Supplementary Material

S1) Nonparametric entropy computation

We considered and compared two computation methods of nonparametric entropy. The first relies on replacing the pdf of equation (5) by a histogram (Beirlant 1997; Wallis 2006, Stone 2015), which we compute using GMT (Wessel and Smith 1998). The second is based on the method of M-Spacings, i.e., estimating the variability (i.e., probabilities) within the series from differences between its elements rather than bins of their values, which was pioneered by Vasicek (1976). This method was later modified and improved by multiple researchers, who improved its convergence properties for samples of small size (Beirlant 1997), e.g., smaller than N=100. We ignore these modifications because our series are long containing hundreds to thousands of elements. Both the histograms and M-spacings methods have one tricky detail, which we needed to understand before we could use it in our analysis and compare it to others. For binning, this detail is the effect of the bin size of the histogram and its relation to the entropy, as explained in section S1.1. Another aspect of histograms we needed to understand is the fact that calculating entropy based on histograms (or, equivalently, probability based on relative frequency) underestimates its value. In the M-Spacings method, we needed to understand the role of the depth of the differencing (i.e., the parameter *M* in equation (9) and (S6) in section S1.2 below), which does slightly affect the computed entropy value.

S1.1) Entropy from histograms

One way for computing entropy of a continuous random variable is by discrete approximations. The approximate entropy, denoted by $H(\Delta)$, is obtained by replacing the integral in (5) by a sum and replacing the pdf by a histogram, where the sum goes over all bins. In that case, however, equation (5) does **not** reduce to:

$$H(\Delta) = \sum_i P_i \cdot \log\left(\frac{1}{P_i}\right) \quad \text{(S1)}$$

This is because while equation (5) depends on the value of the variable through the increment dx, equation (S1) does not. In fact, if the entropy is computed by equation (S1), its numerical value would greatly depend on the bin size, Δx, and not only on the sample's probabilities as it should. It can be shown (Cover and Thomas 2006, Theorem 8.3.1; Stone 2015; Beirlant 1997; Wallis 2006) that the correct discretization of the integral in (5) leads to the following expression for the nonparametric discretized entropy:

$$H(\Delta) = \sum_i [P_i \cdot \log\left(\frac{1}{P_i}\right))] - \log(\frac{1}{\Delta x}) \quad \text{(S2)}$$

and that the discretized entropy value resulting from equation (S2) is practically independent of the bin size.

Finally, it can be shown (Fig S9) that when probability is estimated from relative frequency, entropy is underestimated (Stone 2015). Fig S9 presents entropy values computed from a sample of synthetic colored noise described in section 4.1 based on a histogram and on the M-Spacings method using (9) with different values of M. Clearly, entropy computed by a histogram is underestimated especially for short time series.

S1.2) Entropy based on M-spacings

Another method for computing nonparametric entropy is the so-called "M-Spacings" (Beirlant et al. 1997), which is based on the variability within a sample of the RV. One of the first variations of M-Spacings was developed by Vasicek (1976), who used the cumulative distribution function (cdf), $F(X)$, and the fact that:

$$\frac{1}{f(X)} \equiv \frac{1}{f(F^{-1}(p))} = \frac{d}{dp}F^{-1}(p) \quad \text{(S3)}$$

where p is the probability, to reformulate the integral of differential entropy, (5), as follows:

$$H = \int_{-\infty}^{\infty} f(x) \cdot \log(\frac{1}{f(x)}) \cdot dx \equiv \int_{p=0}^{1} \log\left\{\frac{d}{dp}F^{-1}(p)\right\} dp. \quad \text{(S4)}$$

replacing time domain averaging (i.e., weighting the Shannon information by $f(x)$ and integrating from minus to plus infinity) by probability-space averaging over all possible probabilities from zero to one. The derivative on the right-hand side of (S4) is then replaced by numerical differencing while the cdf is replaced by the empirical cumulative distribution function (ecdf), $F_N(X)$, defined as:

$$F_N(t) = (the\ number\ of\ elements\ in\ the\ sample\ which\ are\ \leq\ t)/N \quad \text{(S5)}$$

where N is the sample size. For example, if $x_1, x_2, \cdots, x_N$ are the sample's elements (or residual time series in our case) and $x_{(1)}, x_{(2)}, \cdots, x_{(N)}$ are its order statistics, i.e., the elements of the sample after they are sorted in an ascending order, $x_{(1)} < x_{(2)} \cdots < x_{(N)}$, then according to definition (S5), $F_N(X_{(i-M)}) = \frac{i-M}{N}$ and $F_N(X_{(i+M)}) = \frac{i+M}{N}$. Thus, equation (S4) can be discretized as:

$$H_{MN} = \frac{1}{N}\sum_{i=1}^{N} \log\{\frac{X_{(i+M)} - X_{(i-M)}}{\frac{i+M}{N} - \frac{i-M}{N}}\} = \frac{1}{N}\sum_{i=1}^{N} \log\left\{\frac{N}{2M}\left(x_{(i+M)} - x_{(i-M)}\right)\right\} \quad \text{(S6)}$$

where M is a positive integer smaller than $N/2$. In equation (S6), we used $x_{(j)} = x_{(1)}$ when $j < 1$ and $x_{(j)} = x_{(N)}$ when $j > N$. Vasicek (1976) proved the convergence of H_{MN} to the true entropy, H, for $N \to \infty$, $M \to \infty$ and $M/N \to 0$. He used equation (S6) to design a statistical test for normality considering that the normal distribution has the maximal entropy among all distributions of the same variance. He concluded that values of M between 2 and 4 lead to tests of maximal power when using small sample sizes (< 50).

If one thinks of the given sample as the discrete function values $\{(F_N(X_{(i)}), X_{(i)});\ i = 1, 2, 3, \cdots, N\}$, the fraction of the middle term in equation (S6) is the slope of this function based on two points only, $(F_N(X_{(i+M)}, X_{(i+M)})$ and $(F_N(X_{(i-M)}, X_{(i-M)})$, although there are $2M + 1$ elements of this function between $i + M$ and $i - M$. Several attempts to improve the expression of the slope of the original Vasicek estimator were made over the years (Beirlant et al. 1997), which improved of the estimator for small sample size, e.g., $N < 100$. We ignored these improvements since the sizes of our time series are much larger.

Figure S9 presents typical entropy values computed based on one of the 100 simulated noise series used in section 4.1, using both the histogram and Vasicek's M-Spacings methods for several values of M. The lowest curve (in light blue) is computed from a histogram with a bin size of 0.01 mm. Using bin sizes of 0.02 to 0.10 mm changes this curve only very slightly. The other four curves in Fig S9 are computed using M-spacings (equation (9) or S6) with $M = 1, 2, 4$ and 10. The underestimation of the value of entropy based on histograms, especially from short time series, is clear and decreases as the length of the time

series increases. In contrast, the M-spacings' entropy values are fairly stable as a function of time series length, increasing only very slightly with longer time series, but also increase with increasing M value. However, the rate of entropy increase as a function of *M* slows down rapidly as *M* increases. We found that the entropy increase with increasing *M* value does not affect the estimated minimum-entropy velocity. However, it does slightly increase the velocity SD based on equations (14) and (16).

Theoretically, Vasicek's estimator requires a large N and M to converge to the true entropy. On the other hand, we found that the smaller *M* is, the less sensitive is the computed entropy value to deterministic, smooth and slowly varying content of the position time series, such as annual and semiannual periodic signals and transients. The smaller M is the more efficiently these signals are removed by the differencing. We therefore proceeded using Vasicek's M-spacings methods with $M = 1$ throughout this paper.

