Pseudo-nonstationarity in the scaling exponents of finite interval time series

K. H. Kiyani* and S. C. Chapman
Centre for Fusion, Space and Astrophysics; Department of Physics, University of Warwick, Gibbet Hill Road, Coventry, CV4 7AL, United Kingdom

N. W. Watkins
British Antarctic Survey, High Cross, Madingley Road, Cambridge CB3 0ET, United Kingdom

The accurate estimation of scaling exponents is central in the observational study of scale-invariant phenomena. Natural systems unavoidably provide observations over restricted intervals; consequently a stationary stochastic process (time series) can yield anomalous time variation in the scaling exponents, suggestive of non-stationarity. The variance in the estimates of scaling exponents computed from an interval of N observations is known for finite variance processes to vary as $\sim 1/N$ as $N \to \infty$ for certain statistical estimators; however, the convergence to this behaviour will depend on the details of the process, and may be slow. We study the variation in the scaling of second order moments of the time series increments with N, for a variety of synthetic and 'real world' time series; and find that in particular for heavy tailed processes, for realizable N, one is far from this $\sim 1/N$ limiting behaviour. We propose a semi-empirical estimate for the minimum N needed to make a meaningful estimate of the scaling exponents for model stochastic processes and compare these with some 'real world' time series.

PACS numbers: 05.45.Tp, 89.75.Da

I. INTRODUCTION

Testing for, and quantifying scaling is central to the application of statistical theories to 'real-world' extended systems. A broad range of theoretical frameworks such as turbulence [1], critical phenomena [2] and complex systems [3] frame their predictions in terms of the statistical properties of (arbitrarily large) ensembles as a function of scale. Under the assumption of ergodicity the statistical scaling property of an extended system is captured to some extent by a reduced (embedded) set of observations or measurements; so that a 1-D cut through a N dimensional system will be sufficient to indicate whether scaling is present, and in a quantitative way can usefully restrict the scaling exponents of the system as a whole. This approach is pragmatic – in physical systems it is generally not practicable to capture and analyze the full spatiotemporal information of all points in the system on all scales. A key observable is then the quantitative scaling properties of such a one dimensional sample or time series. An example of this is the use of the Taylor hypothesis in turbulence, where the time series at a single point is used as a proxy for the statistical properties of the flow [4].

Time series are also often parsed into sub-intervals to isolate processes of interest, or to remove features which might contaminate the calculation of the quantity of interest. Examples of this in the study of solar wind turbulence are the separating of fast and slow wind, and open/closed field line regions [5, 6]; isolating or removing signals of interplanetary shocks, magnetosheath crossings, and coronal mass ejection remnants [7, 8]; or where the interval is restricted by a secular change in parameters as the spacecraft changes location [9, 10]. Examples in the study of the earth's geomagnetic field include isolating 'quasi-stationary' and 'quiet' intervals in magnetic field data [6, 11]; and the effects of finite sample size in the power spectral exponent estimates in the ionosphere by ground-based measurements [12]. 'Locally stationary processes' are also discussed in speech signal analysis [13] and physiological non-stationary signals [14]; and of course statistical forecasting, whether in the context of seasonal weather or the financial markets [15], is based on time series histories which rely on the stationarity assumption. In all of these cases, it is intuitively apparent that smaller data intervals will result in poorer statistics, which will be manifest in the variance of the observed values of the exponents. The observation of a (non-secular) variation in the scaling exponents therefore has two interpretations; either it is due to intrinsic fluctuations as a result of the finite N interval, or it is a consequence of non-time stationarity of the time series $x(t)$ i.e. different scaling behaviour due to different physical phenomena. If the properties of the underlying process are not known *a priori* we need a method to distinguish these two interpretations in a quantitative manner; or at best to put a degree of confidence that it is due to one and not the other.

The most commonly used tool to both establish and quantify scaling in a time series $x(t)$ is to test for scaling of the power spectral density $F(\omega) \sim \omega^{-\beta}$, and obtain the exponent β. In a physical system, such scaling can only be observed over a finite range, limited by the interval (in time t) of N observations over which the system is measured. From large-sample theory (asymptotic limit of $N \to \infty$) the variance in the power spec-

*Electronic address: k.kiyani@warwick.ac.uk

tral exponent β computed using least squares regression from an interval of N samples is known [16, 17] for finite variance processes to vary as $\sim 1/N$ as $N \to \infty$. One method to obtain more complete information about the scaling properties of a stochastic process $x(t)$ is captured by how the statistical properties of the increments $y(t,\tau) = x(t+\tau) - x(t)$ vary with the differencing scale τ; the differencing being a particular type of coarse-graining operation which has been chosen due to the easy analogy with random walks, return probabilities etc. However, there exist other coarse-graining operations which although more involved, possess additional highly desirable properties when studying scaling. In particular, wavelets which (with some wavelet functions) when combined with their detrending capabilities have been shown to be a natural and computationally efficient way of studying scale-by-scale statistical behaviour [13, 18, 19]. In this paper we will discuss the behaviour of the scaling properties of the second order moment $\left\langle y(t,\tau)^2 \right\rangle \sim \tau^{\zeta(2)}$. We may anticipate that the statistical properties of this scaling exponent $\zeta(2)$ follow that of β; indeed there exist many results for a range of different estimators of the $\zeta(2)$ [20, 21] that directly show the asymptotic $\sim 1/N$ behaviour discussed above. In practice, the convergence to this $\sim 1/N$ behaviour will depend on the details of the process and the estimator and, as we shall show in this paper, is often slow.

An essential tool in the analysis of 'real world' time series in the context of scaling is then a prescription for the variance in the scaling exponents of $x(t)$ as a function of the number of observations N in the chosen interval. In this paper we make some first steps in this direction by obtaining empirical estimates from the study of a variety of stochastic processes that have been used as models for physical systems. We focus on finite size N realizations of self-affine cases with Gaussian distributed increments in the form of a standard Brownian motion and fractional Brownian motion (fBm); and those with heavy tails, namely α-stable Lévy motion and linear fractional stable motion (LFSM) [22–24]. A representative case for multifractal scaling is provided by the p-model, often used to characterize observations of turbulence [25, 26].

The fundamental property of ergodicity in systems that exhibit scaling implies time stationarity. In its strong sense time stationarity implies that the probability density function (PDF) of $x(t)$ does not change with time; this is known as strict stationarity. Pragmatically, weak stationarity, that is time independence of the variance or second order moment is usually adopted – the latter convenience is usually assumed due to the special place that the Central Limit Theorem and the Gaussian distribution hold in statistics. In this paper we are concerned specifically with the behaviour of scaling exponents which are characterized through the statistical properties of the increments $y(t,\tau)$, rather than the time series $x(t)$; hence we will use as our test time series examples that have stationarity in $y(t,\tau)$, and not in $x(t)$.

We will focus on the statistics of the scaling exponent of the second order moment of the increments, as this also captures the power spectral exponent β, and for self-affine finite-variance processes the Hurst exponent H (see next section and also [27] for the infinite variance case). We will study these processes for a range of values of N that are feasable for realisable physical systems; and find that in particularly for the heavy-tailed processes, the variance in the exponent with N shows a significant departure from the $1/N$ asymptotic behaviour. Nevertheless, for these heavy-tailed processes, we find empirical evidence of an intermediate range of scaling with $N^{-\gamma}$. We will estimate the time series interval N required to capture the scaling exponent to reasonable precision; this places a lower limit on the sample size. A related study to this was conducted to investigate and compare the effects of finite sample size on different statistical estimators for the Hurst exponent H for a Gaussian white noise process [28]. Stationarity also implies a particular PDF of the values of the exponent obtained from many, length N, realisations of a given process. This is known asymptotically for $N \to \infty$ for the processes based on Gaussian increments (generalizable to finite variance processes) and is also known in this asymptotic sense for infinite variance processes; both processes approaching a Gaussian distribution for the scaling exponents as $N \to \infty$ [16, 17, 20, 24, 29, 30] (using least squares and maximum likelihood estimation schemes). For the intermediate stage of finite N we find intermediate distributions for the exponents; resembling both the asymptotic Gaussian forms and, for heavy-tailed data, Gumbel max-stable (Extreme value type I) distributions. Comparing these results with that found for real-world time series may provide an additional test for stationarity in the increments. In this spirit we finally illustrate these ideas with some examples of real-world time series in the form of *in-situ* observations of magnetic field and velocity in the turbulent solar wind using data from spacecraft at 1AU in the ecliptic; and comment on the statistical properties of their scaling exponents in light of the representative synthetic toy models.

