AGU PUBLICATIONS

Water Resources Research

RESEARCH ARTICLE
10.1002/2015WR016998

Key Points:
- A new statistical scaling model is developed, explored and applied
- Apparent inconsistency between variable and increment statistics is eliminated
- Parameters are estimated using variable and increment moments jointly

Supporting Information:
- Supporting Information S1

Correspondence to:
M. Riva,
monica.riva@polimi.it

Citation:
Riva, M., S. P. Neuman, and A. Guadagnini (2015), New scaling model for variables and increments with heavy-tailed distributions, *Water Resour. Res.*, *51*, 4623–4634, doi:10.1002/2015WR016998.

Received 30 JAN 2015
Accepted 23 MAY 2015
Accepted article online 30 MAY 2015
Published online 26 JUN 2015

New scaling model for variables and increments with heavy-tailed distributions

Monica Riva[1,2], Shlomo P. Neuman[2], and Alberto Guadagnini[1,2]

[1]Dipartimento di Ingegneria Civile e Ambientale, Politecnico di Milano, Milano, Italy, [2]Department of Hydrology and Water Resources, University of Arizona, Tucson, Arizona, USA

Abstract Many hydrological (as well as diverse earth, environmental, ecological, biological, physical, social, financial and other) variables, Y, exhibit frequency distributions that are difficult to reconcile with those of their spatial or temporal increments, ΔY. Whereas distributions of Y (or its logarithm) are at times slightly asymmetric with relatively mild peaks and tails, those of ΔY tend to be symmetric with peaks that grow sharper, and tails that become heavier, as the separation distance (lag) between pairs of Y values decreases. No statistical model known to us captures these behaviors of Y and ΔY in a unified and consistent manner. We propose a new, generalized sub-Gaussian model that does so. We derive analytical expressions for probability distribution functions (pdfs) of Y and ΔY as well as corresponding lead statistical moments. In our model the peak and tails of the ΔY pdf scale with lag in line with observed behavior. The model allows one to estimate, accurately and efficiently, all relevant parameters by analyzing jointly sample moments of Y and ΔY. We illustrate key features of our new model and method of inference on synthetically generated samples and neutron porosity data from a deep borehole.

1. Introduction

Traditional geostatistical models consider data to represent samples of multivariate Gaussian functions. Yet many earth, environmental, ecological, biological, physical, social, financial and other variables, Y, exhibit frequency distributions that are difficult to reconcile with those of their spatial or temporal increments, ΔY. Whereas distributions of Y (or its logarithm) are at times slightly asymmetric with relatively mild peaks and tails, those of ΔY tend to be symmetric with peaks that grow sharper, and tails that become heavier, as the separation distance (lag) between pairs of Y values decreases. Documented examples include porosity [*Painter*, 1996; *Guadagnini et al*., 2014, 2015], permeability [*Painter*, 1996; *Riva et al*., 2013a, 2013b] and hydraulic conductivity [*Liu and Molz*, 1997; *Meerschaert et al*., 2004; *Guadagnini et al*., 2013], electrical resistivity [*Painter*, 2001; *Yang et al*., 2009], soil and sediment texture [*Guadagnini et al*., 2014], sediment transport rate [*Ganti et al*., 2009], rainfall [*Kumar and Foufoula-Georgiou*, 1993], measured and simulated turbulent fluid velocity [*Castaing et al*., 1990; *Boffetta et al*., 2008], and magnetic fluctuation [*von Papen et al*., 2014] data. No statistical model known to us captures these behaviors of Y and ΔY in a unified and consistent manner. A step in that direction is represented by the fractional Laplace motion model of *Meerschaert et al*. [2004] and *Kozubowski et al*. [2006, 2013], which transitions automatically from heavy-tailed to Gaussian with increasing lag.

Heavy tails are important because they control the distributions of extreme values, which are of central interest in hydrology and many other fields [*Katz et al*., 2002; *Riva et al*., 2013c]. We propose a new statistical model that reconciles the behaviors of variables and increments possessing heavy-tailed distributions, the tails and peaks of which scale with lag in the above manner. Let the variable of interest be a stationary random function $Y(x)=\langle Y\rangle+Y'(x)$ defined on a continuum of points, x, in Euclidean space (or time) where $\langle Y\rangle$ is a constant ensemble mean (expectation) and $Y'(x)$ a zero-mean random fluctuation about $\langle Y\rangle$. One way to model heavy tail behavior of $Y'(x)$ is to write it in standard sub-Gaussian form, $Y'(x)=UG(x)$, where $G(x)$ is a zero-mean stationary Gaussian function and U is a nonnegative random variable independent of $G(x)$ [e.g., *Samorodnitsky and Taqqu*, 1994]. This form subordinates $Y'(x)$ to $G(x)$ through action of the subordinator U, rendering it a scale mixture of Gaussian functions with random variances proportional to U^2. The scale mixture is non-Gaussian with distribution that depends on that of U. Yet any single realization of $Y'(x)$,

obtained upon multiplying one realization of $G(x)$ by one draw from the distribution of U, is Gaussian. The standard sub-Gaussian form is thus nonergodic, having different distributions in probability (ensemble) and real spaces. It is therefore incompatible with any single spatial (or temporal) sample (realization) of $Y'(x)$ having a non-Gaussian frequency distribution. For this reason we [*Neuman et al*., 2013; T. Nan et al., Analyzing randomly fluctuating hierarchical variables and extremes, submitted to *Handbook of Groundwater Engineering*, edited by J. Cushman and D. M. Tartakovsky, CRC Press, N. Y., 2015] previously explored various properties of standard sub-Gaussian functions by considering not one but a large number of such samples simultaneously.

To allow dealing with unique non-Gaussian data sets we introduce in this paper a generalized sub-Gaussian model

$$Y'(x)=U(x)G(x) \tag{1}$$

in which $U(x)$ is now a function (not a variable), independent of $G(x)$, consisting of independent and identically distributed (iid) nonnegative values at all points x. Some authors have previously used (1) to generate synthetic samples of non-Gaussian random functions [*Georgiou and Kyriakakis*, 2006; *Neuman*, 2011; *Guadagnini et al*., 2012; *Riva et al*., 2013a], without, however, having at their disposal a formal theory relating the statistical properties of $U(x)$ and $G(x)$ to those of $Y'(x)$ and its increments. The purpose of this paper is to derive and explore such relationships with the aim of developing and demonstrating a method of statistical inference based on (1). Correspondingly, we derive in section 2 analytical expressions for bivariate (two-point) and marginal distributions of $Y'(x)$ together with associated lead moments for the case of lognormal $U(x)$; use them in section 3 to develop analytical expressions for marginal distributions of $\Delta Y(x;s)$ and the variogram of $Y(x)$ as functions of lag, s; explore and demonstrate these results numerically on synthetically generated samples, as well as propose and test a new method of statistical inference based on (1), in section 4; and illustrate our newly proposed theory and method of inference on neutron porosity data from a deep borehole in section 5.

2. Probability Distributions and Lead Moments of Y'

We introduce the following notation to define Y' at two points, x_1 and x_2,

$$Y'(x_1)=U(x_1)G(x_1)=Y_1=U_1G_1, \qquad Y'(x_2)=U(x_2)G(x_2)=Y_2=U_2G_2. \tag{2}$$

The bivariate probability distribution of Y_1 and Y_2 is

$$F_{Y_1,Y_2}(y_1,y_2)=P(Y_1\le y_1,Y_2\le y_2)=\int\int_{D_1}\int\int_{D_2} f_{U_1U_2G_1G_2}(u_1,u_2,g_1,g_2)du_1du_2dg_1dg_2, \tag{3}$$

where $D_i=\{(u_i,g_i)\in R^2: u_ig_i\le y_i\}$ for $i=1, 2$ and $f_{U_1U_2G_1G_2}$ is the joint probability density function (pdf) of (U_1,U_2,G_1,G_2). Since the random variables U_1 and U_2 are independent of each other and of G_i (3) simplifies to

$$F_{Y_1,Y_2}(y_1,y_2)=\int\int_{D_1}\int\int_{D_2} f_{U_1}(u_1)f_{U_2}(u_2)f_{G_1G_2}(g_1,g_2)du_1du_2dg_1dg_2. \tag{4}$$

