at the extended time series. We can equate Equation (23) to f_0 and rearrange to calculate the extension factor $x_{\rm break}$ for which the QPO will be in the white-noise-dominated part of the periodogram,

$$x_{\rm break} = \left(\frac{f_{\rm break}}{f_0}\right)^{\alpha} = \frac{A}{C} f_0^{-\alpha}. \tag{24}$$

Therefore, we are prone to artificially increase the S/N by extending our time series with white noise up to an extension factor of $x_{\rm break}$. Conversely, if our QPO mainly competes with white noise $f_0 > f_{\rm break}$, we use the argument in Section 3.1 to show that we suppress the S/N with increasing x,

$$\begin{aligned}\rho_{{\rm tot},x} &= \sqrt{\sum_{j=1}^{xN/2}\left(\frac{S^{\rm Q}(f_j)/x}{S^{\rm W}}\right)^2} \\ &\approx \frac{1}{\sqrt{x}}\sqrt{\sum_{j=1}^{N/2}\left(\frac{S^{\rm Q}(f_j)}{S^{\rm W}}\right)^2} \\ &\approx \frac{\rho_{\rm tot}}{\sqrt{x}},\end{aligned} \tag{25}$$

where we have used the fact that the PSD is slowly varying with frequency in the first approximation, and that the frequencies where the QPO exceeds the noise continuum are all dominated by white noise in the second approximation.

The most obvious case where this could become a practical problem is when we have to decide where to start and end the segment. Not being aware of this bias may lead us to overselect the data by $\mathcal{O}(10\%)$ and infer a somewhat higher S/N and significance, which by itself is not a big issue. But the issue becomes much worse in practice if the QPO itself is nonstationary and only appears for a part of the transient, which implies that selecting the entire transient may already be overextending the segment in which the QPO is present. Time–frequency representations, such as spectrograms, can only in some circumstances be helpful owing to limitations on their resolution (we discuss this in more detail in Appendix C).

In principle, extending the time series with red noise instead of white noise should not cause the same issue, as we would expect the same suppression by $1/\sqrt{x}$ that we have for white noise in Section 3.1. However, in real transients, the low-frequency noise continuum arises not because of a stationary noise process but because of a combination of deterministic and nonstationary stochastic processes. Different parts of the time series might add up roughly to a single power law or a broken power law, but smaller segments within the transient may have vastly different shapes. For example, the rising and falling edge of the transient may create different power laws, we may have segments that are fairly flat and thus mostly add white noise, and so on. Additionally, deterministic aspects of the time series, if not properly subtracted, are not appropriately modeled with a Whittle likelihood as we show in Section 4.3. The deterministic parts of the light curve can not be safely removed by filtering or smoothing methods, as they are prone to create an artificial structure that looks like oscillatory behavior (Auchère et al. 2016). Clearly, it is not possible to write down a rigorous treatment of this rather large class of possible constellations in which nonstationarity bias could occur.

We note here that using the outlier analysis is not prone to false detections due to the nonstationarity bias since we are testing neighboring bins individually instead of combining S/N from neighboring bins. The opposite is the case: since we add more bins by having extended the time series, it is harder to exceed the Bonferroni-corrected significance level. However, in practice, periodograms are sometimes rebinned to add up neighboring bins (e.g., Huppenkothen et al. 2014a). The motivation behind this is to combine the S/N from neighboring bins, making it easier to detect QPOs. Thus, rebinned outlier analyses are prone to the same nonstationarity bias that affects the Bayesian methods as well. The bins in the rebinned periodogram are distributed as the sum of exponentially distributed random variables, which is a χ^2-distribution with $2k$ degrees of freedom,

$$p(f;\, k,\, 1/S) = \frac{1}{\Gamma(k) S^k} f^{k-2} e^{-f/S}, \tag{26}$$

where Γ is the Gamma function, k is how many bins are rebinned into a single one, and S is the power spectral density at frequency f, which is assumed to be approximately constant.

4. Simulated Data

To demonstrate the impact of nonstationarities empirically, we analyze simulated data using a variety of setups. For all studies on simulated data, we assume white-noise levels to be constant along with the time series. We discuss the impact of the Poissonian nature of the photon-counting process in Appendix B.

4.1. Setup

To produce the time domain data, we use the algorithm by Timmer & König (1995), as it is implemented in the BILBY software package. Concretely, this works by creating a frequency series by randomly drawing both real and imaginary parts from a Gaussian distribution, coloring the noise by multiplying it with the square root of the PSD, and then applying an inverse Fourier transform to obtain a stationary time series. In general, we create a continuous, nonstationary time series by using a combination of addition, concatenation, and convolution with window functions. For example, we start by creating a stationary time series using the QPO model and red noise $S^{\rm RQ} = S^{\rm R} + S^{\rm Q}$. We apply a Hann window to ensure a smooth turn-on from zero. This helps us to avoid Fourier artifacts such as sidelobes in the periodogram that can appear if discontinuities are present. Next, we create a much longer time series containing just white noise using the Timmer & König (1995) method, and we add the $S^{\rm RQ}$ time series in the center. We are also interested in what happens when we extend the time series with zeros instead of white noise. To do this, we mask the parts containing just white noise and set those values to zero. In all cases, we center the part of the time series in which the QPO is present. This is to prevent the Hann window, which is applied when calculating the periodogram, from diminishing the amplitude of the red-noise and QPO components. Note that this is separate from the Hann window we use to ensure the smooth turn-on from zero.

For all simulated data in this section, we assume a constant white-noise level throughout the time series and place the QPO in the center of the time series between −10 s and 10 s. We use a sampling frequency of 40 Hz and extend the time series up to 200 s ($x = 10$) or 400 s ($x = 20$) depending on the scenario. The

specific parameters for each simulation are listed in the tables in Appendix A. We discuss the effect of nonstationary white noise in Appendix B. All parameters and priors are listed in Tables 2 and 3.

4.2. Nonstationary QPO in White Noise

We start by considering a combination of a QPO and white noise where the setup is effectively identical to what we describe in Section 4.1, but we use S^{Q} instead of S^{RQ} to omit red noise. We consider both scenarios discussed in Section 3.1, where we extend the time series either with zeros or with white noise. In terms of the parameter inference, we deviate from the standard practice of using $S^{\mathrm{RWQ}}(f)$, as there is no red-noise component. Instead, we compare S^{WQ} against the white-noise hypothesis $S^{\mathrm{W}} = C$.

First, we consider what happens when we extend the time series with white noise. In principle, the results should follow our discussion in Section 3.1 for QPOs in the white-noise-dominated part of the periodogram, i.e., the S/N should decrease with $1/\sqrt{x}$ and the lnBF should decrease roughly inversely to x. The results, which we display in the top panel of Figure 2, are in agreement with our expectations.

We now consider the case where we extend the time series with zeros, similar to what we did to produce Figure 1. As we discussed in Section 3, we expect the S/N to increase with $\sqrt{x}$ and the ln BF roughly linearly. As we show in the bottom panel of Figure 2, this is almost perfectly the case for the ln BF and is also qualitatively true for the S/N.

4.3. Nonstationary QPO in a Simple Transient

We investigate the effect of analyzing a nonstationary QPO that appears on top of a deterministic transient flare shape in the presence of white noise. The time series is displayed in the top panel of Figure 3. The flare shape mimics a red-noise continuum that overlaps with the QPO in the periodogram (bottom panel of Figure 3), though it is not well described by a Whittle likelihood in the periodogram since it is not a stochastic process. We see that there is conspicuously little fluctuation at low frequencies in Figure 3. For the flare shape we use a continuous exponential rise and fall described by

$$f(t) = A_{\mathrm{flare}} \begin{cases} \exp\left(\dfrac{t - t_0}{\tau_{\mathrm{r}}}\right) & \text{if } t < t_0 \\ \exp\left(-\dfrac{t - t_0}{\tau_{\mathrm{f}}}\right) & \text{if } t \geqslant t_0 \end{cases}, \tag{27}$$

where A_{flare} is the amplitude, t_0 is the peak time, and τ_{r} and τ_{f} are the rise and fall timescales, respectively. This model is similar to the FRED model, which is popular to fit time series of gamma-ray bursts (Norris et al. 2005). The advantage of using an exponential shape for this simulation is that it produces a power law in the periodogram that matches our red-noise model. Since the flare is deterministic, low frequencies in the periodogram are not well described by a Whittle likelihood. To create this time series, we repeat the steps in Section 4.2, though we start with different parameters listed in Table 3 and add the deterministic flare shape in the last step.