II. METHODOLOGY

We will focus attention on the scaling exponent $\zeta(2)$ of the second order moment of the increments also known as the second order structure function:

$$\left\langle y(t,\tau)^2 \right\rangle = \left\langle \left(x(t+\tau) - x(t)\right)^2 \right\rangle = \left\langle y(t,1)^2 \right\rangle \tau^{\zeta(2)} \ , \quad (1)$$

where we have assumed that the increment process is at least second order stationary i.e. $\left\langle y(t,\tau)^2 \right\rangle = \left\langle y(\tau)^2 \right\rangle$ (weak-stationarity). In particular, this implies that the power spectral density of a discrete time random walk $x(t)$ of *i.i.d.* stationary increments with finite variance, scales as [13]

$$F(\omega) \sim \omega^{-(\zeta(2)+1)} \ , \quad (2)$$

where the scaling exponent $\zeta(2)$ is related to the power spectral exponent β of $x(t)$ by $\zeta(2) = \beta - 1$. For self-affine process with Hurst exponent H the PDF $P(y,\tau)$ of the increments obeys the scaling relation (for the case of α-stable processes with finite N see [27], and the discussion to follow)

$$P(y,\tau) = \tau^{-H}\mathcal{P}^s(\tau^{-H}y) \ , \tag{3}$$

where the PDF P at any scale τ can be collapsed onto a unique scaling function $\mathcal{P}^s$. The scaling relation (3) implies that the scaling of the structure functions to all orders p [27] is given by $\langle y(\tau)^p\rangle = \langle y(1)^p\rangle\,\tau^{\zeta(p)}$ where $\zeta(p) = Hp$; and thus we have that $\zeta(2) = 2H$. Our results concerning the statistical behaviour of $\zeta(2)$ with N will thus also apply to the power spectral exponent β for all the models concerned and the Hurst exponent H for the self-affine models; both are commonly used to characterize statistical scaling. Our remarks can also be generalized to the scaling exponents $\zeta(p)$ of structure functions of higher-order positive moments. These are relevant for multifractal processes where the $\zeta(p)$ are a nonlinear function of p and so H or β are not sufficient to determine the complete statistical scaling of the $y(t,\tau)$.

Our study consists of partitioning a given time series $x(t)$ into L equal intervals of sample size N denoted by $x_i(t)$ where $i = 1 \ldots L$. Each of these intervals are then differenced on scale τ to produce a time series of the increments $y_i(t,\tau) = x_i(t+\tau) - x_i(t)$ of the process $x_i(t)$.

We will look for scaling of the second order moment (structure function)

$$M_i^2(\tau) = \left\langle y_i(\tau)^2\right\rangle = \int_{y_i^-}^{y_i^+} y_i^2 P_i(y_i,\tau)dy_i \ , \tag{4}$$

with τ such that $M_i^2(\tau) = M_i^2(1)\tau^{\zeta_i(2)}$. Again, the index i indicates the i^{th} interval over which the exponents are calculated and tracks any (real or statistical) time variation in the value of $\zeta_i(2)$. In an infinitely large interval, $N \to \infty$, the limits of the integral $y_i^{\pm} \to \pm\infty$; here however each i^{th} interval of the time series will impose different finite extremal values $y_i^{\pm}$. For the heavy-tailed processes in particular, the statistics of the $y_i^{\pm}$ can be anticipated to have a significant effect on the statistics of the $\zeta_i(2)$; this has been discussed for the case of α- stable Lévy processes in [27]. These Lévy processes, possess heavy tails in the PDFs of their increments, with tails that fall as $P(y) \sim y^{-(1+\alpha)}$ power-laws. The α-stable Lévy processes have divergent moments for $p = 2$ and above; for a finite sized sample the moments exist but can be dominated by the behaviour of rare outlying points in the tails which introduce a pathological bias when estimating scaling exponents from the moments [27] (for a wider discussion see [31]). We circumvent these difficulties, at least for self-affine time series, by restricting our analysis to the scaling of the second order moment $\zeta(2)$, and by using the iterative conditioning technique [27]. This simple and robust technique for exponent estimation removes a small percentage of the extreme data values which are poorly sampled statistically. In some pathological cases such as the α-stable Lévy distributions these rare extreme points are of the order of and sometimes larger than the whole sum [32]. Because they are so large they tend to dominate the statistics and thus the scaling of the higher order moments. This can be clearly seen if we look at the discrete definition of the moments of order p

$$M_i^p(\tau) = \frac{1}{N}\sum_{j=1}^{N}\left(y_i^p\right)_j \ . \tag{5}$$

The reasoning and full illustration of this iterative conditioning method to heavy-tailed non-Gaussian distributions is discussed in [27]. Although not discussed in this paper, the iterative conditioning technique is also an unbiased robust technique for distinguishing self-affine (monofractal) from multifractal behaviour.

We will focus here on parameter stationarity as opposed to trend stationarity. The former refers to the change in the intrinsic dynamics of the process of interest as characterized by its quantitative statistical properties (the behaviour of the moments); as opposed to the latter which is simply an additive trend to the signal. In particular we will focus on the stationarity of the scaling of the moments as captured by the exponents $\zeta_i(p)$. If secular trends are present in the time series then the time series of increments will be approximately trend-free provided our differencing scale τ is sufficiently small [33]. A secular trend can also be removed by detrending or by the method of studying the scaling of moments of wavelet coefficients where an appropriate wavelet is chosen with a large number of zero-moments [19, 34]. The more complex case of mixed dynamics i.e. two or more intrinsically different processes represented in different sections of a time series will not be considered here.

A. Data generation and sources

We will consider synthetically generated signals that are both stationary and nonstationary with respect to their increments. The signals with stationary independent increments will consist of a standard Brownian motion and standard symmetric α-stable Lévy motion for four values of the exponent α [22, 35]; the latter being highly non-Gaussian and heavy-tailed with very large excursions in their time series. To survey a broad range of such processes we have also included non-Markovian versions of the above processes. These include a long-memory fractional Brownian motion (fBm), and a long-memory persistent and anti-persistent linear fractional stable motion (LFSM) – see [17, 22–24] for more details on these processes and in particular [24] for the algorithm and MATLAB code for the LFSM.

We also investigate a multifractal time series generated from a discrete multiplicative cascade process in the form of the p-model [25, 26]. The p-model is used as a model

for intermittent turbulence [1, 36]. The intermittency of the p-model time series leads to non-time stationary finite N moments; however the set of scaling exponents $\zeta_i(p)$ are stationary.

The nonstationary time series we will consider are a standard Brownian motion with linearly varying standard deviation of its increments with time $\sigma \sim t$, and cyclically varying standard deviation (cyclically stationary) $\sigma \sim sin^2(t)$. All of the above synthetic time series were generated in MATLAB with appropriate random seeding and sample sizes of $N \sim 10^6$.