Here $f_{U_i}(u_i)$ is the pdf of U_i and $f_{G_1G_2}$ is the bivariate pdf of (G_1,G_2), i.e.,

$$f_{G_1G_2}(g_1,g_2)=\frac{1}{2\pi\sigma_G^2\sqrt{1-\rho_G^2}}e^{-\frac{g_1^2+g_2^2-2\rho_G g_1 g_2}{2\sigma_G^2\left(1-\rho_G^2\right)}}, \tag{5}$$

where σ_G^2 is the variance of G and ρ_G is the coefficient of correlation between G_1 and G_2, a function of lag, $s=|x_1-x_2|$. As U_1 and U_2 are nonnegative (4) reduces to

$$F_{Y_1,Y_2}(y_1,y_2)=\int_{u_1=0}^{\infty}\int_{u_2=0}^{\infty}\int_{g_1=-\infty}^{y_1/u_1}\int_{g_2=-\infty}^{y_2/u_2} f_{U_1}(u_1)f_{U_2}(u_2)f_{G_1G_2}(g_1,g_2)dg_1dg_2du_2\,du_1. \tag{6}$$

The bivariate pdf of Y_1 and Y_2 is

$$f_{Y_1,Y_2}(y_1,y_2)=\frac{\partial^2}{\partial y_1\partial y_2}F_{Y_1,Y_2}(y_1,y_2)=\int_{u_1=0}^{\infty}\int_{u_2=0}^{\infty}\frac{1}{u_2u_1}f_{U_1}(u_1)f_{U_2}(u_2)f_{G_1G_2}\left(\frac{y_1}{u_1},\frac{y_2}{u_2}\right)du_2\,du_1. \tag{7}$$

Clark [1973] and *Guadagnini et al.* [2015] found some financial and environmental variables to be well represented by normal-lognormal (NLN) distributions. Correspondingly, we consider U_1 and U_2 to be lognormally distributed according to $\ln N\left(0,(2-\alpha)^2\right)$, i.e.,

$$f_{U_i}(u_i)=\frac{1}{\sqrt{2\pi}u_i(2-\alpha)}e^{-\frac{\ln^2 u_i}{2(2-\alpha)^2}},\qquad \text{with } i=1,\ 2 \quad \text{and } \alpha<2; \tag{8}$$

other potential distributions of U that, like the lognormal, possess finite moments of all orders include the exponential, Weibull and gamma. We show that (8) renders the marginal distribution of Y' NLN. Substituting (5) and (8) into (7) yields the following bivariate pdf of Y_1 and Y_2,

$$f_{Y_1,Y_2}(y_1,y_2)=\frac{1}{4\pi^2(2-\alpha)^2}\frac{1}{\sqrt{1-\rho_G^2}}\int_0^{\infty}\int_0^{\infty}\frac{1}{u_2{}^2u_1{}^2}e^{-\frac{1}{2}\left[\frac{1}{(2-\alpha)^2}\left(ln^2\frac{u_1}{\sigma_G}+\ln^2\frac{u_2}{\sigma_G}\right)+\frac{1}{1-\rho_G^2}\left(\frac{y_1{}^2}{u_1{}^2}+\frac{y_2{}^2}{u_2{}^2}-2\rho_G\frac{y_1y_2}{u_1u_2}\right)\right]}du_2du_1. \tag{9}$$

The marginal pdf of Y' follows from (9) in the form

$$f_{Y'}(Y')=\int_{-\infty}^{\infty}f_{Y_1,Y_2}(y_1,y_2)dy_1=\frac{1}{2\pi(2-\alpha)}\int_0^{\infty}\frac{1}{u^2}e^{-\frac{1}{2}\left(\frac{1}{(2-\alpha)^2}\ln^2\frac{u}{\sigma_G}+\frac{y^2}{u^2}\right)}du, \tag{10}$$

which coincides with NLN [e.g., *Guadagnini et al.*, 2015]. Note that when $\alpha\to 2$ the lognormal distribution (8) tends to a delta function so that (10) coincides with the Gaussian distribution.

As (10) is symmetric, Y' moments of odd orders vanish and those of even orders q become

$$\langle Y'^q\rangle=\int_{-\infty}^{\infty}y'^qf_{Y'}(y')dy'=\frac{2^{\frac{q}{2}}}{\sqrt{\pi}}\Gamma\left(\frac{q+1}{2}\right)e^{\frac{(2-\alpha)^2q^2}{2}}\sigma_G^q, \tag{11}$$

where Γ is the gamma function. In particular the variance and kurtosis of Y' become, respectively,

$$\sigma_Y^2=\langle Y'^2\rangle=e^{2(2-\alpha)^2}\sigma_G^2, \tag{12}$$

$$\langle Y'^4\rangle=3e^{8(2-\alpha)^2}\sigma_G^4. \tag{13}$$

A global measure of how sharp is the peak and how heavy are the tails of $f_{Y'}$ is provided by the standardized kurtosis

$$\kappa_Y=\frac{\langle Y'^4\rangle}{\langle Y'^2\rangle^2}=3e^{4(2-\alpha)^2}. \tag{14}$$

Since $\kappa_Y\geq 3$ the pdf $f_{Y'}$ is leptokurtic, tending to the Gaussian distribution with $\kappa_Y=3$ as $\alpha\to 2$.

3. Probability Distribution and Variogram of Y Increments

We define increments $\Delta Y(s=|x_1-x_2|)=Y_1-Y_2$ for lags, s, equal to or larger than the measurement and/or resolution scale of given data; this scale, represented formally by a lower cutoff λ_I, is introduced in (23) – (24) below. The probability distribution of increments associated with a given lag is

$$F_{\Delta Y}(\Delta y)=P(\Delta Y\leq\Delta y)=\iint_{y_1-y_2\leq\Delta y}f_{Y_1Y_2}(y_1,y_2)dy_1dy_2=\int_{y_2=-\infty}^{\infty}\int_{y_1=-\infty}^{\Delta y+y_2}f_{Y_1Y_2}(y_1,y_2)dy_1dy_2, \tag{15}$$

the corresponding pdf being

$$f_{\Delta Y}(\Delta y)=\frac{d}{d(\Delta y)}F_{\Delta Y}(\Delta y)=\int_{-\infty}^{\infty} f_{Y_1Y_2}(\Delta y+y_2,y_2)dy_2. \tag{16}$$

Substituting (9) into (16) yields

$$f_{\Delta Y}(\Delta y)=\frac{1}{2\pi^2(2-\alpha)^2}\sqrt{\frac{\pi}{2}}\int_0^\infty\int_0^\infty \frac{e^{-\frac{1}{2}\left[\frac{1}{(2-\alpha)^2}\left(ln^2\frac{u_1}{\sigma_G}+\ln^2\frac{u_2}{\sigma_G}\right)+\frac{(\Delta y)^2}{u_1^2+u_2^2-2u_1u_2\rho_G}\right]}}{u_2u_1\sqrt{u_1^2+u_2^2-2u_1u_2\rho_G}}\,du_2du_1, \tag{17}$$

from which one may derive lead statistical moments of ΔY. In particular, all moments of odd order vanish while the variance, kurtosis and standardized kurtosis are given, respectively, by

$$\langle\Delta Y^2\rangle=2\sigma_G^2e^{(2-\alpha)^2}\left[e^{(2-\alpha)^2}-\rho_G\right], \tag{18}$$

$$\langle\Delta Y^4\rangle=6\sigma_G^4e^{4(2-\alpha)^2}\left[e^{4(2-\alpha)^2}+1-4e^{(2-\alpha)^2}\rho_G+2\rho_G^2\right], \tag{19}$$

$$\kappa_{\Delta Y}=\frac{\langle\Delta Y^4\rangle}{\langle\Delta Y^2\rangle^2}=3e^{2(2-\alpha)^2}\left\{1+\frac{1}{2}\left(\frac{e^{2(2-\alpha)^2}-1}{e^{(2-\alpha)^2}-\rho_G}\right)^2\right\}. \tag{20}$$

As $\alpha\to 2$ (the distribution of Y' tends to the Gaussian), $\kappa_{\Delta Y}\to 3$ and the distribution of ΔY likewise tends to the Gaussian. Otherwise $\kappa_{\Delta Y}$ increases (the peak of $f_{\Delta Y}$ sharpens and its tails become heavier) with ρ_G. In other words, the shape of $f_{\Delta Y}$ scales with the correlation coefficient of G or, equivalently, with lag. Figure 1 illustrates how excess kurtosis, $\kappa_{\Delta Y}-3$, varies with ρ_G and α. At small lags (large ρ_G) $\kappa_{\Delta Y}-3$ exceeds zero by a significant margin, even at large values of α (when the pdf of Y' is near-Gaussian). Excess kurtosis decreases as ρ_G decreases (lag increases), rendering the peak of $f_{\Delta Y}$ less sharp and its tails lighter. When $\alpha\geq 1.8$, the asymptotic value of $\kappa_{\Delta Y}-3$ at large lags is very small ($<<1$) and $f_{\Delta Y}$ is virtually Gaussian.