We perform the same test as in the previous section of extending the time series starting from the 20 s segment in which the QPO is present (inset in the top panel of Figure 3). It is less intuitive why nonstationarity bias may occur and how it will manifest in this scenario. Extending beyond the peak of the flare might reduce the apparent significance, as it will provide the strongest contribution to the low-frequency noise continuum. At the same time, we extend toward the tail of the flare, which contributes little more than white noise. As we can see in Figure 4, the S/N and ln BF do indeed increase sharply until $x = 3$, which corresponds to the point where we extend the selected window to the peak of the transient. Past this point low-frequency contributions from the rising edge of the transient obscure the QPO in the periodogram and quickly lead to the QPO becoming undetectable for $x > 7$. We note that using a different set of parameters to create the data can easily create situations where the ln BF continues to increase past $x = 3$.

4.4. Nonstationary QPO in Nonstationary Red Noise

In this scenario, we create a time series using a combination of red noise, white noise, and a QPO using the S^{RWQ} model. We extend this time series with either white noise or zeros. Figure 5 displays the time series for $|t| < 30$ s (top panel) and the periodogram as obtained for $|t| < 10$ s.

In the case of extending with zeros, we expect the same relation as in Section 4.2, which we recover in the bottom panel of Figure 6. On the other hand, when we extend with white noise, the S/N first rises quickly until the extension factor reaches around $x_{\mathrm{break}} = 2$ (see Equation (24)) and then turns toward a slow descent proportional to $1/\sqrt{x}$. We note that the behavior in the case of $x < x_{\mathrm{break}}$ for white noise is not well captured, but greater extension factors would be less realistic and the simulated data set aims to qualitatively show all effects.

We use models S^{RW} and S^{RWQ} as described earlier to calculate posterior distributions, Bayes factors, and ΔBICs. As we show in the bottom panel of Figure 6, the ln BF increases almost perfectly linearly when we extend with zeros. In the top panel, we see that if we extend with white noise, the ln BF also increases with x even past x_{break}, though it turns around eventually. The ln BF may be increasing past x_{break} if the QPO feature in the periodogram is better fit by the model for higher x compared to the feature's shape at x_{break}.

We have shown that overall we can see a nonstationarity bias if white noise remains stationary, but red noise vanishes for part of the light curve. This is in principle different from the effect in Section 4.3, as the red noise is due to a stochastic rather than a deterministic process.

4.5. Tests of Stationarity

One noticeable property of the bias we have found is that it passes regular tests employed to detect fitting validity. Nita et al. (2014) demonstrate a χ^2-like test for exponentially distributed data

$$\chi^2_\nu = \frac{1}{\nu} \sum_{j=1}^{N/2} \left(1 - \frac{I_j}{\hat{S}_j}\right)^2, \tag{28}$$

where ν are the number of degrees freedom and $\hat{S}$ is the maximum likelihood PSD. Analogously to the regular χ^2-test, $\chi^2_\nu \approx 1$ indicates an appropriate fit to the data, whereas $\chi^2_\nu > 1$ indicates underfitting and $\chi^2_\nu < 1$ overfitting. Poor fits should in general occur with nonstationary behavior in the red-noise- or QPO-dominated part of the PSD. That is because neighboring bins will

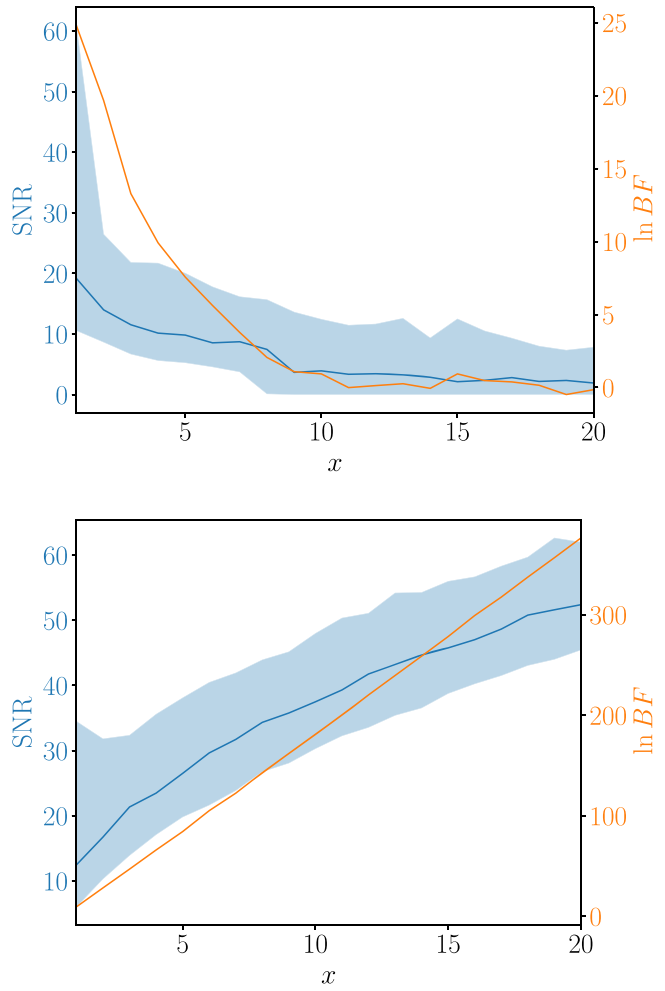

Figure 2. S/N (blue) and lnBF (orange) vs. the extension factor for a simulated QPO plus white-noise signal that is extended with white noise (top) or zeros (bottom). The solid blue line displays the S/N of the maximum likelihood point, and the shaded blue region displays the S/N's 90% credible interval. In the top panel both S/N and lnBF decrease with increasing x, which corresponds to the QPO vanishing in noise. In the bottom panel the lnBF increases almost perfectly linearly owing to the increased number of frequency bins that make up the QPO, whereas the S/N increases roughly with the square root.

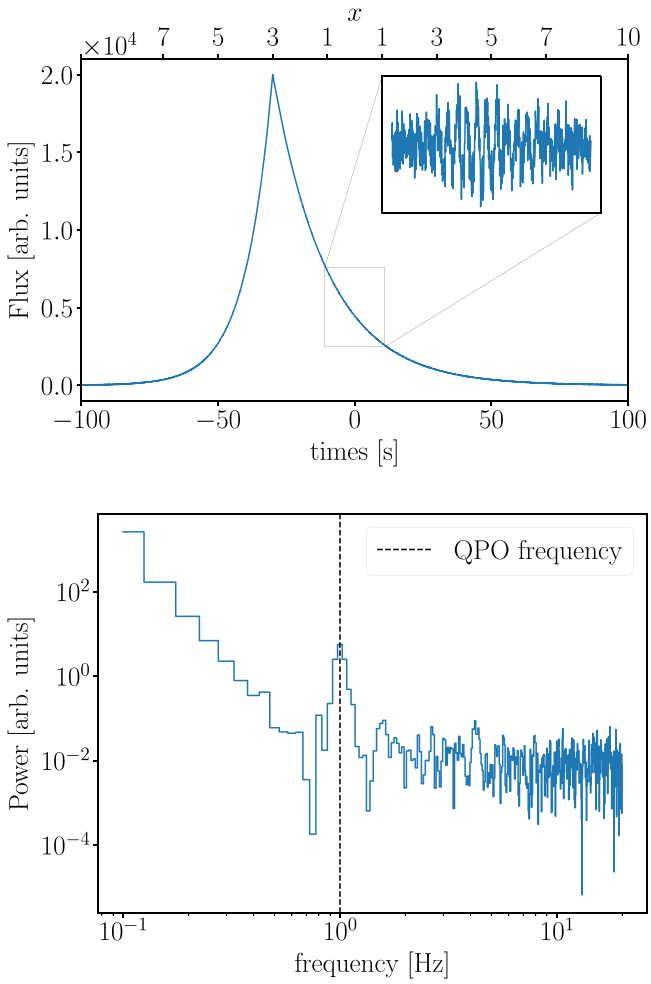

Figure 3. Simulated data of a nonstationary QPO in a deterministic transient flare shape in the presence of white noise. The time series (top) shows the dominant transient and the QPO that is present for 20 s on the tail of the flare (inset). We remove the deterministic trend of the flare in the inset to make the QPO easily visible. The periodogram (bottom) shows a visible QPO at 1 Hz and displays a low-frequency continuum that arises as a result of the transient flare shape. The periodogram corresponds only to the segment of the light curve that contains the QPO. Note that there is much less variability than we would expect from the χ^2_2-distribution for the bins at low frequencies since the noise continuum is due not to a stationary stochastic process but to a deterministic process.

be covariant and a fit may align either more closely or farther away than should be possible for independent bins, though $\chi^2 \approx 1$ may still randomly occur. However, in practice, nonstationarities will affect the low-frequency part of the PSD, while high frequencies are dominated by white noise. Since frequency bins are linearly spaced, the vast majority of all bins are almost always fitted well in terms of the χ^2-statistic if the periodogram reaches a white-noise floor at high frequencies. As we show in Figure 7, the χ^2-value for the entire PSD is very close to the χ^2-value for all the frequencies above the QPO frequency plus twice its width. For comparison, setting $f_0 = 1$ Hz with a 40 Hz sampling frequency, as is the case in the bottom panel of Figure 7, about 95% off all frequency bins are greater than f_0. Thus, even if we overfit locally around the QPO, it is unlikely that this greatly impacts the overall χ^2-value.