Lastly, we will consider three real-world time series which have been found to exhibit scaling [37–40]. These consist of two time series of 100 second resolution magnetic field B_z and speed v from the NASA WIND spacecraft at 1AU in the solar minimum year 1996; and a 64 second resolution one year long time series of the magnetic field energy density B^2 from the NASA ACE spacecraft in the solar maximum year 2000. All of these time series consist of $N \sim 5 \times 10^5$ data samples and can be downloaded from CDA web `http://cdaweb.gsfc.nasa.gov/`.

III. RESULTS

We study the variation of the scaling exponent of the second order moment $\zeta_i(2)$ with sample size N. The process by which the exponent $\zeta_i(2)$ is estimated for L contiguous intervals of N points of a time series is illustrated in Figure 1 for the p-model. We begin with the time series in Figure 1(a) which we parse into L intervals. For each of these intervals we obtain an estimate of $\zeta_i(2)$ as the gradient of a linear least squares regression to a log-log plot of the second-order moment $M_i^2(\tau)$ versus the scale or differencing parameter τ. This method of obtaining the scaling exponents is also known as the structure function technique [1, 5] and is closely related to variance plot, correlogram and log-periodogram techniques [17, 30] – in the latter reference [30] it is identical to the variogram technique. We focus on this particular method to estimate $\zeta_i(2)$ as it provides a point of contact with asymptotic $N \to \infty$ estimates of the variance of the power spectral exponent β which are based on linear regression over a finite range power law power spectrum [16, 17]. In both cases, the variance in the estimated exponent will depend upon the details of the linear regression. For the second order moment these details include the range of values of τ over which $M_i^2(\tau)$ is a power law; the number of different τ for which we calculate $M_i^2(\tau)$ and use in the linear regression; and the uncertainty of each $M_i^2(\tau)$ value. In all cases considered here we optimize these details to minimize the linear regression error but importantly use the same algorithm for all of the sample time series that we discuss.

The linear fit is obtained by linear least-squares regression which also provides an estimate of the error. We augment this estimate of the error by varying the start and end points of the regression by a few points and obtaining the difference in the exponents. The linear regression was done over ~ 20 values of the scale parameter τ, where τ was increased geometrically as $\tau = base^k$, where $k \in \{0, \cdots, 40\}$ and $base$ was chosen to be 1.2. The fit was done over this reduced set of measurements at ~ 20 values of τ so that a fair comparison can be made with the real-world data (to be discussed later) where only a limited power-law range is seen.

Due to its highly intermittent nature the $p = 0.6$ p-model is not time stationary in its finite N moments and this can be seen in Figure 1(b) where we plot consecutive values of the second order moment $M_i^2(\tau)$ obtained for each of the L intervals of N points, shown for $\tau = 1$ and two values of N. For the p-model time series shown here, the second order moment follows the local amplitude of fluctuations in the time series itself; comparing the ratio of the amplitude of these fluctuations to the signal amplitude is one of the classical 'first base' techniques for establishing whether the signal is stationary (in the weak sense)[33]. As one would expect from (5), this variation of the second-order moment $M_i^2(\tau)$ with the amplitude of the time series is emphasized as we decrease N as any estimates of the statistics from smaller sample size will naturally mimic the more local features of the time series. This behaviour is more pronounced in very intermittent signals i.e. those with heavy-tailed fluctuation PDFs.

We also plot in Figures 1(c) and (d) the corresponding estimates of $\zeta_i(2)$ for each interval. These two panels show the same data, that is, the estimates of $\zeta_i(2)$ plotted without (c) and with (d) error bars obtained from the linear regression and the error augmentation outlined above. As intuitively expected, if we decrease the sample size N over which the $\zeta_i(2)$ are computed, the scatter increases. However unlike the moments, there is no clear trend with the amplitude of the signal, indicating stationarity of the scaling exponent $\zeta_i(2)$. This latter phenomenon will also be encountered in the non-stationary Brownian time series we will study. The estimates of $\zeta_i(2)$ can be seen to vary by up to a factor of two for $N = 10^4$ for this realization of the p-model. This underlies the difficulty of obtaining physically meaningful estimates of scaling exponents for physically realizable N. We can see that the error bars approximately capture the fluctuations in the estimates of $\zeta_i(2)$ for the case of the p-model. As we wish to include strongly non-Gaussian processes in our study, we will henceforth present numerical estimates of the variance of $\zeta_i(2)$ obtained directly from computing many values of $\zeta_i(2)$ rather than the linear regression error.

The essential point is that quite significant variation in the scaling exponents can arise in time stationary, but intermittent, time series; even when these are estimated over intervals of data that might intuitively be considered to be of adequate length. In order to distinguish variation in the scaling exponents that is statistical (finite N effect) as shown above, from that which reflects intrinsic non-time stationarity, some estimate of the N dependence

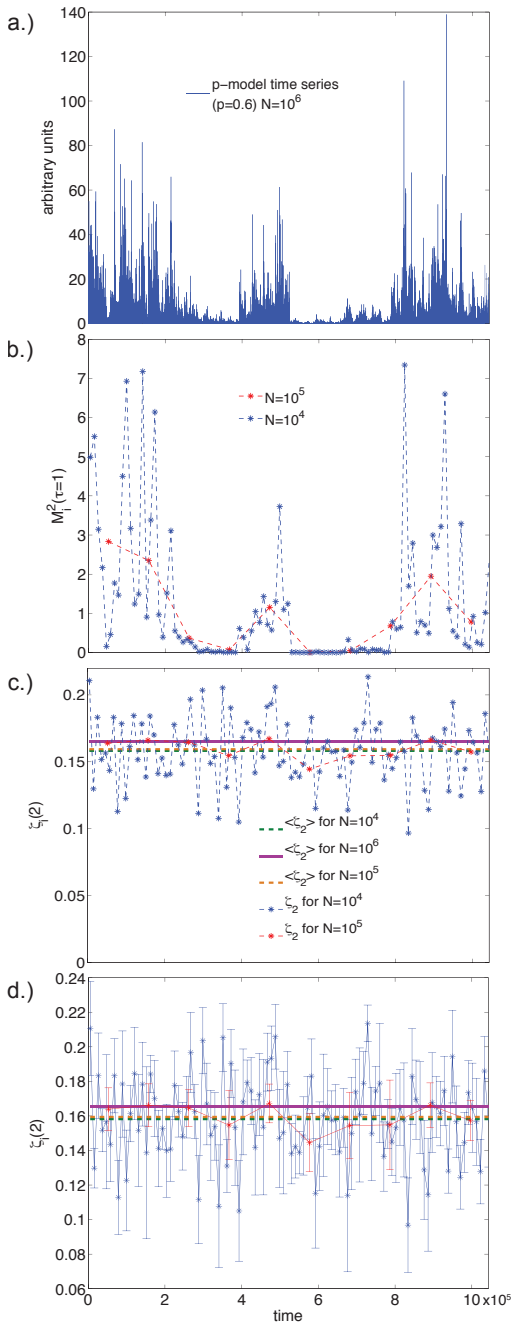

Figure 1: (Color online) a.) Time series of length $N = 10^6$ for the p-model ($p = 0.6$). b.) Variation of the second order moment of the increments $y_i(t, \tau)$ for time-scale $\tau = 1$ of the above time series where the original time series has been partitioned into $L = 100$ and $L = 10$ equal sized intervals. c.) Variation of conditioned $\zeta_i(2)$ with time for the p-model with the same segmentation as in b.) – also shown are the mean values of the exponents for different partitioning corresponding to different sample sizes; d.) Same as (c.) but with errors explicitly shown.