Included in Figure 1 are horizontal lines depicting excess kurtosis of the pdf of Y', κ_Y-3. From (14) and (20) it follows that when $\alpha>2-\sqrt{\ln 3}\approx 0.95$, $f_{\Delta Y}$ at small lags has sharper peaks and heavier tails than does $f_{Y'}$, the opposite being true at large lags. When $\alpha<2-\sqrt{\ln 3}$ the pdf of Y' has higher peaks and heavier tails than the pdfs of ΔY regardless of lag. This behavior of $f_{\Delta Y}$ is indicated by Figure 1. Figure 2 depicts on arithmetic and semi logarithmic scales $f_{\Delta Y}$ for $\sigma_G=1.0$, $\alpha=1.8$ and three values of ρ_G. Also shown for comparison is a Gaussian distribution having the same mean and variance as ΔY. As noted earlier, $f_{\Delta Y}$ exhibits sharp peaks and heavy tails when $\rho_G=0.99$ (at small lag) and becomes virtually Gaussian as ρ_G decreases (lag increases).

The variogram of Y is obtained directly from (18) as

$$\gamma_Y=\frac{\langle\Delta Y^2\rangle}{2}=\sigma_G^2e^{(2-\alpha)^2}\left(e^{(2-\alpha)^2}-\rho_G\right)=\sigma_G^2e^{(2-\alpha)^2}\left(e^{(2-\alpha)^2}-1\right)+\gamma_Ge^{(2-\alpha)^2}, \tag{21}$$

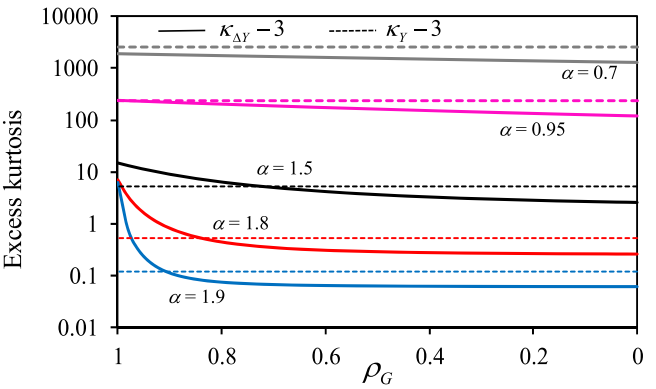

Figure 1. Excess kurtosis of ΔY (continuous curves) and of Y' (horizontal dashed lines) versus ρ_G for five values of α.

where γ_G is the variogram of G. Note that γ_Y includes a nugget effect (a constant independent of lag) that vanishes only in the Gaussian limit $\alpha\to 2$. From (21) and (12) we obtain an expression for the covariance of Y,

$$C_Y=\sigma_Y^2-\gamma_Y=e^{(2-\alpha)^2}C_G, \tag{22}$$

where $C_G=\sigma_G^2\rho_G$ is the covariance of G. This in turn yields the integral scale of Y, $I_Y=e^{-(2-\alpha)^2}I_G$, where I_G is the integral scale of G. It is thus seen that a log-normal subordinator dampens, but does not destroy, the covariance structure of G; the smaller is α the shorter is

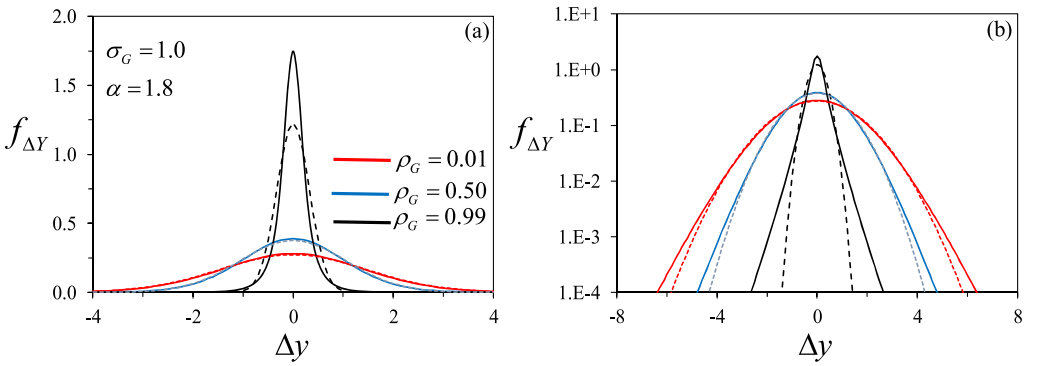

Figure 2. $f_{\Delta Y}$ (17) on (a) arithmetic and (b) semilogarithmic scales for $\sigma_G = 1.0$, $\alpha = 1.8$ and three values of ρ_G (continuous curves). Also shown are Gaussian distributions having the same mean and variance as ΔY (dashed curves).

the integral scale of Y. When $\alpha \to 2$, $I_Y \to I_G$. The integral scale of Y vanishes only in the limit as $\alpha \to -\infty$, i.e., when the variance of subordinator U tends to infinity.

4. Synthetic Examples

We generate synthetic realizations of $Y'(x)$ according to (1) at $N_e = 30{,}000$ discrete points spaced a distance $\delta = 10^{-4}$ (measured in arbitrary consistent length units) apart, on a line of length $l = 3$. As a first step we use a version of SGSIM [*Deutsch and Journel*, 1998] modified to generate a zero-mean stationary Gaussian random function, $G(x)$, constituting truncated fractional Brownian motion (tfBm) with truncated power variogram (TPV) [*Di Federico and Neuman*, 1997]

$$\gamma_G^2(s) = \gamma^2(s;\lambda_u) - \gamma^2(s;\lambda_l), \tag{23}$$

where

$$\gamma^2(s;\lambda_m) = \frac{A\lambda_m^{2H}}{2H}\left[1 - \exp\left(-\frac{s}{\lambda_m}\right) + \left(\frac{s}{\lambda_m}\right)^{2H}\Gamma\left(1-2H, \frac{s}{\lambda_m}\right)\right] \qquad m = l, u, \tag{24}$$

A is a coefficient, H is a Hurst scaling exponent, λ_l and λ_u being lower and upper cutoff scales proportional, respectively, to the resolution and sampling domain scales of given data. In the limits as $\lambda_l \to 0$ (perfect data resolution) and $\lambda_u \to \infty$ (infinite sampling domain) $G(x)$ becomes (nonstationary) fBm. Our samples are generated with $A = 1$, $H = 0.33$, $\lambda_l = 10^{-4}$ and $\lambda_u = 1$, which in turn yield $\sigma_G = 1.22$ and $I_G = 0.40$.

Our next step in generating synthetic $Y'(x)$ samples is to multiply each discrete value of $G(x)$, generated in the first step, by a random lognormal draw of $U(x)$. We do so twice, once by setting $\alpha = 1.8$ (to obtain $\sigma_Y = 1.27$ and $I_Y / I_G = 0.92$) and again by setting $\alpha = 1.5$ (resulting in $\sigma_Y = 1.57$ and $I_Y / I_G = 0.61$). Results

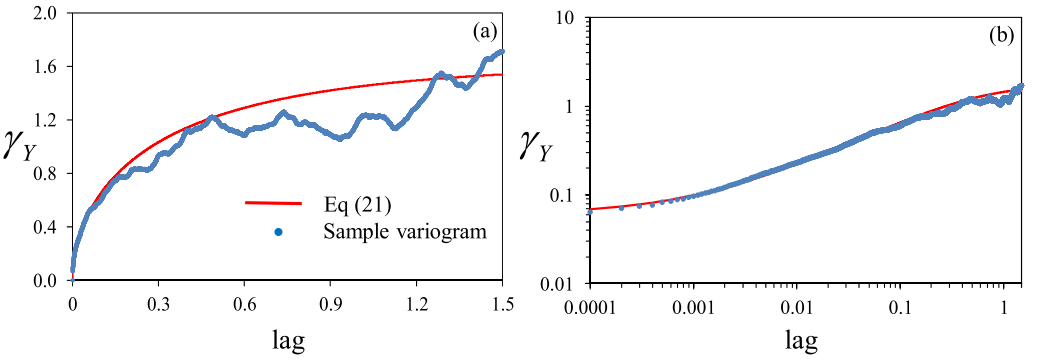

Figure 3. Analytical expression (21) and sample variogram, γ_Y, of a single Y realization generated with $\alpha = 1.8$ on (a) arithmetic and (b) log-log scales.