We propose a local χ^2-test, where we calculate the χ^2-value for just the frequencies within 2σ around the QPO frequency. The 2σ width provides a reasonable trade-off between having a range in which noise is much smaller than the QPO and covering enough bins to be able to perform the test. We show this test in Figure 7 for data sets used in Sections 4.2 (top) and 4.4 (bottom). We note in the top panel that we obtain a reasonable $\chi^2 \approx 0.92$ for the local test if we just look at the stationary time series at $x = 1$ and obtain $\chi^2 < 0.6$ for large x. This method has the downside that it will not work if there are only very few bins available as is the case for $x \lesssim 6$ in the bottom panel of Figure 7. We need at least seven data points to compensate for the six degrees of freedom in the S^{RWQ} model. If more bins are available, this test can flag whether the QPO has been overfitted or underfitted. However, in some cases we may find $\chi^2 \approx 1$ even for QPOs that are strongly affected by nonstationarity bias. For example, a time series may in principle contain two QPOs at similar frequencies and nonstationarity bias. In that instance, the overfitting due to covariance of neighboring bins may be compensated by underfitting of the underlying multiple QPO shapes and yield $\chi^2 \approx 1$ for the local test.

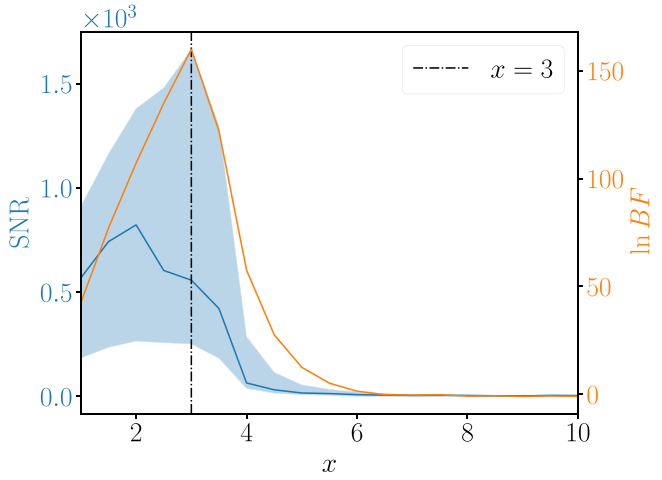

Figure 4. S/N (blue) and ln BF (orange) vs. the extension factor for the QPO shown in Figure 3. The solid blue line displays the S/N of the maximum likelihood point, and the shaded blue region displays the S/N's 90% credible interval. Both the ln BF and the upper edge of the S/N's credible interval peak at $x = 3$ (black dashed vertical line), which corresponds to the point where the selected data segment exceeds past the peak of the flare. The strong differences in ln BF and S/N across extension factors shows that differences in data selection can greatly influence the results.

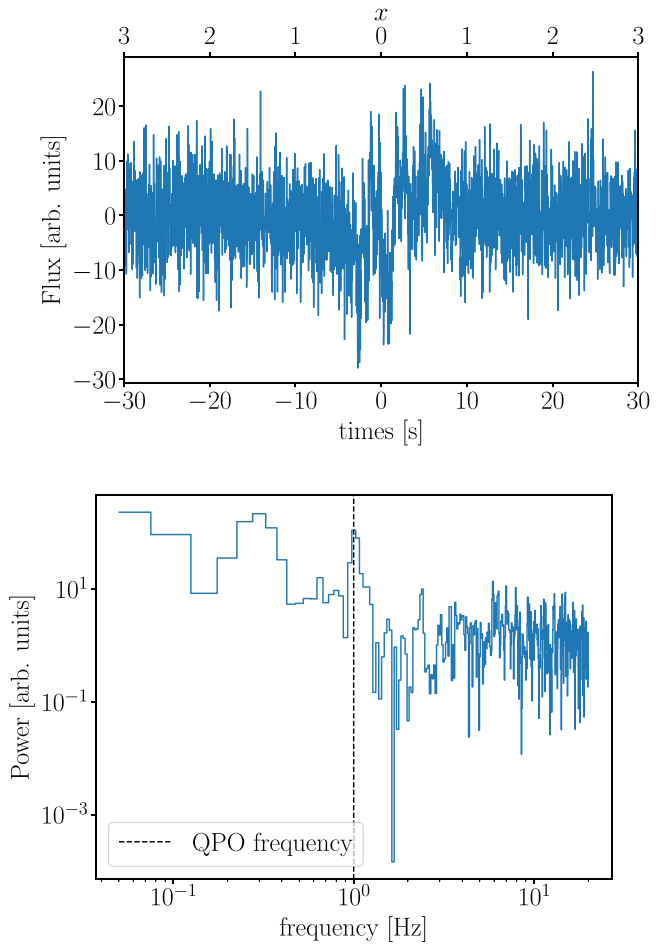

Figure 5. Simulated data of a nonstationary QPO and red noise in surrounding stationary white noise. In the time series (top) we show how a QPO and red noise are transitioning smoothly into the surrounding part of the time series that is pure white noise ($|t| > 10$ s) owing to the Hann window that we have applied. In the periodogram (bottom) for $|t| \leqslant 10$ s the QPO is clearly visible by eye. The Hann window also produces an additional broad feature at ≈ 0.3 Hz in the periodogram.

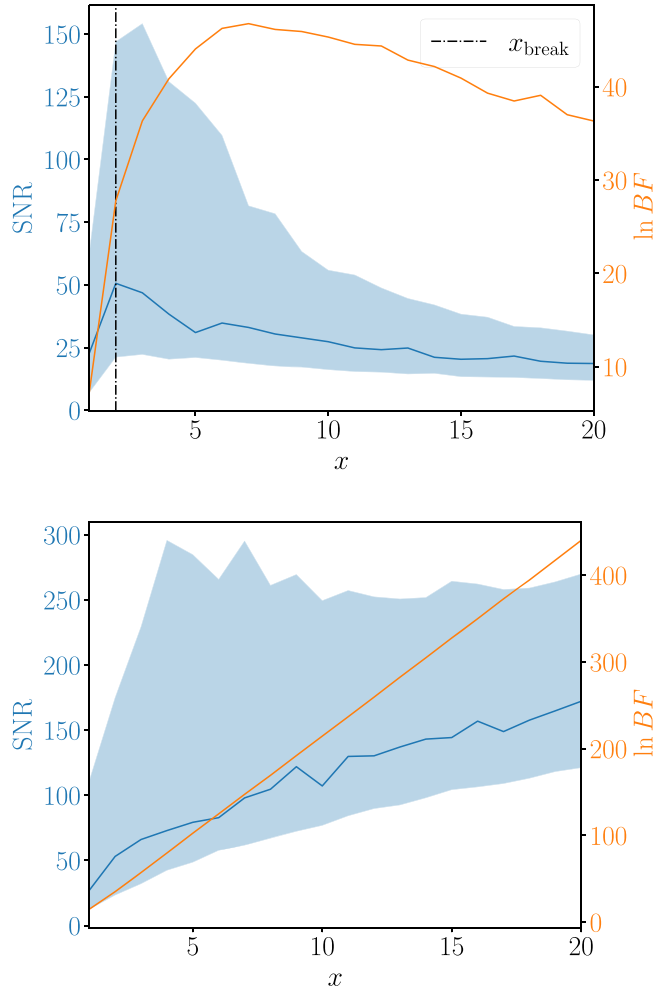

Figure 6. S/N (blue) and ln BF (orange) vs. the extension factor for simulated data of a nonstationary QPO and red noise with white noise as described in Section 4.4 for extension with white noise (top) and zeros (bottom). The solid blue line displays the S/N of the maximum likelihood point, and the shaded blue region displays the S/N's 90% credible interval.