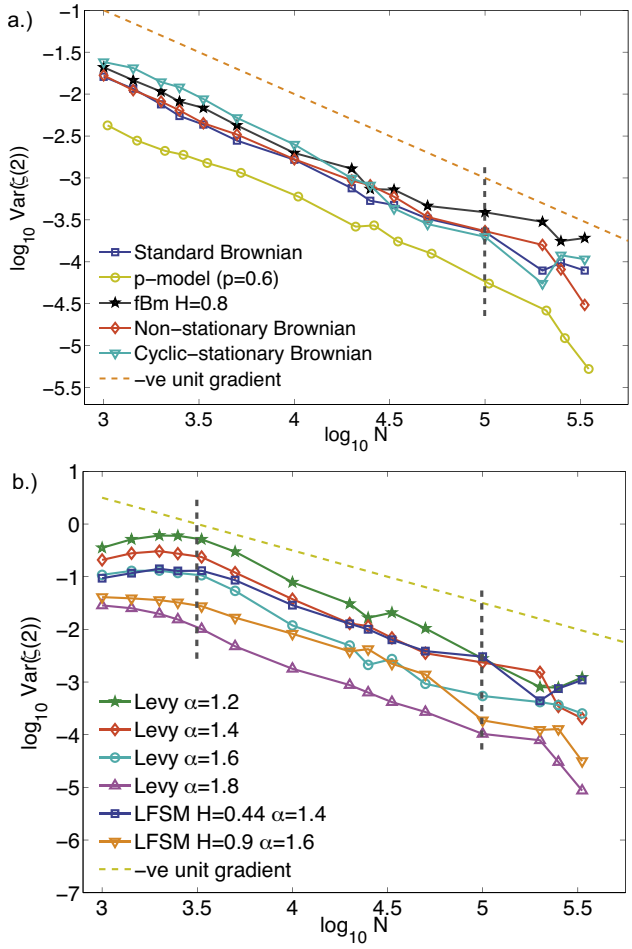

Figure 2: (Color online) a.) The variance of conditioned $\zeta(2)$ with sample size N for all the synthetic finite variance processes studied shown on a log-log plot. b.) same as in a.) for all the synthetic infinite variance processes studied. The diagonal dashed line on both these plots indicates a negative slope of unit gradient so that comparison with theoretically expected asymptotic behaviour can be made. The vertical black dashed lines indicate the areas outside of which errors begin to dominate due to i.) (bottom vertical line) lack of values of $\zeta_i(2)$ to make a decent estimate of $Var(\zeta(2))$; and ii.) (top vertical line) failure of the iterative conditioning technique to obtain unbiased estimates of $\zeta(2)$.

of the variance of $\zeta(2)$ is needed; this will also point to an estimate of the minimum number of observations N needed to obtain a 'reasonably accurate' estimate of $\zeta(2)$. We will now explore the variance of $\zeta(2)$ as a function of N.

In the limit $N \to \infty$, β, when estimated via a log-periodogram varies as $Var[\beta] \sim 1/N$ [16, 17]. This limiting behaviour is also known for some other estimation schemes of the self-similarity parameter [20] (as we discuss later here). Thus we would anticipate that for sufficiently large N, $Var[\zeta_i(2)] \sim 1/N$ for our moment scaling estimation also. However, we do not know the rate of convergence with N to this limiting behaviour and can also

anticipate that this will depend upon whether the PDF of the increments is heavy-tailed, and whether or not the increments are dependent – both of which introduce further difficulties in obtaining an unbiased estimator.

In Figure 2 we plot the variance of $\zeta_i(2)$ against the sample size N on log-log axes, for a range of N that are feasable in realistic realizations of physical systems. Figure 2(a) shows the behaviour of a subset of our synthetic time series that are intrinsically finite variance processes; Figure 2(b) shows all the synthetic time series from infinite variance processes that we consider. Plotting these on log-log axes reveals a characteristic power law trend for all the processes:

$$Var[\zeta(2)] = CN^{-\gamma} \ . \tag{6}$$

We see that indeed, $\gamma \sim 1$ for the intrinsically finite variance processes. More pragmatically, we can use this plot to make an estimate of the minimum sample size N_{min} needed in order to estimate $\zeta(2)$ such that the error introduced from the small sample size $N = N_{min}$ is, say, $\sim 5\%$. We propose a simple criterion

$$\frac{\sqrt{Var[\zeta(2)]}}{\zeta(2)|_{L=1}} \lesssim 0.05 \ , \tag{7}$$

where $\zeta(2)|_{L=1}$ is the value of $\zeta(2)$ estimated for the entire time series (assuming that the scaling is stationary). This leads to

$$Var[\zeta(2)] \lesssim \left(0.05\zeta(2)|_{L=1}\right)^2 \ , \tag{8}$$

where the value of N_{min} is extrapolated from the plot of $Var[\zeta(2)]$ Vs. N, from Figure 2(a). For these finite variance processes expressions (7) and (8) yield $N_{min} \sim 10^3$ for the fBm model; $N_{min} \sim 10^4$ for the standard Brownian motion (stationary and non-stationary); and $N_{min} \sim 10^5$ for the the p-model.

One can invert these relationships to obtain the approximate error on $\zeta_i(2)$ given a sample size N from which it was calculated. The constant C in (6) is also intrinsic to our estimate of N_{min}; operationally the procedure for obtaining the error on $\zeta_i(2)$ in this manner would also include estimating C from the plot in Figure 2(a).

Processes that show scaling often have increments drawn from a heavy tailed PDF, these may also not intrinsically have finite variance as is the case for the α-stable Lévy processes. Figure 2(b) shows the N dependence of all the infinite variance synthetic time series that we have considered, including those with long-range memory. The curves are all generated from time series which possess heavy-tailed PDFs for their increments. These include both the ordinary and fractional Lévy increments. The curves in Figure 2(b) have a range of γ values close to but also clearly distinct from $\gamma = 1$. As will be discussed later this is due to slow convergence to the asymptotic N^{-1} behaviour; from Figure 2(b) we can see that the Lévy process which is closest to Gaussian, namely with $\alpha = 1.8$, has behaviour closest to $\gamma \sim 1$. In a similar way to the method used above for the finite variance synthetic processes, we make empirical estimates of N_{min} required to obtain an estimate of $\zeta(2)$ to within $\sim 5\%$ for the infinite variance processes. For the Lévy processes $\alpha = 1.2$ and $\alpha = 1.4$, and LFSM ($H = 0.44$, $\alpha = 1.4$) $N_{min} \sim 10^5$; for the $\alpha = 1.6$ case $N_{min} \sim 10^4$; and for the $\alpha = 1.8$ case and LFSM ($H = 0.9$, $\alpha = 1.6$) $N_{min} \sim 10^3$. The relevant property in the context of estimating the uncertainty on $\zeta_i(2)$ is that for realizable N, these processes do not show an N^{-1} dependence. Also, unlike the Gaussian processes in Figure 2(a) which cluster around a similar C value, the infinite variance processes have noticeably different values of C which depends on both the the tail exponent α and also on the degree of memory in the process given by $H - (1/\alpha)$ [20].

Finally, combining equations (6) and (8) we obtain

$$N_{min} = C^{1/\gamma} \left(0.05\zeta(2)|_{L=1}\right)^{-2/\gamma} \ , \tag{9}$$

where both C and γ depend on the process in question; and for finite variance processes $\gamma = 1$.

Error analysis

To estimate the errors in the estimates of $Var(\zeta(2))$ a small monte-carlo type study was performed in which different random seeds were used to generate 10 different realisations of the two archetypal processes studied here i.e. finite and infinite variance processes in the form of 10 different realisations of a standard Brownian motion and a standard α-stable Lévy process ($\alpha = 1.4$). The computation of the $\log Var(\zeta(2))$ Vs. $\log N$ plots were then calculated for each of these realisations; these are shown in Figures 3(a) and (b). We then average over these realisations to obtain an average value of $Var(\zeta(2))$ for each N, shown on log-log axes in Figures 3(c) and (d); the ensemble of realisations also provides an error on this value via the maximum deviation from this average.