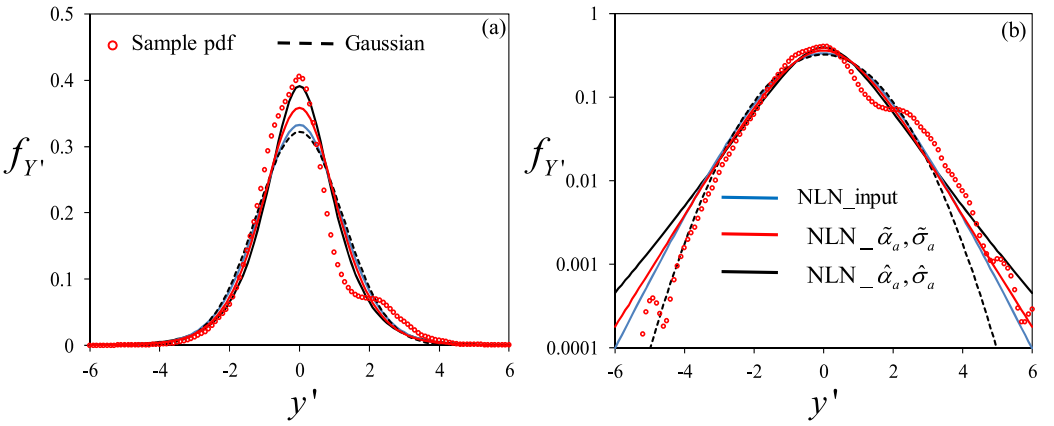

Figure 4. Sample pdf of Y' on (a) arithmetic and (b) semi-logarithmic scales obtained on a single realization generated with $\alpha = 1.8$. Also shown are: the theoretical (input) NLN pdf (NLN_input), a Gaussian pdf with variance equal to that of the generated Y' sample, NLN with parameters $\tilde{\alpha}_a$ and $\tilde{\sigma}_a$ (NLN_$\tilde{\alpha}_a,\tilde{\sigma}_a$), NLN with parameters $\hat{\alpha}_a$ and $\hat{\sigma}_a$ (NLN_$\hat{\alpha}_a,\hat{\sigma}_a$).

corresponding to $\alpha = 1.8$ are presented below and those obtained with $\alpha = 1.5$ are provided as supporting information.

Figure 3 compares on arithmetic and logarithmic scales the sample variogram of a single Y realization generated with $\alpha = 1.8$ at lags $\delta \le s \le l/2$, as is common in practice, with the corresponding analytical expression (21) evaluated with the generating parameters. Agreement between the two variograms is seen to be very good at small lags and acceptable at large lags. A relatively small theoretical nugget of 6.4×10^{-2} is correctly

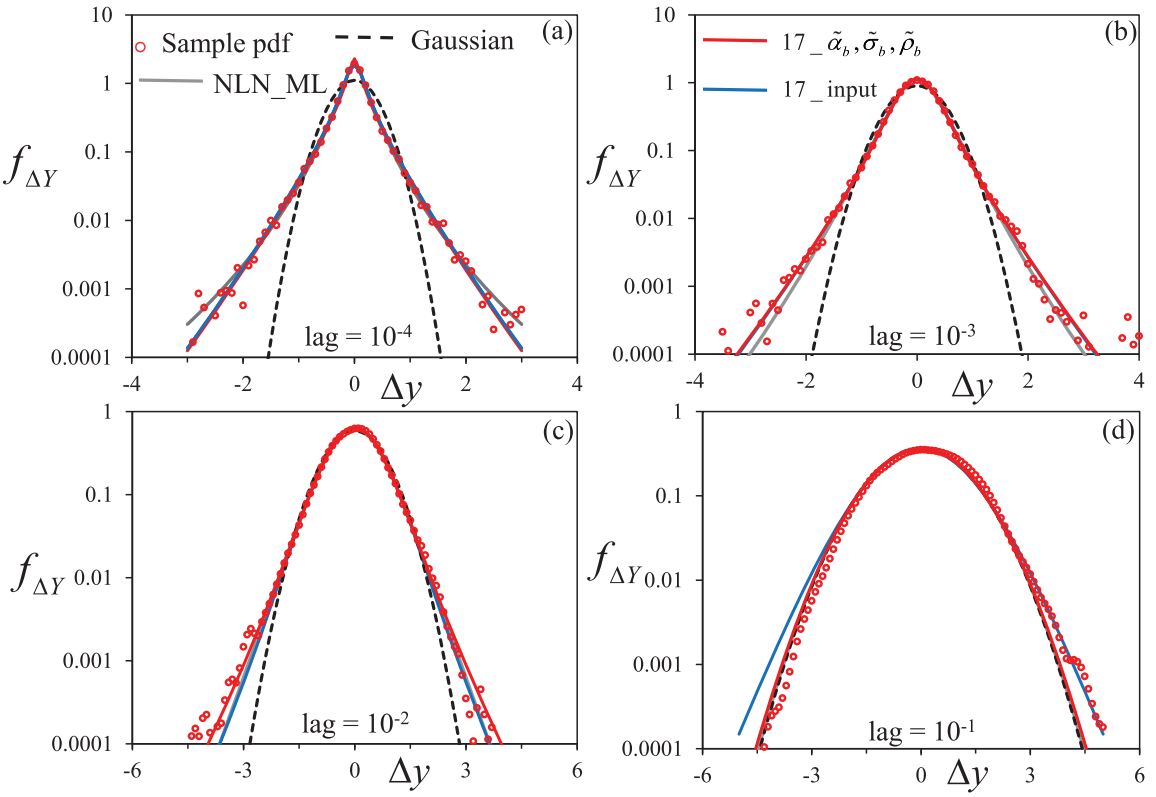

Figure 5. Sample pdf of increments associated with four lags obtained on a single realization generated with $\alpha = 1.8$; Gaussian pdf with variance equal to that of the generated ΔY sample; ML fit of NLN (NLN_ML). Also shown is the pdf (17) evaluated using (*i*) input parameters (17_input), (*ii*) $\tilde{\alpha}_b$, $\tilde{\sigma}_b$, $\tilde{\rho}_G$ (17_$\tilde{\alpha}_b$, $\tilde{\sigma}_b$, $\tilde{\rho}_G$). Some curves are not clearly distinguishable since they overlap, e.g., 17_$\tilde{\alpha}_b,\tilde{\sigma}_b,\tilde{\rho}_G$ and 17_input in (a) and (b), NLN_ML and 17_input in (c), 17_$\tilde{\alpha}_b,\tilde{\sigma}_b,\tilde{\rho}_G$ and NLN_ML in (d).

Table 1. Estimates of α and σ_G Computed With *Methods a* and *b* for Two Synthetic Test Cases[a]

	Synthetic Test Case $\alpha = 1.8$; $\sigma_G = 1.22$	Synthetic Test Case $\alpha = 1.5$; $\sigma_G = 1.22$
$\tilde{\alpha}_a$	1.73 (3.8%)	1.53 (2.2%)
$\hat{\alpha}_a$	1.63±0.01 (9.6%)	1.60±0.02 (6.8%)
Mean of $\tilde{\alpha}_b$	1.78 (1.4%)	1.55 (3.2%)
CV of $\tilde{\alpha}_b$	0.05	0.02
$\tilde{\sigma}_a$	1.15 (5.8%)	1.21 (0.9%)
$\hat{\sigma}_a$	1.09±0.01 (10.8%)	1.27±0.02 (4.0%)
Mean of $\tilde{\sigma}_b$	1.19 (3.0%)	1.23 (0.3%)
CV of $\tilde{\sigma}_b$	0.04	0.03

[a]The relative percentage differences between estimates and reference values are listed in parentheses.

reproduced and clearly visible on logarithmic scale. One can verify in supporting information Figure S1 that, in line with (21), the nugget obtained with $\alpha = 1.5$ is larger (i.e., it is equal to 0.55). For illustration purposes, we present here results associated with a single realization for each value of α analyzed. Results of similar quality are obtained for diverse realizations (not shown).