5. Solar Flare Data

QPOs are a regularly reported phenomenon that have been observed in solar flares over a wide range of wavelengths for decades, from radio waves to EUV, as well as soft and hard X-rays (see Nakariakov & Melnikov 2009; Van Doorsselaere et al. 2016, for recent reviews). In this domain, they are usually referred to as quasiperiodic pulsations. Due to their ubiquity, progress has been undertaken to automate their analysis using Whittle likelihoods, e.g., with the Automated Flare Inference of Oscillations (AFINO) method (Inglis et al. 2015, 2016; Hayes et al. 2020). An extensive catalog of AFINO analyses is available online.[8] AFINO defines models similar to S^{RW}, S^{RWQ}, and S^{BW}, although it uses a Gaussian instead of a Lorentzian pulse as the QPO component and a slightly different broken power-law model. AFINO uses a computationally efficient SCIPY fitting routine to determine the maximum likelihood values and calculate the ΔBIC with Equation (8), though it does not calculate Bayes factors or posterior distributions. The QPO model is "strongly favored" if $\Delta\mathrm{BIC} < -10$, relative to both the red noise and broken power-law model. This ensures a

[8] https://aringlis.github.io/AFINO/

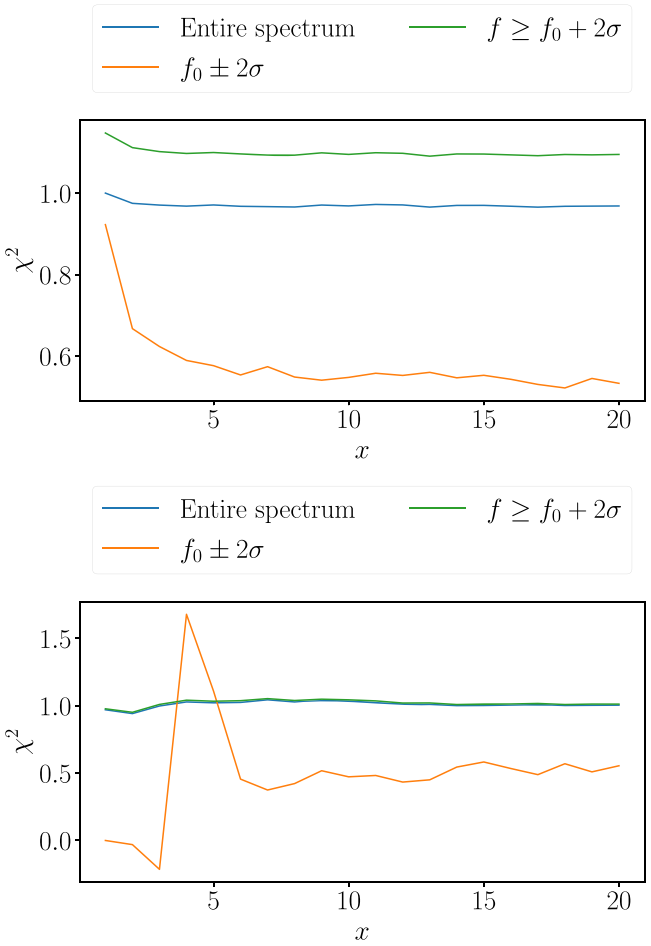

Figure 7. Nita et al. (2014) χ^2-tests in the case of a QPO and white noise extended with zeros as described in Section 4.2 (top), and the case of QPO, red noise, and white noise extended with white noise as described in Section 4.4 (bottom). We see in the top panel that the overall fit (blue) and the fit for frequencies above the QPO frequency (green) have $\chi^2 \approx 1$, indicating a good fit. The local χ^2-test around the QPO frequency f_0 indicates a good fit only for $x = 1$ and falls much below that for $x > 1$, indicating overfitting. This implies that the local χ^2-test can detect nonstationarity bias in this instance. In the bottom panel, we display some of the limitations of the local χ^2-test when we deal with narrow QPOs. For $x \lesssim 6$ the local χ^2-test (orange) around f_0 swings widely, as we have too few frequency bins within the range. Above $x \approx 6$, enough frequency bins accumulate within the range, and we detect the overfitting. We also note that the χ^2-test for the entire PSD (blue) and frequencies above the QPO frequency (green) indicate an overall good fit despite us overfitting the QPO.

small number of false-positive values, as has been demonstrated on simulated data in Broomhall et al. (2019). Additionally, AFINO performs the χ^2-test for exponentially distributed data using Equation (28). AFINO flags the results if the implied p-value from the χ^2_ν test is below or above a set threshold.

For a comparative study on the impact of nonstationarity bias, it is instructive to consider some of the most significant QPOs. This is because these QPOs are visually identifiable from the light curve, which guides us in their analysis. We look at a solar flare observed by the X-Ray Sensor (XRS) on board the GOES-15 satellite on 2013 May 12, at 20:17 UT, for which AFINO reports the highest significance for a QPO with $\Delta\mathrm{BIC} = -451.6$ relative to red noise and $\Delta\mathrm{BIC} = -278.2$ relative to the broken power law. This flare was of GOES magnitude M1.9 and originated from NOAA Active Region

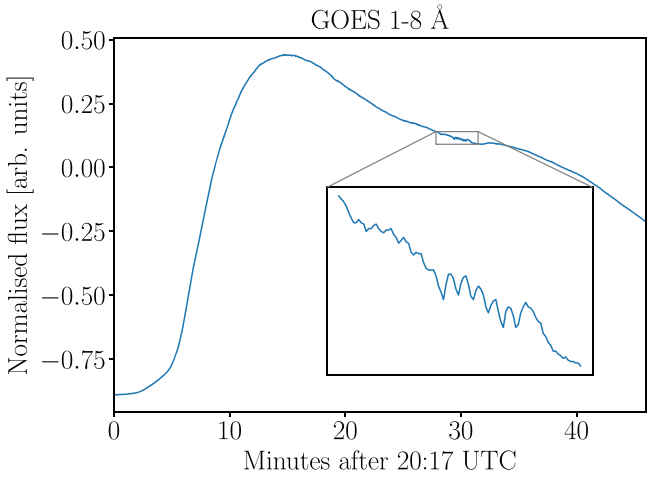

Figure 8. Solar flare X-ray light curve as observed with GOES in the 1–8 Å band. The figure shows the time selected as done with AFINO. The inset zooms in on the tail of the light curve, where the QPO is clearly visible. The left and right edges of the small inset box demarcate the three segments noted in Table 1.

11748, which was located on the east limb of the Sun at the time. AFINO finds a high-quality fit for the QPO model with $\chi^2 = 1.06$ but flags the fits with the broken power law and red noise owing to their high χ^2 values, likely due to the very pronounced QPO present at $P = 12.6$ s. We show the 46-minute-long GOES X-ray light curve we analyze in Figure 8. The QPO is visible by eye on the tail end of the distribution (inset) and persists for about 200 s, much shorter than the 2760 s data segment. This particular QPO may not be of solar origin, since it is not observed in other flare-observing instruments. However, a similar signature is coincidentally seen in the GOES magnetometer data, suggesting a possible artifact. Regardless of the origin, the QPO is clearly present and confined to a small portion of the data, providing an ideal test case.

To demonstrate the presence of an artificial amplification of the S/N, we split the light curve into three segments. The first segment covers the time before the QPO (0–1680 s) (after 20:17 UT), the second selects just the QPO (1680–1880 s), and the third segment selects everything after the QPO (1880–2760 s). We perform Bayesian inference separately on each of these segments and the combined segment independently with the models S^{RW} and S^{RWQ}. We use the same priors as in the studies on simulated data that we list in Table 2. If the light curve were stationary, it would be valid to combine the Bayes factors and f_0 posteriors, as they represent independent draws from the same distribution. However, as we show in Table 1, we are unable to detect the reported QPO in the first and third segments. There is some weak evidence toward a QPO in the first segment, though not at the reported 12.6 s period. The third segment shows some weak support for a QPO with $\ln \mathrm{BF} = 1.4$, but the maximum likelihood fit indicates that this is rather due to a broad feature in the periodogram that is better fitted with a broken power law. As we show in Table 1, both the first and third segments are better fitted with S^{BW} than with S^{RWQ}, which indicates that there is likely no QPO present. In the second segment we find the QPO independently with a very high significance $\ln \mathrm{BF} = 27.5$, and it is clearly visible in the fitted periodogram in Figure 9. Finally, we analyze the combined segments together and find the QPO with $\ln \mathrm{BF} = 229.4$, a significance much higher than in segment 2. We note that we find the QPO at a slightly different period of

Table 1
Results from Analyzing the Selected Segments of the Solar Flare Detected by GOES

Segment	ln BF S^{RWQ}	ln BF S^{BW}	ΔBIC S^{RWQ}	ΔBIC S^{BW}
0–1680 s	3.0 ± 0.3	6.8 ± 0.3	-11.3	-5.0
1680–1880 s	27.5 ± 0.2	-0.7 ± 0.2	-66.8	7.8
1880–2760 s	1.4 ± 0.3	7.7 ± 0.2	-0.7	-6.8
0–2760 s	229.4 ± 0.3	93.0 ± 0.3	-465.6	-175.8

Note. All errors are given based on 1σ confidence or credible intervals. The segments are given in seconds after 20:17 UT. All ln BF and ΔBIC values are calculated relative to S^{RW}. We reiterate that either a positive ln BF or a negative ΔBIC indicates that S^{RWQ} or S^{BW} is preferred over S^{RW}. We find broadly that the ΔBIC values in the last row are in agreement with what AFINO reported. Deviations are likely due to differences in our model and slightly different data selection. While in all instances Bayes factors and ΔBICs give the same indication about the preferred model, the significance differs to some extent. This may be in part due to our wide prior choices.