At large N errors are dominated by there being fewer values of computed $\zeta_i(2)$ and at small N, by poor resolution of the PDF from which we ultimately compute $\zeta_i(2)$. In particular, at small N we can see from the plots for the α-stable processes that there is a systematic deviation from power law behaviour in N. This arises from a breakdown in the iterative conditioning technique [27] at small N.

In the next section we will discuss the PDFs of the scaling exponents $\zeta_i(2)$ obtained from this study. When these are close to Gaussian, standard Chi-squared distributions and F-test techniques could provide methods of obtaining errors for values of $Var[\zeta(2)]$, even from a single realisation. In this context we should mention the use of bootstrap re-sampling methods for providing distributions, confidence intervals and statistical significance for parameter estimates in situations when one is limited by a single realisation [41, 42]. Although the convergence

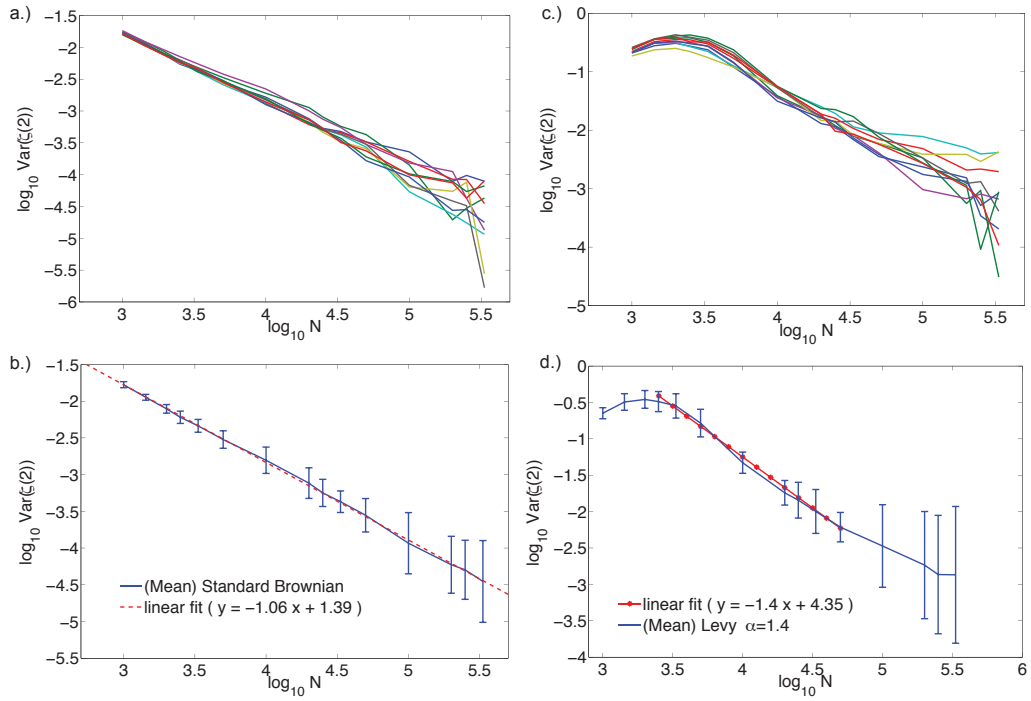

Figure 3: (Color online) Plots showing how errors can be ascribed to the plots in Figure 2. The top plots show the results of the study for 10 different randomly seeded realisations of sample size $N = 10^6$ for a.) a standard Brownian motion and b.) an α-stable Lévy motion (α=1.4). Plots c.) and d.) are the mean averages of the realisations in a.) and b.) respectively, where the error bars are calculated from the maximum deviation from this mean in the 10 realisations.

and consistent properties of such techniques in the case of heavy-tailed distributions [43–45] and especially infinite-variance processes are still unclear we envisage the use of such methods in future research.

Finally, one could in principle increase the available number of values of $\zeta_i(2)$ by overlapping intervals of size N within a given single realisation. We have, however, found that this introduces a significant systematic bias in the computed values of $Var[\zeta(2)]$.

Real-world data

We calculate $Var[\zeta(2)]$ values for the examples of real-world data sets discussed earlier in the introduction. The plot detailing this study is shown in Figure 4. For comparison we have also included on this plot the variation of $Var[\zeta(2)]$ with N for the two archetypal cases of finite and infinite variance processes in the form of a standard Brownian motion and an α-stable Lévy process ($\alpha = 1.4$); we also plot a negative unit slope for the asymptotic $N \to \infty$ behaviour obtained from large sample theory, this is indicated by the dashed line. Figure 4 shows that the real-world data can show significant departures from the synthetic data.

The WIND data illustrates the effect of large data gaps which are not present in the ACE data shown; this limits the amount of data available for certain N which is reflected in the corresponding estimations of $Var[\zeta(2)]$. For the ACE B^2 data we can see a clear systematic departure from the synthetic models. We will discuss this latter data set in the next section below.

The problem with a single length N realisation is that we cannot calculate the errors on $Var[\zeta(2)]$ as done in the previous section; and thus have no way of gauging how close these graphs are to the expected asymptotic behaviour predicted by large-sample theory. However, one can still estimate an error for measurements of $\zeta(2)$ obtained from a finite data size N, in the same way as was done in equations (7)-(9). For example, in the case of the ACE B^2 data this would indicate that a $N \sim 10^5$ sample size would introduce an error of $\sim 12\%$ in the estimated values of $\zeta(2)$ using the iteratively conditioned moment scaling technique.

A. Underlying statistics of $\zeta(2)$

We plot in Figure 5 the PDFs $H(\zeta(2))$ for three of the representative models we have studied along with the PDFs $H(\zeta(2))$ for one of the real-world data sets. For each of these time series, PDFs have been constructed for two different sample sizes N. We see that apart from the α-stable case, these PDFs are well described by a Gaussian distribution, as can be seen by the Maximum Likelihood fits. The α-stable Lévy case is shown in Fig-

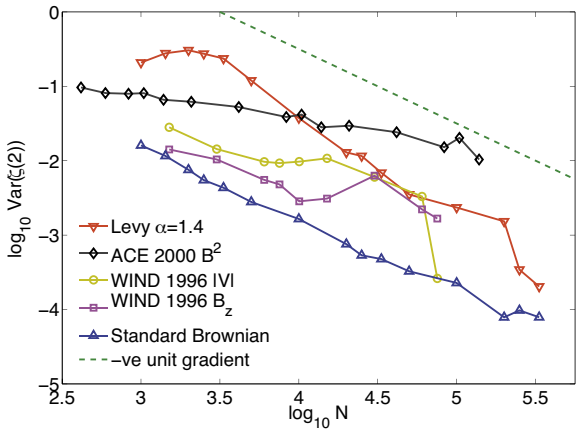

Figure 4: (Color online) Plot of the variance of conditioned $\zeta(2)$ with sample size N for all the real-world data sets studied shown on a log-log plot. The dashed line on both of these plots indicates a negative slope of unit gradient. Also included for comparison are the archetypal synthetic data sets for the finite variance and infinite variance processes.

ures 5b i.) and b ii.) to be well described by a Gumbel max-stable Extreme Value distribution [46]. To see why nearly all of our PDFs corresponding to finite variance processes are close to Gaussian we appeal to large sample-theory.