Figure 4 depicts on arithmetic and semilogarithmic scales the sample frequency distribution of Y' obtained with $\alpha = 1.8$ and $\sigma_G = 1.22$. Also shown for comparison are the theoretical (input) NLN pdf (10) and a Gaussian pdf with variance equal to that of the generated Y' sample (which we calculate as $M_2^Y = 1.53$). Whereas the input NLN is unimodal and symmetric, the frequency distribution of Y' is slightly bimodal and asymmetric with relatively sharp peak. Even though the sample mean ($= -0.09$) and standard deviation ($= 1.24$) of the sample are quite close to their input values of 0 and 1.27, the sample and input distributions differ markedly. We attribute this difference to insufficient size of the generated sample, noting that an increase in sample size by one order of magnitude from $N_e = 30{,}000$ to $N_e = 300{,}000$ in supporting information Figure S2 results in much improved agreement between the distributions. This notwithstanding, the peak and fine tail details of the input NLN are not captured fully by the very large sample in supporting information Figure S2. As the sizes of real data samples are often relatively small, we expect corresponding sample frequencies to provide less than perfect representations of their parent sub-Gaussian pdfs.

In contrast to somewhat irregular frequency distributions of Y' samples, those of increments ΔY are found to be symmetric with sharp peaks and heavy tails at small lags, milder peaks and lighter tails at larger lags, as seen in Figure 5 (and supporting information Figure S4). These increment frequency distributions are represented closely by our novel pdf expression (17) when one evaluates it using input parameters of the Y'-generating NLN distribution. These inputs consist of constant parameters α, σ_G and a lag-dependent parameter ρ_G. We describe two methods of estimating these parameters, *method a* using only Y data to estimate α and σ_G and *method b* using Y and ΔY data jointly to estimate α, σ_G and ρ_G.

4.1. Parameter Estimation Method a

Method a relies on the marginal frequency distribution and moments of Y' which depend only on two parameters, α and σ_G. One therefore cannot estimate ρ_G by this method.

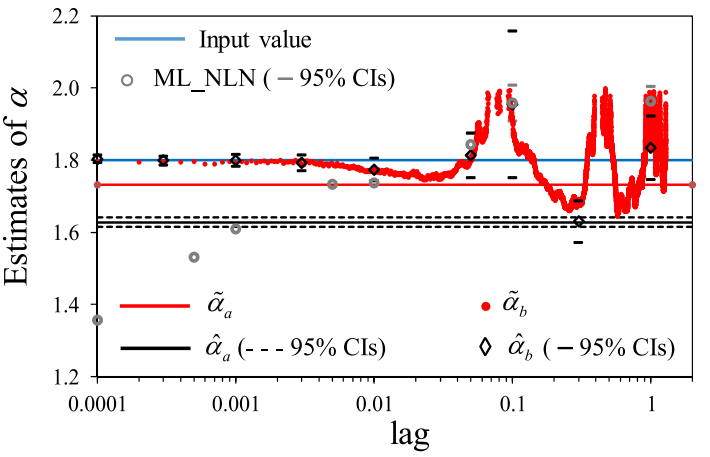

Figure 6. Estimates of α versus lag obtained on a single realization generated with $\alpha = 1.8$. Input value and ML estimates computed by fitting NLN to ΔY data are also reported (ML_NLN). The 95% confidence intervals (CIs) of the ML estimates are plotted only when not negligible.

Fitting the NLN pdf (10) of Y' by maximum likelihood (ML) to a frequency distribution of mean-removed Y data yields estimates $\hat{\alpha}_a$ and $\hat{\sigma}_a$ of α and σ_G. Doing so for synthetic Y' data generated with $\alpha = 1.8$ and $\sigma_G = 1.22$ yields $\hat{\alpha}_a = 1.63 \pm 0.01$ and $\hat{\sigma}_a = 1.09 \pm 0.01$, the half-width of the confidence intervals being calculated as twice the square root of the diagonal components of the Cramer-Rao lower-bound approximation of the parameter estimation covariance matrix. Alternatively one can replace the second and fourth moments, $\langle Y'^2 \rangle$ and $\langle Y'^4 \rangle$, of Y' in (12) and (13) by their sample

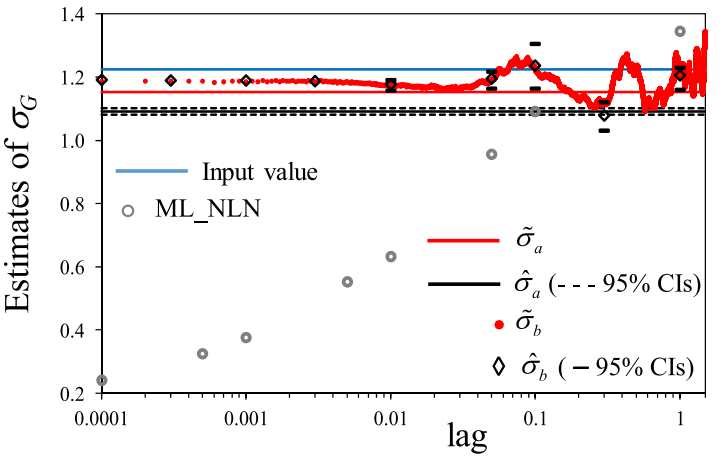

Figure 7. Estimates of σ_G versus lag obtained on a single realization generated with $\alpha = 1.8$. Input value and ML estimates computed by fitting NLN to ΔY data are also reported (ML_NLN). The 95% confidence intervals (CIs) of the ML estimates are plotted only when not negligible.

counterparts, M_2^Y and M_4^Y, to obtain explicit estimates $\tilde{\alpha}_a = 1.73$ and $\tilde{\sigma}_a = 1.15$ of the same parameters. These as well as estimates obtained in the case of $\alpha = 1.5$, and associated estimation errors, are listed in Table 1. Inserting each pair of estimates into (10) yields pdf curves illustrated in Figure 4 (and supporting information Figure S3).

4.2. Parameter Estimation Method b

Method b provides estimates of all three parameters α, σ_G and ρ_G characterizing Y' and ΔY by relying jointly on samples of both functions. Replacing $\langle Y'^2 \rangle$, $\langle \Delta Y^2 \rangle$ and $\langle \Delta Y^4 \rangle$ in (12), (18) and (19) by their sample counterparts M_2^Y, $M_2^{\Delta Y}$ and $M_4^{\Delta Y}$ provides explicit estimates $\tilde{\alpha}_b$, $\tilde{\sigma}_b$ and $\tilde{\rho}_G$ of the three parameters. Whereas α and σ_G are constant, their estimates $\tilde{\alpha}_b$ and $\tilde{\sigma}_b$, obtained jointly with the lag-dependent parameter ρ_G, do vary slightly with lag at small values of s and oscillate at larger lags (within the range $\delta \le s \le l/2$), as seen in Figures 6 and 7 (also supporting information Figures S5 and S6). Included in these figures are horizontal lines indicating input values of these parameters and estimates obtained by *method a*. Mean values of $\tilde{\alpha}_b$ and $\tilde{\sigma}_b$ are listed in Table 1 together with their coefficients of variation (CV). All estimates are close to the true input values and are characterized by small CVs. Estimates of the correlation coefficient, $\tilde{\rho}_G$, decrease with lag as do their input values (Figure 8 and supporting information Figure S7). The same figures also depict sample ρ_G of simulated G. Agreement between all three curves is remarkable.

We recommend using *methods a* and *b* in tandem to verify that they yield comparable estimates of α and σ_G.