$12.34\,\mathrm{s} \pm 0.06\,\mathrm{s}$ compared to the value reported by AFINO. This may be due to slightly different QPO modeling choices between AFINO and our method.

We perform the local χ^2-test we introduced in Section 4.5 for the entire light curve and find $\chi^2 = 0.29$ for the frequency bins surrounding the QPO. This indicates that the QPO has been overfitted and is nonstationary.

It is evident that the solar flare segment we have analyzed contains a QPO and an overestimation of its significance is not as critical as it would be for a marginal detection. Trying to find an instance where a marginal detection was turned into a very confident detection due to nonstationarity bias would be much harder. We would not be able to determine the location of the QPO in the light curve by eye. Instead, we would have to take a systematic approach and split the light curve into several smaller segments and determine whether the QPO exists in them individually. Of course, the significance in the individual segments would always be lower, so it would be hard to determine whether this is due to the reduced nonstationarity bias or because we split a persistent QPO into multiple segments. As mentioned previously, we do not expect AFINO to report many false detections, as has been established in Broomhall et al. (2019). This is due to its conservative detection threshold $\Delta\mathrm{BIC} < -10$. Thus, if AFINO is coupled with a more careful data selection method, it could be possible to find more QPOs.

The nonstationarity bias in solar flare data arises most likely as a combination of the effects we describe in Section 3, i.e., a nonstationary QPO, nonstationary noise processes, and a deterministic overall flare shape all contribute to some extent. We cannot easily discern what the impact of each of these effects is.

Nonautomated analyses of QPOs in solar flares are also likely less affected than the stretch of GOES data we analyze in this section. Typically, parts of the light curve that are of interest are selected manually for detailed analyses (e.g., Hayes et al. 2019), though this by no means provides a guarantee that such results are without bias. As we have established previously, the ln BF or the ΔBIC can grow linearly with x; thus, even a mild overselection of the light curve can lead to erroneous detections and inferences.

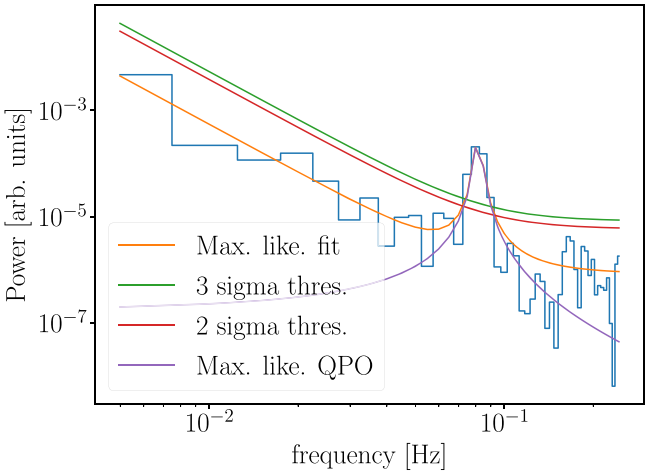

Figure 9. Maximum likelihood fit of the periodogram just for the second segment of the GOES light curve. The QPO is clearly visible and highly significant. We draw the 2σ and 3σ detection limits based on the Bonferroni-corrected frequentist statistics. We note that the QPO is located in the red-noise-dominated part of the PSD.

6. Discussion and Outlook

In this paper, we show that analyses based on Whittle likelihoods likely overstate the significance of QPOs in light curves if the QPO or noise features in time series are nonstationary. Specifically, this is the case if data are included in the time series that contains only white noise when the QPO is in the red-noise-dominated part of the PSD. The effect can also occur if the QPO is in the white-noise-dominated part of the PSD but the white noise is nonstationary (e.g., if the noise arises as a result of photon counting) and we include weak white noise from parts of the light curve that have low photon-counting rates. Selecting an appropriate time segment for the analysis of transients is thus important but remains difficult to do generically and without bias. A scientist's natural intuition may be to select data conservatively, i.e., select a longer time series while being aware that the QPO may only be present for a part of it. The erroneous reasoning may be that by selecting a longer segment they are not "cherry-picking" the segment with the most pronounced QPO features. As we demonstrate in this paper, this choice may increase the significance of the QPO artificially and lead to false detections due to the nonstationarity bias.

There are some possible mitigation strategies that one may apply when the use of periodograms is still preferred. However, this means that it is a lot harder to devise a generic process with which to analyze time series.

First, if we suspect that the QPO is indeed only present for part of the transient and the white-noise level is relatively low compared to the low-frequency continuum, as is indeed the case for X-ray light curves of solar flares, it is reasonable to seek out the part of the light curve that looks by eye most likely to contain a QPO and then set the limits on the start and end of the segment manually. However, this method requires significant human supervision and thus will not scale to large data sets. It is open to other forms of selection bias too.

Second, for such transients, we also suggest splitting the light curve up into several parts to see whether one can detect the QPO in all segments or only in some of them. For the final analysis, one should only use segments if there is some evidence for the QPO. Next, for shorter transients where we

suspect the QPO to be present for most of the time, we should conservatively trim the light curve after the onset and before the end of the segment. This way we are less likely to make the light curve nonstationary.

Finally, we have found that regular χ^2-like tests are unsuitable to detect nonstationary bias since only a few bins around the QPO are affected. The overall χ^2-value is mostly determined by the far greater share of frequency bins that lie in the white-noise-dominated part of the PSD. We outline that alternatively calculating χ^2 based solely on the frequencies surrounding the QPO can detect overfitting, which is a hallmark of statistically nonindependent frequency bins.

Aside from periodograms, other methods can be used for the search for QPOs. One alternative are spectrograms that can be used to visualize nonstationary behavior. Spectrograms are created by segmenting the light curve and calculating the periodogram for each section to obtain a time–frequency representation of the power. Thus, the issues highlighted in this paper also apply to spectrograms. We provide a discussion in Appendix C. Wavelet transforms are a popular way to analyze the time series by convolving a wavelet function with a time series of interest (Torrence & Compo 1998). Wavelet analyses are also used in gravitational-wave detection, where they are implemented in the *coherent-wave burst* search pipeline (Klimenko & Mitselmakher 2004; Drago et al. 2021). Wavelet analyses are not restricted to stationary data sets and may help us to find the specific times when a QPO occurs. However, the statistics of detecting QPOs with wavelets in the presence of red noise remains contentious since detrending methods are likely to lead to false detections (Auchère et al. 2016). Alternatively, we can avoid frequency domain methods altogether and model the time series as a Gaussian process with some mean function. The main drawback of Gaussian processes is that they have computational complexity $\mathcal{O}(N^3)$ and thus are only suitable for short time series in the general case (Rasmussen & Williams 2006), though progress has been made to reduce complexity to $\mathcal{O}(N)$ for stationary, complex exponential kernels or combinations thereof (Foreman-Mackey et al. 2017). We note that we expect the same nonstationarity bias if we apply a stationary Gaussian process kernel to a nonstationary time series since the Whittle likelihood explicitly derives from a Gaussian process likelihood. We have also found this result empirically in some preliminary analyses (M. T. Huebner et al. 2022, in preparation).