To facilitate understanding we employ more heuristic arguments at the expense of mathematical rigour. Interested readers can find more on the mathematical details and proofs in [16, 47] which deal with spectral parameter estimates of strong long-range dependent Gaussian stationary time series; [29] for a non-stationary generalization of these; [21] for a pseudo-variogram estimator (similar to the method in this paper) to long-range dependent Gaussian processes with stationary increments; and the more recent and extensive paper by Stoev, Pipiras and Taqqu [20] which extends the proofs and arguments of [21] to infinite variance processes in the form of α-stable and linear fractional stable processes. This latter reference will be our main source and point of contact for what follows. A survey of many of these papers and parameter estimation techniques can be found in [17].

As mentioned above, we have chosen Stoev *et. al.* [20] as a point of contact from amongst the extensive literature concerning asymptotic large sample behaviour of parameter estimates. This is primarily because this work has dealt with infinite variance processes of the type discussed here; also our moment scaling technique is a particular form of one of the main estimators used in [20] (see also [21]). We have also used Stoev's MATLAB algorithm for generating the LFSM realisations used in this paper. Similar to our study Stoev *et. al.* use a least squares regression on the moments which they refer to as the 'power' estimator. However, instead of taking moments of the increments as we do, they take the moments of coefficients for a Finite Impulse Response Transformation (FIRT) of the discrete time-series, which is characterised by a discrete filter of n members. Our increments are one of the simplest forms of these FIRT coefficients if we take the filter to be comprised of a set of $n = 2$ members $\{-1, 1\}$. However, any extra benefits of having more than one zero-moment (moments which are equal to zero) will be lost due to this simplicity. This also applies to the wavelet coefficients used in the study of Stoev *et. al.* where our increments result from taking the mother wavelet to be the superposition of two delta functions (one positive, one negative) separated by a scale τ – also known in the literature as the 'poor man's wavelet' [48]. Also, an important fact to note when comparing the methods of Stoev *et. al.* to ours is that we differ with the 'power' method of these authors by using the iterative conditioning technique which by censoring and excluding the poorly sampled large extreme events, correct for the bias which is pathological in the case of heavy-tailed distributions [27].

Recall that the second order moment is scaling as $M_i^2(\tau) = M_i^2(1)\tau^{\zeta(2)}$ and we will be estimating $\zeta(2)$ from the gradient of a log-log plot

$$\log M_i^2(\tau) = \zeta(2)\log\tau + \log M_i^2(1)\ . \tag{10}$$

For the discrete data the gradient can be estimated via least squares linear regression, and the problem can be set out as

$$\mathbf{M^2_{\log}} = \mathbf{T}_{\log}\mathbf{Z} + \frac{1}{\sqrt{N}}\epsilon\ , \tag{11}$$

where

$$\mathbf{M^2_{\log}} = \begin{pmatrix} \log M_i^2(\tau_1) \\ \vdots \\ \log M_i^2(\tau_k) \end{pmatrix}\ , \tag{12}$$

is the vector of the observations (or dependent variables);

$$\mathbf{T}_{\log} = \begin{pmatrix} \log\tau_1 & 1 \\ \vdots & \vdots \\ \log\tau_k & 1 \end{pmatrix}\ , \tag{13}$$

contains the vector of the scales (or independent variables);

$$\mathbf{Z} = \begin{pmatrix} \zeta(2) \\ \log M_i^2(1) \end{pmatrix}\ , \tag{14}$$

is the vector of the parameters needed to be estimated; and

$$\epsilon = \begin{pmatrix} \sqrt{N}\left(\log M_i^2(\tau_1) - \log\hat{M}_i^2(\tau_1)\right) \\ \vdots \\ \sqrt{N}\left(\log M_i^2(\tau_k) - \log\hat{M}_i^2(\tau_k)\right) \end{pmatrix}\ , \tag{15}$$

is the vector of estimation errors between the sample measurements and those of the true expected values (denoted by $\hat{M}$) for which the scaling relation in (10) actually holds. The solution to equation (11) is then given by

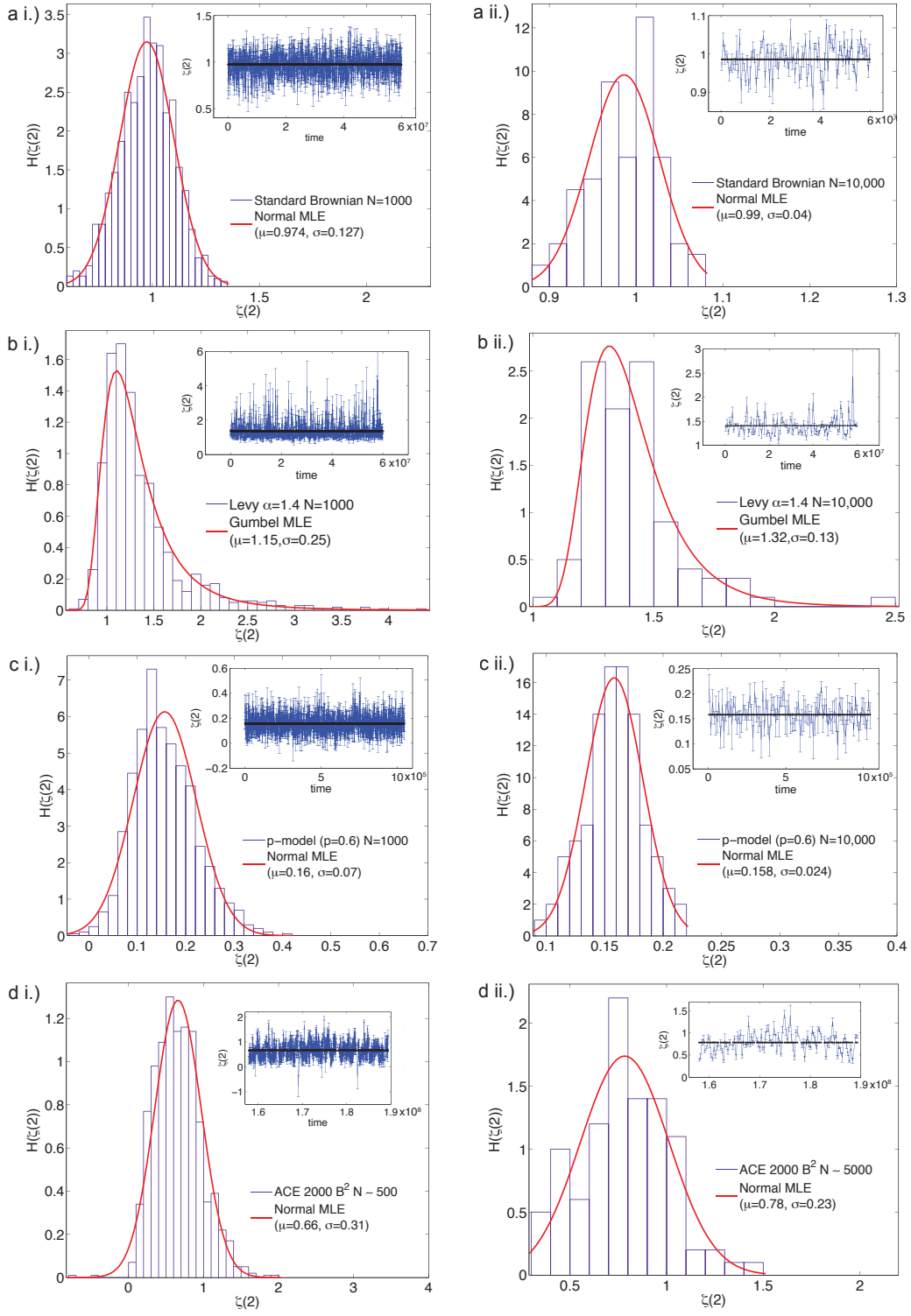

Figure 5: (Color online) PDF's $H(\zeta(2))$ obtained for (a) a standard Brownian motion, (b) α-stable Lévy process ($\alpha = 1.4$), (c) p-model (p=0.6) and (d) ACE 2000 B^2 for two different sample sizes (i) $N = 1000$ ($N = 500$ for ACE) and (ii) $N = 10,000$ ($N = 5000$ for ACE). The sample PDF's are overlaid with Maximum Likelihood Estimate (MLE) model fits of a Normal distribution for a, c and d; and Gumbel max-stable (extreme value type I) fits for b. For both these models μ is the location parameter and σ is the scale parameter, which for the Normal distribution coincide with the mean and standard deviation. The samples of $\zeta_i(2)$ (varying with time t) from which these PDFs are constructed, are shown in the corresponding inserts of each plot.