Yet another possibility is to estimate α and ρ_G upon fitting (17) to frequency distributions of ΔY at various lags by ML while setting the variance of Y' equal to its sample value, M_2^Y. Due to CPU time constrains, we tried this by resampling the available data at a subset of 1000, 2000, and 3000 points using a polyphase filter, as embedded in the "resample" subroutine implemented in Matlab. Whereas parameter estimates obtained with each subset were virtually identical, their 95% confidence intervals decreased slightly with number of points. CPU time ranged from 10 to more than 100 h (using 3000 points) on a 2.80 GHz Intel i7–860 processor. Figures 6–8 (and supporting information Figures S5–S7) show ML estimates $\hat{\alpha}_b$, $\hat{\sigma}_b$, $\hat{\rho}_G$ obtained in this manner, using 3000 points, together with their 95% confidence intervals. For clarity, we have not plotted the 95% confidence intervals when negligible. All these estimates are close to

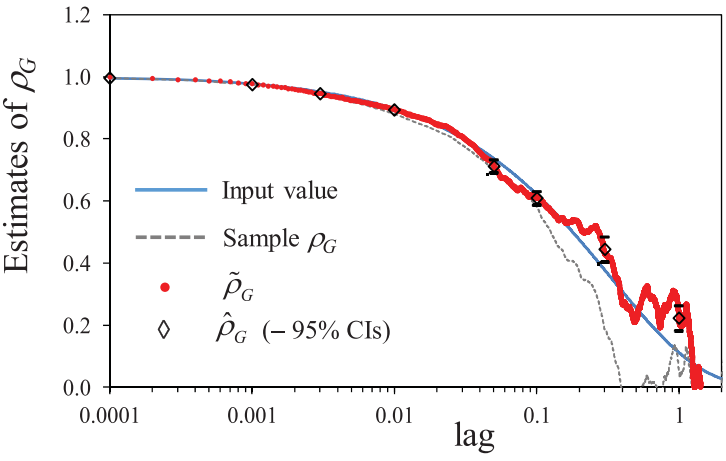

Figure 8. Estimates of ρ_G obtained on a single realization generated with $\alpha = 1.8$. Input values and sample ρ_G of simulated G are also reported. The 95% confidence intervals (CIs) of the ML estimates are plotted only when not negligible.

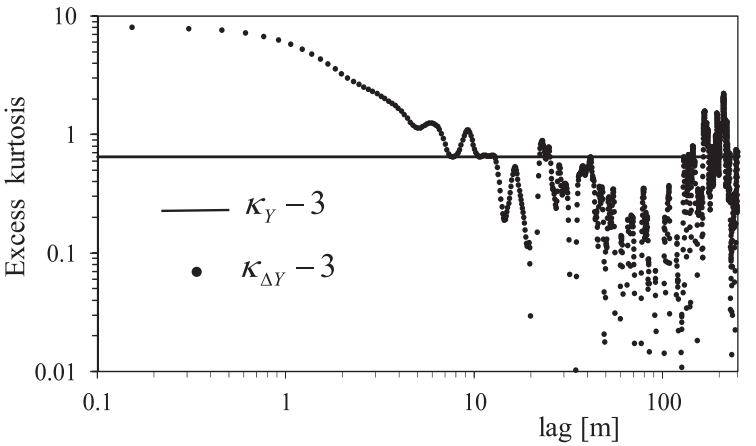

Figure 9. Excess kurtosis of mean-removed porosity data (continuous line) and of porosity increments (symbols) versus lag.

those obtained much more efficiently (requiring negligible CPU time) on the basis of (12) and (18) and (19). Optimizing the ML computation of $\hat{\alpha}_b$, $\hat{\sigma}_b$ and $\hat{\rho}_G$ was outside the scope of this paper.

We end our analysis of synthetic data by noting that it is possible, though theoretically unjustified, to represent increment frequency distributions at various lags quite closely by NLN pdfs, as illustrated in Figure 5 (and supporting information Figure S4). This is what we did in some of our own previous studies, most recently in *Guadagnini et al.* [2015], of which more is said in section 5. The practice has led us, and others (see section 1), to the incorrect conclusion that parameters controlling the peaks and tails of increment distributions, such as our α and σ_G, vary with lag as do their ML NLN estimates in Figures 6 and 7 (and/or supporting information Figures S5 and S6). It is now clear that this phenomenon is an artifact of using an inappropriate (in this case NLN, in some previous cases Lévy stable or other) model to interpret incremental data.

5. Field Application

To conclude we let Y represent neutron porosity data from a deep vertical borehole in southwestern Iran recently analyzed by *Dashtian et al.* [2011] and *Guadagnini et al.* [2015]. The well is drilled in the Maroon field within which gas drive is used to produce oil and natural gas. A large number (3,567) of neutron porosity data taken at a distance of about 15 cm apart are available, having sample mean $M_1^Y = 14\%$ and sample standard deviation = 6.4%. Figure 9 plots excess kurtosis of porosity increments, ΔY, versus lag which ranges from 15 cm to $l/2$ where $l = 543$ m is the total depth of the well segment along which data are

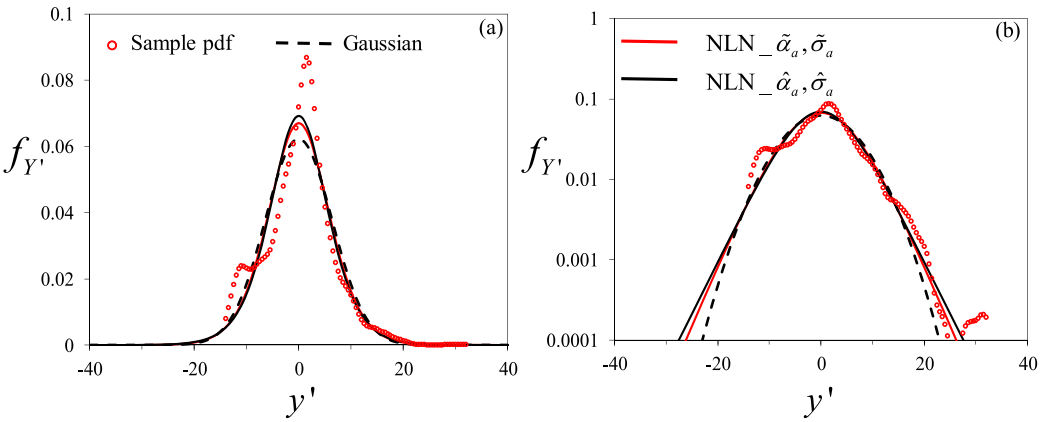

Figure 10. Sample pdf of neutron porosity data, Y', on (a) arithmetic and (b) semilogarithmic scales. Also shown are: a Gaussian pdf with variance equal to that of the Y' sample, NLN with parameters $\tilde{\alpha}_a$ and $\tilde{\sigma}_a$ (NLN_$\tilde{\alpha}_a, \tilde{\sigma}_a$), NLN with parameters $\hat{\alpha}_a$ and $\hat{\sigma}_a$ (NLN_$\hat{\alpha}_a, \hat{\sigma}_a$).