One advantage of Gaussian processes is that they allow us to fit the overall shape of the transient and the stochastic process simultaneously, instead of lumping the shape of the transient into the red noise, which can help us to prevent the bias demonstrated in Section 4.3. It thereby allows us to avoid nonparametric methods that are likely to create false periodicities (Auchère et al. 2016). Additionally, the Gaussian process likelihood can explicitly take in the known variance for flux values within a light curve if they are available. Thus, they also provide a natural resolution to the bias shown in Appendix B. Finally, Gaussian processes allow us to model relatively simple, nonstationary extensions to the fast, stationary models that are popular now (M. T. Huebner et al. 2022, in preparation). Further development of Gaussian process methods may eventually allow us to create more complex models of QPOs and their possible intermittency within transients.

This work is supported through ARC Centre of Excellence CE170100004. P.D.L. is supported through the Australian Research Council (ARC) Future Fellowships FT160100112, ARC Discovery Project DP180103155. D.H. is supported by the Women In Science Excel (WISE) program of the Netherlands Organisation for Scientific Research (NWO). This work was performed on the OzSTAR national facility at Swinburne University of Technology. The OzSTAR program receives funding in part from the Astronomy National Collaborative Research Infrastructure Strategy (NCRIS) allocation provided by the Australian Government.

Appendix A
Priors and Simulated Data Parameter Tables

We list the priors and parameters used for all the studies in Sections 4 and 5 in Tables 2 and 3.

Table 2
Parameters and Priors Used throughout All Studies in This Paper

Parameter	Description	Prior	Minimum	Maximum
A	Red-noise amplitude	`LogUniform`	$\exp(-30)$	$\exp(30)$
α, α_1, α_2	Red noise / BPL spectral index	`Uniform`	0	10
B	QPO amplitude	`LogUniform`	$\exp(-60)$	$\exp(60)$
f_0	QPO central freq.	`LogUniform`	$2\Delta f$	$f_{\max}$
σ	QPO HWHM	`ConditionalLogUniform`	$\Delta f/\pi$	$0.25f_0$
C	White-noise amplitude	`LogUniform`	$\exp(-30)$	$\exp(30)$
δ	Breaking freq. for BPL	`LogUniform`	Δf	$f_{\max}$

Note. Priors are referred to by their implementation in BILBY. We select wide priors in A, B, and C for simplicity. Although the selection of priors this wide may have a slight impact on the Bayes factor calculation, it will not qualitatively change the scaling of Bayes factor with x. We also set a wide prior on α, or α_1 and α_2, and enforce $\alpha_2 < \alpha_1$ using a `Constraint` prior in BILBY to avoid degeneracies. The prior on f_0 is motivated by the available frequencies in the periodogram. We set the minimum f_0 at twice the difference between neighboring frequencies Δf to ensure better convergence, and we refer to the highest frequency in the periodogram as $f_{\max}$. We set σ, the width of the QPO at half-width at half-maximum, to be conditionally uniform in log between $\Delta f/pi$, i.e., on the scale of a single frequency bin, and $0.25f_0$, to prevent it from converging toward wider features in the periodogram. For the broken power-law analysis of solar flare data, we use priors for $\alpha_{1,2}$ identical to α and the listed δ prior.

Table 3
Values Used to Create the Simulated Data in Section 4

Parameter	Section 4.2 Zeros	Section 4.2 White Noise	Section 4.3	Section 4.4	Appendix B
A	⋯	⋯	⋯	4	⋯
α	⋯	⋯	⋯	2	⋯
B	3	15	80	8	100
f_0	5	5	1	1	5
σ	0.1 Hz	0.1 Hz	0.01 Hz	0.02 Hz	0.1 Hz
C	2	2	2	2	⋯
$A_{\rm flare}$	⋯	⋯	20,000	⋯	$1000\ {\rm s}^{-1}$
t_0	⋯	⋯	70 s	⋯	200 s
$\tau_{\rm r}$	⋯	⋯	10 s	⋯	⋯
$\tau_{\rm f}$	⋯	⋯	20 s	⋯	⋯
$\sigma_{\rm flare}$	⋯	⋯	⋯	⋯	20 s
c_0	⋯	⋯	⋯	⋯	$10\ {\rm s}^{-1}$
Segment length	400 s	400 s	200 s	400 s	400 s
Sampling frequency	40 Hz	40 Hz	40 Hz	40 Hz	40 Hz

Note. The column heads refer to the subsection in which this set of parameters was used. We use two different sets of parameters in Section 4.2 depending on whether we extend using zeros or more white noise.

Appendix B
Deterministic Processes Impact White Noise

White noise observed in astrophysical light curves using photon counting does not arise as a result of intrinsic properties of the source, but rather as a result of the Poissonian noise nature of photon counting. Concretely, given a rate λ, the distribution of the number of observed photons k in a unit of time is Poissonian,

$$\mathrm{Pois}(k;\,\lambda) = \frac{\lambda^k e^{-\lambda}}{k!}. \tag{B1}$$

Photon-counting noise thus scales proportionally to the standard deviation of the Poisson distribution $\sqrt{\lambda}$. This relation implies that photon-counting noise throughout a transient does not remain constant. We demonstrate a limit of this in Section 3.1 when we consider what happens if we extend a time series with zeros. This limit corresponds to the case in which a detector sees no photons at all. In realistic detectors, we will not reach this limit, as there are at least some background photons, though we can get close enough to it for it to lead to wrong inferences about QPO significance.

The above effect points to a deficiency in periodograms more generally. For the calculation of a periodogram we use the values $x(t_i)$ of a discrete time series, but we do not include their variances $\Delta x(t_i)$ if they are available. In effect, this implies a loss of information, as the white noise is already encoded in the variance associated with the photon counts. Instead, we infer it after the fact via power spectral density estimation independently and lose knowledge about nonstationarity in the white noise.

To demonstrate this effect, we construct a transient similar to Section 4.3, though in a way that its low-frequency contributions do not overlap with the QPO frequency, as to demonstrate that this effect arises owing to nonstationarity in the white noise. Concretely, we construct the transient by generating a Gaussian profile plus a constant

$$f(t;\,A_{\rm flare},\,t_0,\,\sigma_{\rm flare}) = A_{\rm flare}\exp\left(-\frac{(t-t_0)^2}{2\sigma_{\rm flare}^2}\right) + c_0, \tag{B2}$$

where $A_{\rm flare}$ is the amplitude, t_0 is the time of the peak, $\sigma_{\rm flare}$ is the width, and the constant c_0 represents a possible background count rate. We add a 20 s nonstationary QPO in the same manner as we describe in Section 4.3. We create photon counts, which we show in Figure 10, by simulating the Poisson process using the implementation in `scipy.stats.poisson.rvs`. The specific parameters of this time series are listed in Table 3. By construction, the QPO is set on the top of the Gaussian profile, which corresponds to the highest level of photon-counting noise. The Gaussian profile, unlike the exponential from Section 4.3, contributes powers at lower frequencies than the exponential, roughly up to $1/\sigma_{\rm flare} \approx 0.05$ Hz. Additionally, we set our QPO frequency at 5 Hz and cut off frequencies below 0.5 Hz so that we can avoid using a red-noise component in our modeling in the same way we did in Section 4.2.

On the tails of the flare, we expect the same effect as we observe when we extend with zeros in our other simulations, though Figure 11 clearly shows that we artificially increase the significance already with very small extension factors. This is possible since the flare quickly falls off on either side and the added photon-counting noise does not compensate for the increased number of frequency bins that contain the QPO. Overall, this effect would be hard to account for just based on better methods for setting data cuts because varying levels of white noise are expected in many transient light curves.