the well known ordinary linear least squares estimator to the parameters as

$$\mathbf{Z} = \left(\mathbf{T}^{\mathbf{t}}_{\log}\mathbf{T}_{\log}\right)^{-\mathbf{1}}\mathbf{T}^{\mathbf{t}}_{\log}\mathbf{M}^{\mathbf{2}}_{\log} \ , \tag{16}$$

where superscript t represents a matrix transpose. (16) is simply a linear combination of the dependent variable $\log M_i^2(\tau)$ i.e. a sum, so that for $\zeta_i(2)$ one can write the ordinary least squares estimate as

$$\zeta_i(2) = \sum_{j=1}^{k} a_j \left(\log M_i^2(\tau_j)\right) \ , \tag{17}$$

where the a_j are all the elements of the appropriate vector from $\left(\mathbf{T}^{\mathbf{t}}_{\log}\mathbf{T}_{\log}\right)^{-\mathbf{1}}\mathbf{T}^{\mathbf{t}}_{\log}$. We will return to this form of the ordinary least squares solution shortly.

Adapted to the notation used in this paper, Theorem 3.1 in [20] states that if $\zeta(2)$ is the FIRT coefficient estimator (using least squares regression) for the scaling exponent and $\hat{\zeta}(2)$ the true expected value; then as $\lim N \to \infty$

$$\sqrt{N}\left(\zeta(2) - \hat{\zeta}(2)\right) \to \mathcal{N}\left(0, \sigma^2\right) \tag{18}$$

where $\mathcal{N}\left(0, \sigma^2\right)$ is a Normal distribution with mean 0 and variance σ^2. Strictly speaking this theorem requires that the FIRT coefficients obey the following inequality between the number of zero-moments of the FIRT filter, the self-similarity parameter H and the tail (stability) index α

$$Q > H + \frac{1}{\alpha\left(\alpha - 1\right)} \ , \tag{19}$$

which in the case of the ordinary Brownian motion and the α-stable Lévy processes, where $H = 1/\alpha$, generalises to

$$Q > \frac{1}{\left(\alpha - 1\right)} \ . \tag{20}$$

The moment scaling scheme based on the raw increments has only one zero moment, hence $Q = 1$; and as a result, except for the p-model for which α and H are unknown (or not applicable), the above inequality does not hold for any of the synthetic models. However as mentioned in [21] for Gaussian processes, where in equation (19) $\alpha = 2$, the $Q = 1$ case is sufficient for the above theorem to hold as long as $H < 0.75$. For our fBm model $H = 0.8$ and we find that the results of the theorem (18) still hold. Thus we can conjecture that the criterion in (19) and (20) can be relaxed a little so that the inequality becomes an approximate inequality. Also, Stoev *et. al.* [20] show via simulations that the estimators continue to work well even when the criterion in (19) and (20) are not satisfied. The essential reason why this criteria was initially introduced, was so that the estimator could distinguish between actual long-memory effects and trends [21].

We now consider the exception that we have found to this behaviour – that of the infinite variance processes at finite N. We would expect that in the $N \to \infty$ limit the results of the above theorem will also hold for the infinite variance processes. However, we believe that in this case the convergence will be slow and will depend upon the number of scales τ that were used to conduct our linear least squares regression.

We will now go beyond the above asymptotic arguments to discuss the non-Gaussian intermediate finite N behaviour that we see here in Figure 5 (b-i) and (b-ii). Recall the expression for $M_i^2(\tau)$ given in equation (5) (for $p = 2$). We will discuss in detail here the case where the sum in (5) consists of *i.i.d.* random variables; this is the case for some of our synthetic time series – these arguments can be developed for other cases. The PDF of this sum will by the Central Limit Theorem tend to a Gaussian and for finite N will probably take the form of a Pearson χ^2_ν type variable with ν degrees of freedom (see [21] for more details). However, for an infinite variance process, the sum in (5) will be dominated by the largest extreme events, which in some cases can be of the order of the rest of the sum [32]. This will still be the case even when we have excluded some of these extreme events due to the iterative conditioning scheme. Thus the sum will be distributed as the extreme values of an α-stable Lévy distribution – which is given by a max-stable Frechét distribution (see [27]). Note that this will be the case for any N, even in the asymptotic large N case. Without too much detail the form of the PDF $P\left(M_i^2\right)$ of M_i^2 will be of the type

$$P\left(M_i^2\right) = \frac{\Lambda}{2\left(M_i^2\right)^{1+\alpha/2}} \exp\left(-\frac{\Lambda}{\alpha\left(M_i^2\right)^{\alpha/2}}\right) \ , \tag{21}$$

where any scale parameters have been absorbed into the Λ. One can then convert this Frechét PDF to a PDF $\tilde{P}\left(M^2_{i,\log}\right)$ corresponding to the dependent variable $\log M_i^2(\tau)$ in the least squares scheme in (17), which under a simple conservation of probability will be given by

$$\tilde{P}\left(M^2_{i,\log}\right) = \frac{\Lambda}{2}\exp\left(-\frac{\alpha}{2}M^2_{i,\log} - \frac{\Lambda}{\alpha}\exp\left(-\frac{\alpha}{2}M^2_{i,\log}\right)\right) \ , \tag{22}$$

which is in the form of a Gumbel extreme value distribution; this is another max-stable distribution [49]. The Gumbel max-stable PDF has a long slow exponentially decreasing right tail; this will imply that a sum of random variables such as (17), each distributed with this PDF, will eventually tend to a Gaussian distributed random variable, but slowly due to the heavy right tail. This then captures our result in Figures 5 (b-i) and (b-ii), and may also explain why we do not obtain the $\sim N^{-1}$ behaviour in the plots of $\log Var(\zeta(2))$ Vs. $\log N$ for the α-stable Lévy cases.

As discussed above, for the case of the finite variance synthetic time series the M_i^2 will be well described by a

Gaussian or (more realistically for finite N) a Pearson χ^2_ν PDF. In the same way as was done with the infinite variance processes above, it can be shown that the PDF of $\log M_i^2$ can be written as a Gumbel min-stable PDF. Gumbel min-stable PDF's have long slow exponentially decreasing left tails, which in our case will be limited by the fact that $\zeta(2) > 0$. The right tail of Gumbel min-stable PDFs is more compact with a rapidly decaying exponential tail. Due to the more compact nature of the PDF, a sum of $\log M_i^2$ variables will tend to a Gaussian under the Central Limit Theorem much faster than the infinite variance processes above, hence explaining why the PDFs $H(\zeta(2))$ for the finite variance processes are well described by a Gaussian.

Finally, there is the open question of the behaviour of the real-world data. The ACE B^2 data PDFs of $H(\zeta(2))$ show that they can be well described by a Gaussian; however the scaling of $Var(\zeta(2))$ with N using our estimation shows a significant deviation from the N^{-1} behaviour implied by (18). This will be the topic of future work.