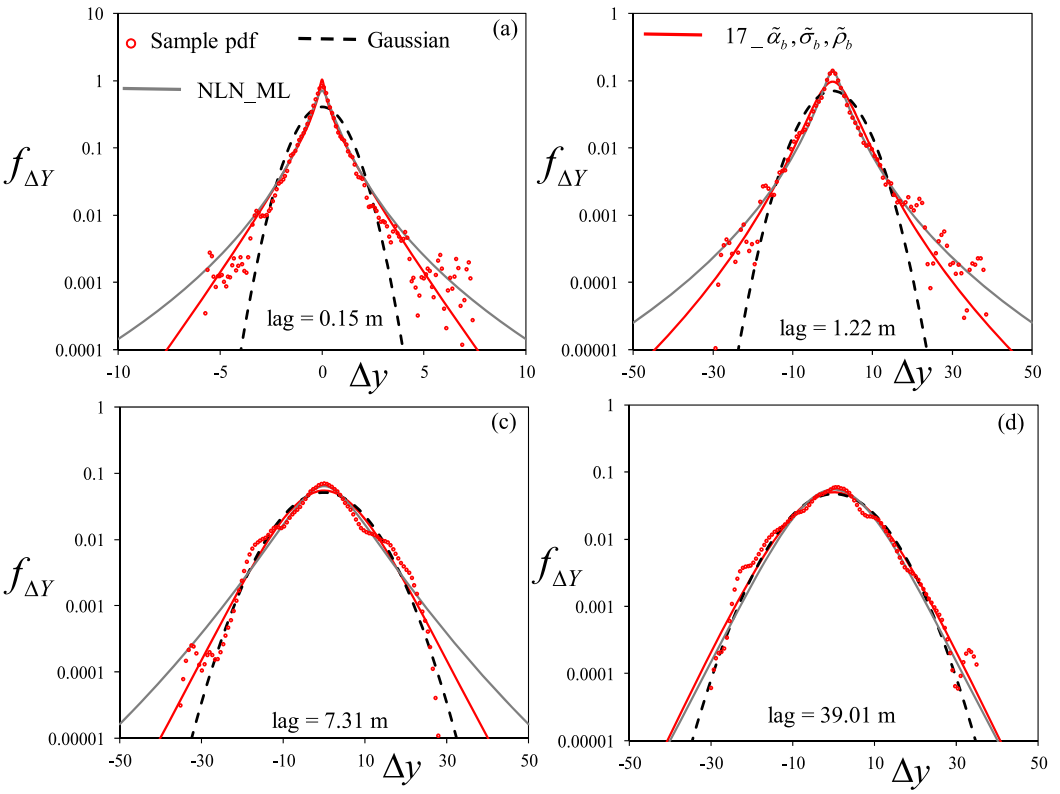

Figure 11. Sample pdf of increments of neutron porosity data, ΔY, at four lags. Also shown are Gaussian pdf with variance equal to that of the sample, ML fit of NLN (NLN_ML) and expression (17) evaluated using $\tilde{\alpha}_b$, $\tilde{\sigma}_b$, $\tilde{\rho}_G$ (17_$\tilde{\alpha}_b$, $\tilde{\sigma}_b$, $\tilde{\rho}_G$).

available. Excess kurtosis $\kappa_{\Delta Y}-3$ is significantly larger than zero at small lags, then decreases with increasing lags to oscillate around relatively small values ($<<1$) at the largest lags. Included in Figure 9 is a horizontal line denoting excess kurtosis, $\kappa_Y-3=0.64$, of mean-removed porosities $Y'=Y-M_1^Y$. As predicted by our generalized sub-Gaussian model and observed in our synthetic examples, whereas at small lags $\kappa_{\Delta Y}-3>\kappa_Y-3$ implying that frequency distributions of ΔY exhibit sharper peaks and heavier tails than does that of Y', the opposite happens at large lags. Indeed, these frequency distributions and models fitted to them in Figures 10 and 11, respectively, are closely reminiscent of their synthetic counterparts in Figures 4 and 5 (supporting information Figures S3 and S4): the pdf of Y' appears slightly asymmetric while distributions of ΔY seem symmetric with peaks and tails that decay with lag. Applying parameter estimation *method a* to these porosity data we find $\tilde{\alpha}_a=1.78$, $\tilde{\sigma}_a=6.10\%$, $\hat{\alpha}_a=1.73\pm 0.05$ and $\hat{\sigma}_a=5.98\%\pm 0.19\%$. The corresponding pdfs are depicted in Figure 10. Parameter estimates $\tilde{\alpha}_b$ and $\tilde{\sigma}_b$ obtained by *method b*, plotted respectively versus lag in Figures 12 and 13, oscillate at 15 cm $\le s \le I/2$ in an irregular fashion about mean values of $\tilde{\alpha}_b=1.75$ and $\tilde{\sigma}_b=6.15\%$. The

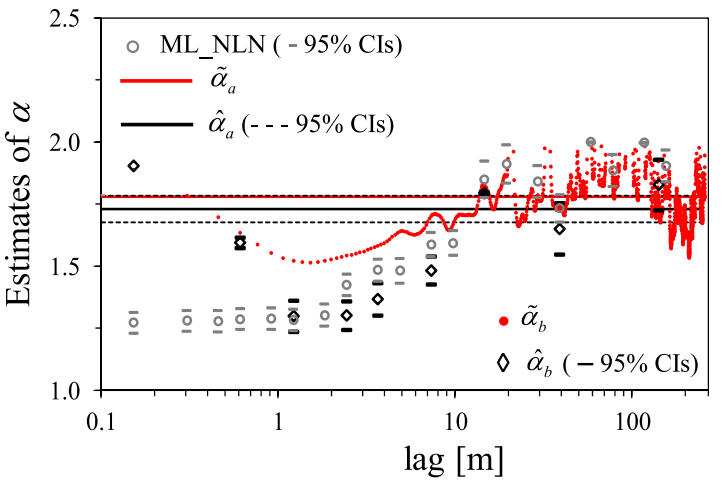

Figure 12. Estimates of α versus lag for the neutron porosity data. ML estimates computed by fitting NLN to ΔY data are also reported (ML_NLN). The 95% confidence intervals (CIs) of the ML estimates are plotted only when not negligible.

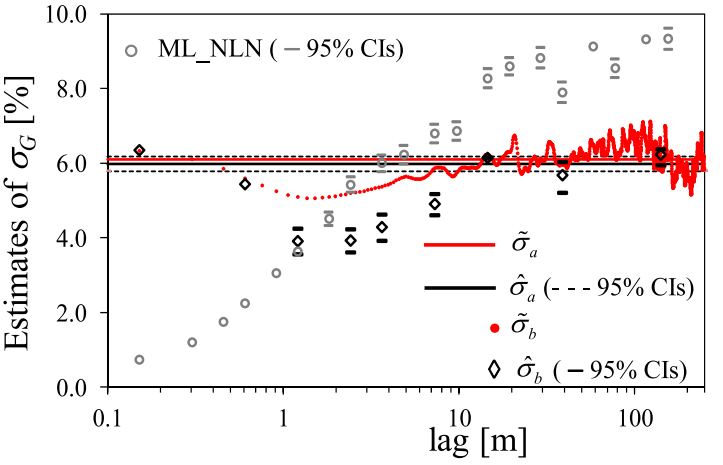

Figure 13. Estimates of σ_G versus lag for the neutron porosity data. ML estimates computed by fitting NLN to ΔY data are also reported (ML_NLN). The 95% confidence intervals (CIs) of the ML estimates are plotted only when not negligible.

latter, characterized by small coefficients of variation (0.05 and 0.07, respectively), are very close to values estimated by *method a*. Included in Figures 12 and 13 are estimates of α and σ_G, respectively, obtained on the basis of NLN fits by ML to frequency distributions of porosity increments at various lags. Once again we see these estimates to increase systematically with lag in a way that is not supported by our theory. Estimates of ρ_G are shown in Figure 14. Maximum likelihood estimates $\hat{\alpha}_a$ and $\hat{\alpha}_b$, $\hat{\sigma}_a$ and $\hat{\sigma}_b$ and $\hat{\rho}_G$ are also shown in Figures 12–14, respectively. As in the case of synthetic data, *methods a* and *b* yield comparable estimates of model parameters α and σ_G, thus strengthening our confidence in the quality of the results.

6. Summary and Conclusions

We proposed, explored and applied to synthetic and real data a new model that reconciles within a unique theoretical framework the probability distributions of spatial (or temporal) variables (including physical, hydrogeological, geophysical, environmental, biological, as well as financial data) and the way distributions of their increments change with scale. When viewed in the context of theoretical developments and synthetic as well as field data analyses in our related earlier works [e.g., *Guadagnini et al.*, 2015 and references therein], our model is seen to be unique in providing a comprehensive, self-consistent and rigorous explanation of the above statistical scaling and related phenomena that have puzzled analysts for decades. These related phenomena include power-law scaling of sample structure functions (statistical moments of absolute increments) in midranges of lags, extended power-law scaling (linear relations between log structure functions of successive orders) at all lags, and nonlinear (and eventually anisotropic) scaling of power-law exponent with order of sample structure function.