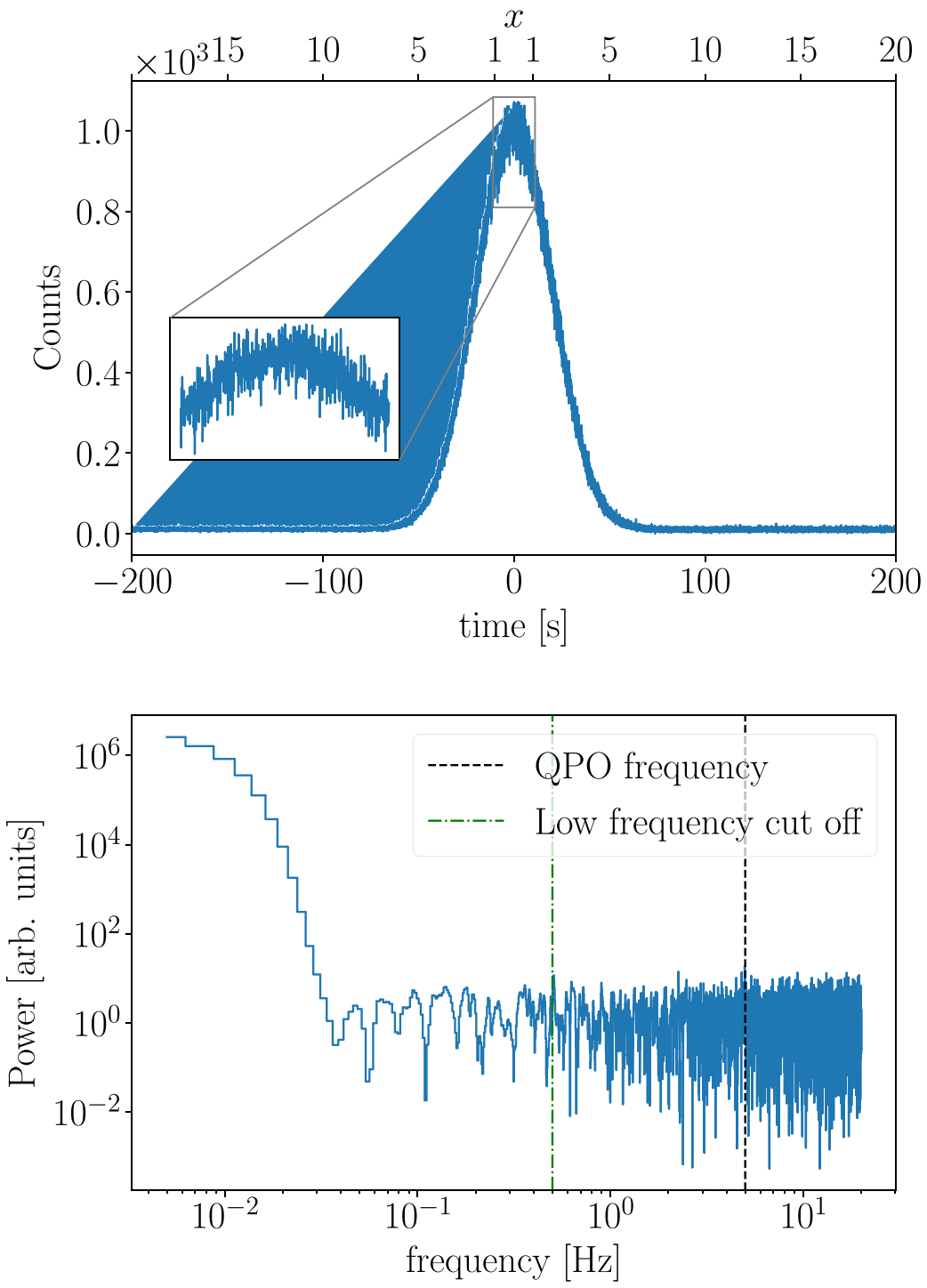

Figure 10. Simulated data of a nonstationary QPO in a deterministic transient flare shape in the presence of white photon-counting noise. The time series (top) shows the dominant transient and the QPO that is present for 20 s on the top of the flare (inset). The periodogram for the entire 400 s time series (bottom) has a relatively wide ($\sigma = 0.1$) QPO at 5 Hz (barely visible). For our analysis, we cut off all frequencies below 0.5 Hz to avoid any effects arising from the power in low frequencies due to the flare shape.

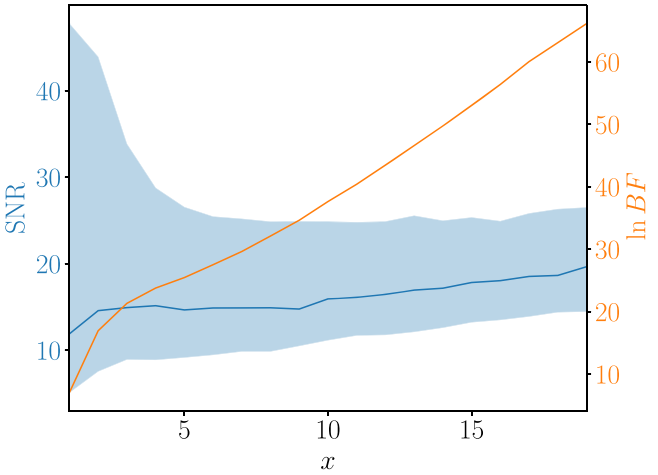

Figure 11. S/N and Bayes factor for the presence of a QPO in a transient flare for different extension factors, with $x = 1$ corresponding to the 20 s inset from the top of Figure 10. While we already start with very strong evidence of a QPO at $x = 1$, the Bayes factors continue to increase as we extend the time series to include more of the transient.

Appendix C
A Note on Spectrograms

Spectrograms[9] are a popular way to identify QPOs in X-ray binaries (see, e.g., Tse et al. 2021, for a recent application). We can obtain a spectrogram by splitting a light curve into several smaller segments that may overlap, calculate an ordinary periodogram for each segment, and stack the segments next to each other. QPOs can be identifiable as horizontal features in the plot. Spectrograms thus may also reveal nonstationary properties of a QPO, such as when it starts or if it drifts in frequency over time. However, there are several caveats to using spectrograms.

First, we may not choose overlapping segments to fit the spectrogram to search for QPOs since that would count some data points multiple times. Using overlapping segments is helpful to enhance features over the noise visually, but it also amplifies statistical flukes and gives us a false sense of the significance of observed features.

Second, we need to set a segment length a priori to calculate the spectrogram. This implies that we have to make a priori adjudications about the QPO frequency we are interested in and for how long the QPO persists in the data. It is thus prone to the same nonstationarity bias as the ordinary periodogram. In Figure 12, we plot spectrograms of the solar flare light curve using 200 s and 2760 s segments. Figure 12 shows that the QPO is far more visually pronounced if we use an overly long segment length. Moreover, we cannot identify the length of features shorter than the segment length. On the other hand, if we use segments that are too short, we risk missing more extended duration features that are only significant if we capture them in a single segment.

Finally, many light curves are not long enough to sensibly create a spectrogram with nonoverlapping segments while also capturing enough QPO cycles to identify features that stand out over the red-noise continuum. While the light curves shown in this manuscript are relatively long, many light curves of interest may only have a few cycles, and segmentation is not helpful. For example, GRB 090709A has attracted some interest because of an apparent periodicity of $T = 8.06$ s in its ≈ 100 s prompt emission (Cenko et al. 2010; De Luca et al. 2010; Iwakiri et al. 2010; Tarnopolski & Marchenko 2021). Given these parameters, we can at best expect to observe about 12 oscillation cycles. Thus, segmenting the light curve even just into two or three nonoverlapping segments leaves us with six or four oscillation cycles, respectively, which makes the identification of the QPO much harder. Moreover, we lose information about possible phase coherence between the segments, making finding a QPO more difficult.

[9] These are sometimes called "dynamical periodograms" in the field of astrophysics.