IV. CONCLUSIONS

We have investigated finite-sample size (N) effects on quantifying the statistical scaling properties of time series. We focus on the scaling exponent $\zeta(2)$ of the variance or second moment which for a wide class of processes also gives the spectral exponent β of the (power law) power spectrum. If too small a sample size is used then these fluctuations effectively introduce a pseudo-nonstationarity in the estimates for the scaling exponents. To achieve an error in the exponent of less than $\sim 5\%$, we find that the number of data points N needed varies significantly with the details of the underlying process and is in the range of $10^3 - 10^5$ for the synthetic models used in this paper. The variance in the exponent when computed from an interval of N samples is known to vary as $\sim 1/N$ for $N \to \infty$; however, the convergence to this behaviour will also depend on the details of the process and more importantly on the parameter estimating technique used. In particular we have shown that heavy tailed processes can be far from this limiting behaviour for observationally realisable N.

We have also considered the case where the scaling exponents are time independent, but where there is a secular time dependence in other parameters such as the standard deviation. For the case of a Brownian motion, the estimate of the scaling exponent is not affected by this time dependence. It may thus be too premature to reject a time series for scaling analysis simply because of the non-stationarity of certain parameters i.e. a running mean or standard deviation. This also highlights the need to distinguish time variation in the moments from that in scaling exponents that are derived from them.

We have focussed here on the moment order scaling technique to calculate the scaling exponents in order to highlight the issue of apparent non-time stationarity. Although there exists extensive statistics literature on the asymptotic $N \to \infty$ limit of various estimation techniques, further work is needed to investigate how these details pass over to the more pressing and pragmatic need for their implications on quantifying scaling in finite and realisable N-sized samples.

Acknowledgments

The authors would like to thank B. Hnat, G. Rowlands, F. M. Poli, T. Dudok De Wit and V. Keinhorst for helpful discussions and suggestions. We acknowledge the financial support of the UK STFC and EPSRC; and R. P. Lepping and K. Oglivie for ACE and WIND data.

[1] U. Frisch, *Turbulence* (Cambridge University Press, 1995).
[2] J. P. Sethna, K. A. Dahmen, and C. R. Myers, Nature **410**, 242 (2001).
[3] D. Sornette, *Critical Phenomena in Natural Sciences* (Springer-Verlag, 2000).
[4] S. B. Pope, *Turbulent Flows* (Cambridge University Press, 2000).
[5] B. Hnat, S. C. Chapman, and G. Rowlands, Physics of Plasmas **11**, 1326 (2004).
[6] C. W. Smith, K. Hamilton, B. J. Vasquez, and R. J. Leamon, Astrophys. J. **645**, L85 (2006).
[7] A. Retinò, D. Sundkvist, A. Vaivads, F. Mozer, M. André, and C. J. Owen, Nature Physics **3**, 236 (2007).
[8] D. Sundkvist, A. Retinò, A. Vaivads, and S. D. Bale, Phys. Rev. Lett. **99**, 025004 (2007).
[9] L. Sorriso-Valvo, R. Marino, V. Carbone, A. Noullez, F. Lepreti, P. Veltri, R. Bruno, B. Bavassano, and E. Pietropaolo, Phys. Rev. Lett. **99**, 115001 (2007).
[10] R. M. Nicol, S. C. Chapman, and R. O. Dendy, Astrophys. J. **679**, 862 (2008).
[11] D. Jankovicova, Z. Voros, and J. Simkanin, Nonlin. Proc. Geophys. **15**, 53 (2008).
[12] G. A. Abel and M. P. Freeman, J. Geophys. Res. **107**, 1470 (2002).
[13] S. Mallat, *A wavelet tour of signal processing* (Academic Press Inc., 1999).
[14] D. Popivanov and A. Mineva, Mathematical Biosciences **157**, 303 (1999).
[15] A. Leontitsis and C. E. Vorlow, Physica A **368**, 522 (2006).
[16] P. M. Robinson, Ann. Stat. **23**, 1048 (1995).
[17] J. Beran, *Statistics for Long-Memory Processes* (Chapman & Hall, 1994).
[18] J. F. Muzy, E. Bacry, and A. Arneodo, Phys. Rev. E **47**, 875 (1993).
[19] P. Abry, P. Flandrin, M. Taqqu, and D. Veitch, *Theory and applications of long-range dependence* (Birkhauser,

2003), p. 527.
[20] S. Stoev, V. Pipiras, and M. S. Taqqu, Signal Proc. **82**, 1873 (2002).
[21] J.-M. Bardet, J. Time Series Analysis **21**, 497 (2000).
[22] G. Samorodnitsky and M. S. Taqqu, *Stable non-Gaussian random processes* (Chapman & Hall, 1994).
[23] S. Stoev and M. S. Taqqu, Adv. Appl. Prob. **36**, 1085 (2004).
[24] S. Stoev and M. S. Taqqu, Fractals **12**, 95 (2004).
[25] C. Meneveau and K. R. Sreenivasan, Phys. Rev. Lett. **59**, 1424 (1987).
[26] C. Meneveau, K. R. Sreenivasan, P. Kailasnath, and M. S. Fan, Phys. Rev. A **41**, 894 (1990).
[27] K. Kiyani, S. C. Chapman, and B. Hnat, Phys. Rev. E **74**, 051122 (2006).
[28] R. Weron, Physica A **312**, 285 (2002).
[29] C. Velasco, J. Econometrics **91**, 325 (1999).
[30] P. Embrechts and M. Maejima, *Selfsimilar Processes* (Princeton University Press, 2002).
[31] T. Dudok De Wit, Phys. Rev. E **70**, 055302(R) (2004).
[32] F. Bardou, J. Bouchaud, A. Aspect, and C. Cohen-Tannoudji, *Lévy Statistics and Laser Cooling* (Cambridge University Press, 2002).
[33] H. Kantz and T. Schreiber, *Nonlinear Time Series Analysis* (Cambridge University Press, 1997).
[34] P. Abry and D. Veitch, IEEE Trans. Inf. Th. **44**, 2 (1998).
[35] S. Siegert and R. Friedrich, Phys. Rev. E **64** (2001).
[36] C. Pagel and A. Balogh, J. Geophys. Res. **107**, 1178 (2002).
[37] M. P. Freeman, N. W. Watkins, and D. J. Riley, Phys. Rev. E **62**, 8794 (2000).
[38] B. Hnat, S. C. Chapman, and G. Rowlands, J. Geophys. Res. **110** (2005).
[39] B. Hnat, S. C. Chapman, G. Rowlands, N. W. Watkins, and W. M. Farrell, Geophys. Res. Lett. **29** (2002).
[40] K. Kiyani, S. C. Chapman, B. Hnat, and R. M. Nicol, Phys. Rev. Lett. **98**, 211101 (2007).
[41] H. Wendt and P. Abry, IEEE Trans. Sig. Proc. **55**, 4811 (2007).
[42] A. M. Zoubir and B. Boashash, IEEE Sig. Proc. Mag. **15**, 56 (1998).
[43] P. Hall, Ann. Prob. **18**, 1342 (1990).
[44] K. B. Athreya, Ann. Stat. **15**, 724 (1987).
[45] K. Knight, Ann. Stat. **17**, 1168 (1989).
[46] E. J. Gumbel, *Statistics of Extremes* (Columbia University Press, 1967).
[47] R. Fox and M. S. Taqqu, Ann. Stat. **14**, 517 (1986).
[48] M. Vergassola and U. Frisch, Physica D **54**, 58 (1991).
[49] S. C. Chapman, G. Rowlands, and N. W. Watkins, Nonlin. Proc. Geophys. **9**, 409 (2002).