The model has generalized sub-Gaussian form, subordinating variables to truncated fractional Brownian motion (tfBm) through the action of a lognormal subordinator (leaving open the possibility of other choices). Statistics of increments are functions of, and thus scale with, lower and upper cutoffs proportional to data resolution and sampling domain scales, respectively. These statistics can thus be scaled up or down to reflect changing resolution and/or sampling domain scales. For given cutoffs the statistics are fully defined in terms of two constant parameters and a lag-dependent correlation function that depends, in a known way, on a coefficient and a Hurst scaling exponent. Decay

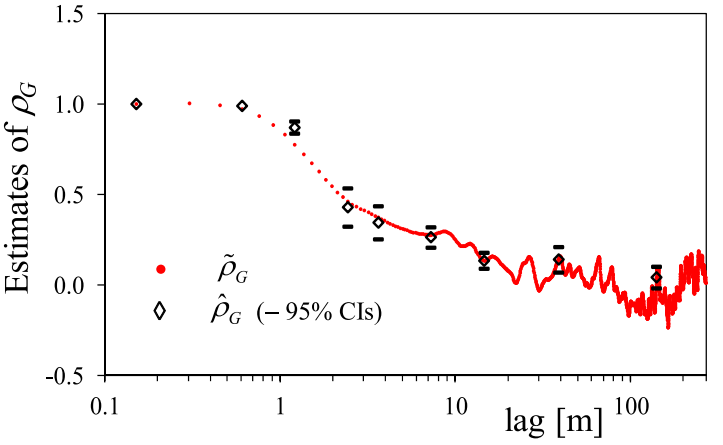

Figure 14. Estimates of ρ_G versus lag for the neutron porosity data. The 95% confidence intervals (CIs) of the ML estimates are plotted only when not negligible.

of this correlation function with lag was shown to be the reason why sharp peaks and heavy tails of increment probability density functions flatten and lighten with lag. We proposed and tested relatively simple ways to estimate model parameters by considering separately or jointly samples of the variable and its increments at various lags. Our model opens the way for conditional and unconditional simulation, and interpolation, of non-Gaussian random variables in one or more space-time dimensions.

Acknowledgments
Our work was supported in part through a contract between the University of Arizona and Vanderbilt University under the Consortium for Risk Evaluation with Stakeholder Participation (CRESP) III, funded by the U.S. Department of Energy. Funding from MIUR (Italian Ministry of Education, Universities and Research—PRIN2010-11; project: "Innovative methods for water resources under hydro-climatic uncertainty scenarios") is acknowledged. All data used in the paper will be retained by the authors for at least 5 years after publication. The synthetic data will be available to the readers upon request. We thank Professor Muhammad Sahimi, University of Southern California, for having generously shared with us borehole geophysical log data some of which we analyze in this paper.

References

Boffetta, G., A. Mazzino, and A. Vulpiani (2008), Twenty-five years of multifractals in fully developed turbulence: A tribute to Giovanni Paladin, *J. Phys. A Math. Theor.*, *41*, 363001.

Castaing, S., Y. Gagne, and E. J. Hopfinger (1990), Velocity probability density functions of high Reynolds number turbulence, *Physica D*, *46*, 177–200.

Clark, P. K. (1973), A subordinated stochastic process model with finite variance for speculative prices, *Econometrica*, *41*, 1.

Dashtian, H., G. R. Jafari, M. Sahimi, and M. Masihi (2011), Scaling, multifractality, and long-range correlations in well log data of large-scale porous media, Physica A, *390*, 2096–2111, doi:10.1016/j.physa.2011.01.010.

Deutsch, C., and A. Journel (1998), *GSLIB: Geostatistical Software Library and User's Guide*, 2nd ed., Oxford Univ. Press, N. Y.

Di Federico, V., and S. P. Neuman (1997), Scaling of random fields by means of truncated power variograms and associated spectra, *Water Resour. Res.*, *33*, 1075–1085, doi:10.1029/97WR00299.

Ganti, V., A. Singh, P. Passalacqua, and E. Foufoula-Georgiu (2009), Subordinated Brownian motion model for sediment transport, *Phys. Rev. E*, *80*, 011111, doi:1539-5663755/2009/80(1)/011111(9).

Georgiou, P. G., and C. Kyriakakis (2006), Maximum likelihood parameter estimation under impulsive conditions, a sub-Gaussian signal approach, *Signal Process.*, *86*, 3061–3075, doi:10.1016/j.sigpro.2006.01.007.

Guadagnini, A., S. P. Neuman, and M. Riva (2012), Numerical investigation of apparent multifractality of samples from processes subordinated to truncated fBm, *Hydrol. Processes*, *26*, 2894–2908, doi:10.1002/hyp.8358.

Guadagnini, A., S. P. Neuman, M. G. Schaap, and M. Riva (2013), Anisotropic statistical scaling of vadose zone hydraulic property estimates near Maricopa, Arizona, *Water Resour. Res.*, *49*, 8463–8479, doi:10.1002/2013WR014286.

Guadagnini, A., S. P. Neuman, M. G. Schaap, and M. Riva (2014), Anisotropic statistical scaling of soil and sediment texture in a stratified deep vadose zone near Maricopa, Arizona, *Geoderma*, *214*, 217–227, doi:10.1016/j.geoderma.2013.09.008.

Guadagnini, A., S. P. Neuman, T. Nan, M. Riva, and C. L. Winter (2015), Scalable statistics of correlated random variables and extremes applied to deep borehole porosities, *Hydrol. Earth Syst. Sci.*, *19*, 1–17, doi:10.5194/hess-19-1-2015.

Katz, R. A., M. B. Parlange, and P. Naveau (2002), Statistics of extremes in hydrology, *Adv. Water Resour.*, *25*, 1287–1304, doi:10.1016/S0309-1708(02)00056-8.

Kozubowski,T. J., M. M. Meerschaert, and K. Podgorski (2006), Fractional Laplace motion, *Adv. Appl. Probab.*, *38*, 451–464.

Kozubowski, T. J., K. Podgórski, and I. Rychlik (2013), Multivariate generalized Laplace distribution and related random fields, *J. Multivariate Anal.*, *113*, 59–72.

Kumar, P., and E. Foufoula-Georgiou (1993), Multicomponent decomposition of spatial rainfall fields: 2. Self-similarity in fluctuations, *Water Resour. Res.*, *29*, 2533–2544.

Liu, H. H., and F. J. Molz (1997), Comment on "Evidence for non-Gaussian scaling behavior in heterogeneous sedimentary formations" by S. *Painter*, *Water Resour. Res.*, *33*, 907–908, doi:10.1029/96WR03788.

Meerschaert, M. M., T. J. Kozubowski, F. J. Molz, and S. Lu (2004), Fractional Laplace model for hydraulic conductivity, *Geophys. Res. Lett.*, *31*, L08501, doi:10.1029/2003GL019320.

Neuman, S. P. (2011), Apparent multifractality and scale-dependent distribution of data sampled from self-affine process, *Hydrol. Processes*, *25*(11), 1837–1840, doi:10.1002/hyp.7967.

Neuman, S. P., A. Guadagnini, M. Riva, and M. Siena (2013), Recent advances in statistical and scaling analysis of earth and environmental variables, in *Advances in Hydrogeology*, edited by P. K. Mishra and K. L. Kuhlman, pp. 1–25, Springer, N. Y.

Painter, S. (1996), Evidence for non-Gaussian scaling behavior of heterogeneous sedimentary formations, *Water Resour. Res.*, *32*, 1183–1195.

Painter, S. (2001), Flexible scaling model for use in random field simulation of hydraulic conductivity, *Water Resour. Res.*, *37*, 1155–1163.

Riva, M., S. P. Neuman, and A. Guadagnini (2013a), Sub-Gaussian model of processes with heavy tailed distributions applied to permeabilities of fractured tuff, *Stochastic Environ. Res. Risk Assess.*, *27*, 195–207, doi:10.1007/s00477-012-0576-y.

Riva, M., S. P. Neuman, A. Guadagnini, and M. Siena (2013b), Anisotropic scaling of Berea sandstone log air permeability statistics, *Vadose Zone J.*, *12*(3), 1–15, doi:10.2136/vzj2012.0153.

Riva, M., S. P. Neuman, and A. Guadagnini (2013c), On the identification of Dragon Kings among extreme-valued outliers, *Nonlinear Processes Geophys.*, *20*, 549–561, doi:10.5194/npg-20-549-2013.

Samorodnitsky, G., and M. S. Taqqu (1994), *Stable Non-Gaussian Random Processes*, Chapman and Hall, N. Y.

von Papen, M., J. Saur, and O. Alexandrova (2014), Turbulent magnetic field fluctuations in Saturn's magnetosphere, *J. Geophys. Res. Space Physics*, *119*, 2797–2818, doi:10.1002/2013JA019542.

Yang, C.-Y., K.-C. Hsu, and K.-C. Chen (2009), The use of the Levy-stable distribution for geophysical data analysis, *Hydrogeol. J.*, *17*, 1265–1273, doi:10.1007/s10040-008-0411-1.