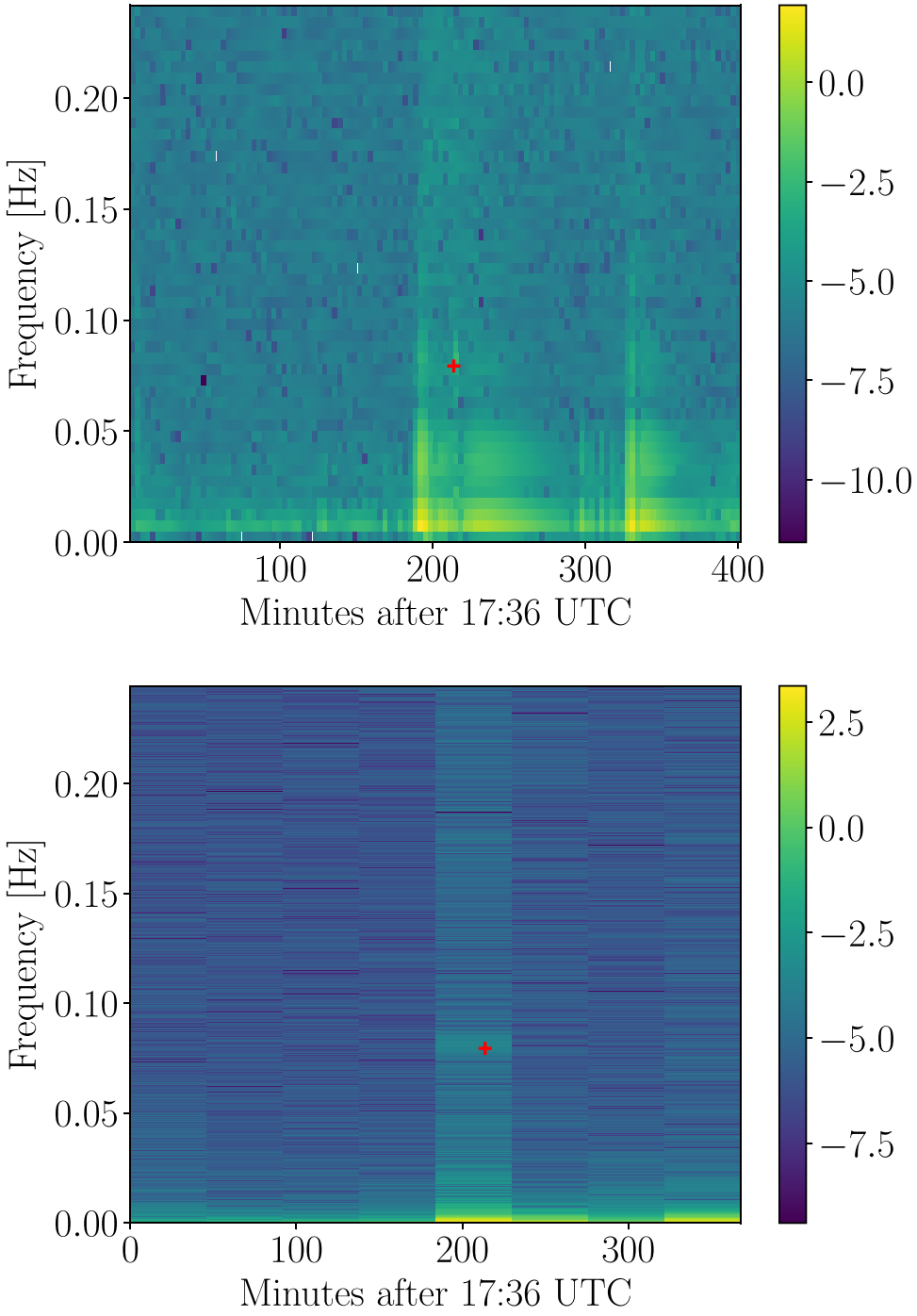

Figure 12. Spectrograms of the solar flare light curve using 200 s (top) and 2760 s nonoverlapping segments (bottom), corresponding to the segment durations in Section 5. We highlight the central location of the QPO in the time–frequency plot with the red plus sign. The color bar shows the log of the power in each time–frequency bin. The QPO is hardly identifiable in the top panel, as it only appears for 1–2 time bins. Meanwhile, the QPO is strongly pronounced in the bottom panel owing to the nonstationarity bias.

ORCID iDs

Moritz Hübner https://orcid.org/0000-0002-9642-3029
Daniela Huppenkothen https://orcid.org/0000-0002-1169-7486
Paul D. Lasky https://orcid.org/0000-0003-3763-1386
Andrew R. Inglis https://orcid.org/0000-0003-0656-2437

References

Ashton, G., Hübner, M., Lasky, P. D., et al. 2019, ApJS, 241, 27
Auchère, F., Froment, C., Bocchialini, K., Buchlin, E., & Solomon, J. 2016, ApJ, 825, 110
Barret, D., & Vaughan, S. 2012, ApJ, 746, 131
Blackman, R. B., & Tukey, J. W. 1958a, BSTJ, 37, 185
Blackman, R. B., & Tukey, J. W. 1958b, BSTJ, 37, 485
Bonferroni, C. 1936, Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze, 8, 3
Broomhall, A. M., Davenport, J. R., Hayes, L. A., et al. 2019, ApJS, 244, 44
Cenko, S. B., Butler, N. R., Ofek, E. O., et al. 2010, AJ, 140, 224
Chatfield, C., & Xing, H. 2019, The Analysis of Time Series (7th ed.; Boca Raton, FL: CRC Press)
Covino, S., Landoni, M., Sandrinelli, A., & Treves, A. 2020, ApJ, 895, 122
De Luca, A., Esposito, P., Israel, G. L., et al. 2010, MNRAS, 402, 1870
Drago, M., Klimenko, S., Lazzaro, C., et al. 2021, SoftX, 14, 100678
Foreman-Mackey, D., Agol, E., Ambikasaran, S., & Angus, R. 2017, AJ, 154, 220
Foster, G. 1996, AJ, 112, 1709
Gelman, A., Carlin, J. B., Stern, H. S., et al. 2013, Bayesian Data Analysis (3rd ed.; New York: Chapman and Hall/CRC), 677
Hayes, L. A., Gallagher, P. T., Dennis, B. R., et al. 2019, ApJ, 875, 33
Hayes, L. A., Inglis, A. R., Christe, S., Dennis, B., & Gallagher, P. T. 2020, ApJ, 895, 50
Huppenkothen, D., D'Angelo, C., Watts, A. L., et al. 2014a, ApJ, 787, 128
Huppenkothen, D., Watts, A. L., & Levin, Y. 2014b, ApJ, 793, 129
Huppenkothen, D., Watts, A. L., Uttley, P., et al. 2012, ApJ, 768, 87
Huppenkothen, D., Younes, G., Ingram, A., et al. 2017, ApJ, 834, 90
Inglis, A. R., Ireland, J., Dennis, B. R., Hayes, L., & Gallagher, P. 2016, ApJ, 833, 284
Inglis, A. R., Ireland, J., & Dominique, M. 2015, ApJ, 798, 108
Ingram, A. R., & Motta, S. E. 2019, NewAR, 85, 101524
Ireland, J., McAteer, R. T., & Inglis, A. R. 2015, ApJ, 798, 1
Israel, G. L., Belloni, T., Stella, L., et al. 2005, ApJ, 628, L53
Iwakiri, W., Ohno, M., Kamae, T., et al. 2010, in AIP Conf. Proc. 1279, Deciphering the Ancient Universe with Gamma-Ray Bursts (Melville, NY: AIP), 89
Klimenko, S., & Mitselmakher, G. 2004, CQGra, 21, S1819
Lachowicz, P., & Done, C. 2010, A&A, 515, A65
Miller, M. C., Chirenti, C., & Strohmayer, T. E. 2019, ApJ, 871, 95
Morris, D., Battista, F., Dhuga, K., & MacLachlan, G. 2010, in AIP Conf. Proc. 1279, Deciphering the Ancient Universe with Gamma-Ray Bursts (Melville, NY: AIP), 394
Nakariakov, V. M., & Melnikov, V. F. 2009, SSRv, 149, 119
Nita, G. M., Fleishman, G. D., Gary, D. E., Marin, W., & Boone, K. 2014, ApJ, 789, 152
Norris, J. P., Bonnell, J. T., Kazanas, D., et al. 2005, ApJ, 627, 324
Percival, D. B., & Walden, A. T. 1993, Spectral Analysis for Physical Applications (Cambridge: Cambridge Univ. Press)
Rasmussen, C., & Williams, C. 2006, Gaussian Processes for Machine Learning, Vol. 7 (Cambridge, MA: MIT Press)
Romero-Shaw, I. M., Talbot, C., Biscoveanu, S., et al. 2020, MNRAS, 499, 3295
Skilling, J. 2006, BayAn, 1, 833
Speagle, J. S. 2020, MNRAS, 493, 3132
Strohmayer, T. E., & Watts, A. L. 2005, ApJ, 632, L111
Strohmayer, T. E., & Watts, A. L. 2006, ApJ, 653, 593
Tarnopolski, M., & Marchenko, V. 2021, ApJ, 911, 20
Timmer, J., & König, M. 1995, A&A, 300, 707
Torrence, C., & Compo, G. P. 1998, BAMS, 79, 61
Tse, K., Galloway, D. K., Chou, Y., Heger, A., & Hsieh, H.-E. 2021, MNRAS, 500, 34
van der Klis, M. 1989, ARA&A, 27, 517
Van Doorsselaere, T., Kupriyanova, E. G., & Yuan, D. 2016, SoPh, 291, 3143
Vaughan, S. 2005, A&A, 431, 391
Vaughan, S. 2009, MNRAS, 402, 307
Watts, A. L., & Strohmayer, T. E. 2006, ApJ, 637, L117
Weakliem, D. L. 1999, Sociol. Methods Res., 27, 359
Whittle, P. 1951, Hypothesis Testing in Time Series Analysis (Uppsala: Almqvist & Wiksell)
Zhu, X. J., & Thrane, E. 2020, ApJ, 900, 117
Zimovets, I. V., McLaughlin, J. A., Srivastava, A. K., et al. 2021, SSRv, 217, 66