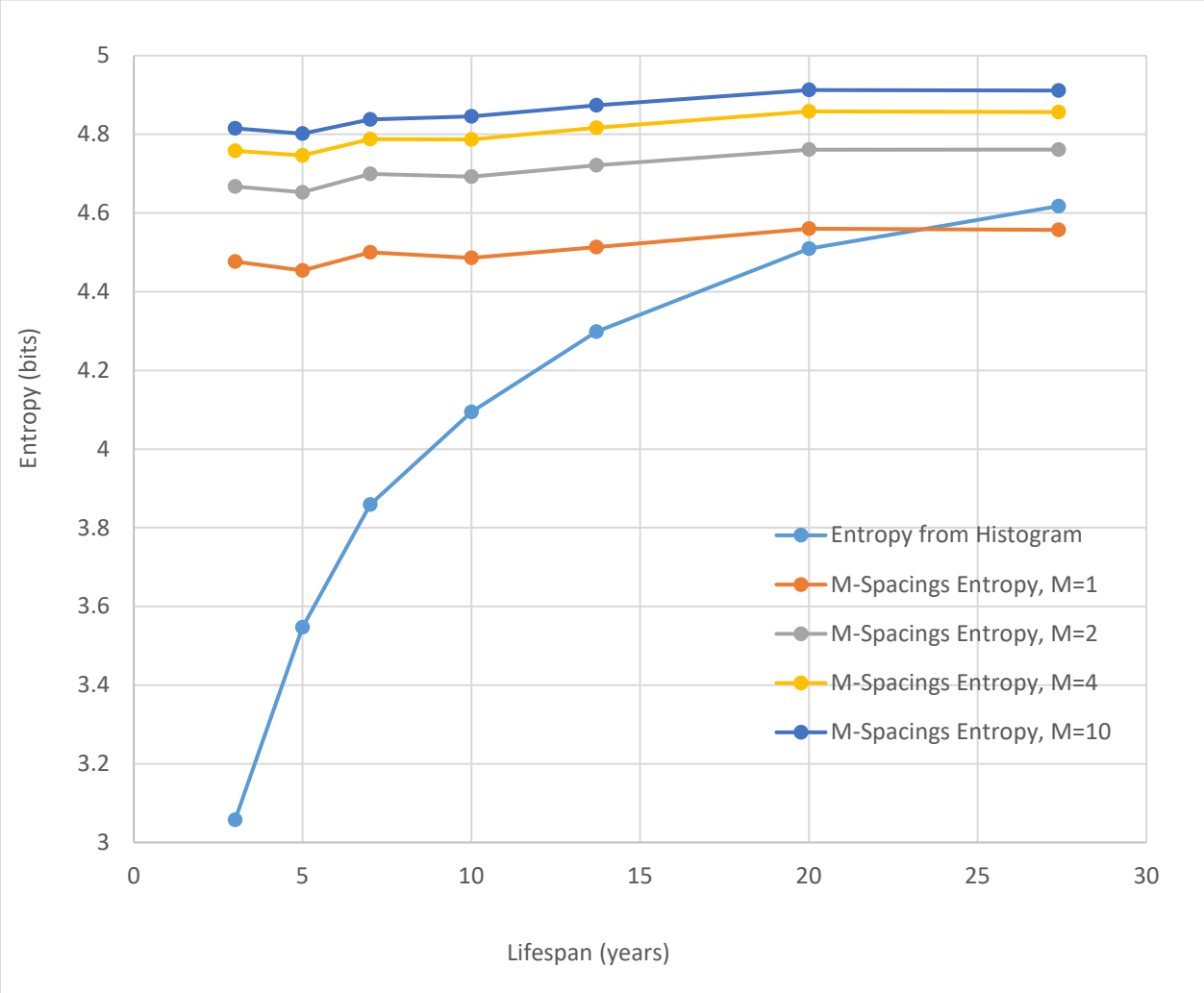

Fig S9: Nonparametric entropy values for a typical simulated noise series (2.7 mm white noise and 5 mm power-law noise of spectral index -0.9). The lowest (light blue) curve is computed from a histogram with a bin size of 0.01 mm. The other four are computed using the M-spacings method with $M = 1, 2, 4$ and 10.

S2) Relevant properties of entropy used in formulating the proposed method

This section lists and discusses some additional details which complement section 2:

1) An important information-theoretic measure, related to entropy, is the "Mutual Information" (MI) between two random variables (Cover and Thomas 2006, chapter 2). MI is a measure of the statistical dependence of one RV on the other. The MI, $I(X;Y)$, of RVs X and Y is related to their entropies by:

$$I(X;Y) = H(X) + H(Y) - H(X,Y) \quad \text{(S7)}$$

where $H(X,Y)$ is the joint entropy of X and Y. If X and Y are independent, their joint entropy is the sum of their marginal entropies (see equation (6)) and the MI vanishes. If the two RVs are completely dependent to become identical, their entropies become identical and their joint entropy is equal to either of their entropies (because if one is given, there is no additional uncertainty introduced by the other), thus the MI is equal to the entropy of either variable. Like entropy, MI is measured in bits and can be computed nonparametrically from reasonably long data samples of both random variables.

2) If $X' = g(X)$ and $Y' = h(Y)$ are invertible functions of X and Y, then (Stone 2015):

$$I(X';Y') = I(X;Y) \quad \text{(S8)}$$

This equation implies that regardless of how complex is the (linear or nonlinear) relation between two RVs, their MI reflects the precise measure of dependence of one on the other. In other words, unlike (linear) correlation coefficients, the MI is sensitive to nonlinear dependencies between variables, which do not necessarily manifest as covariances.

3) If a sinusoid with small (several mm) constant amplitude, constant period (e.g., annual) and constant phase angle is added to the signal, it does not affect its entropy. This is because entropy is a measure of uncertainty (rather than the spread of values), and a constant sinusoid is deterministic and hence adds no uncertainty. However, the sinusoids we deal with in geodetic time series are usually modeled as: $g(\delta t) = A \cdot \sin(\omega \cdot \delta t + \Phi)$ assuming A, ω and Φ are a constant amplitude (mm), angular frequency and phase angle, respectively, but δt is a random timing variable. This latter form of sinusoids is then stochastic and hence has a certain nonzero entropy. However, according to (S8):

$$I(\delta t; g(\delta t)) = I(\delta t; \delta t) = H(\delta t) \quad \text{(S9)}$$

which means that δt and $g(\delta t)$ are stochastically identical, i.e., no additional uncertainty beyond that of δt is introduced by the stochastic sinusoid.

4) Adding a step discontinuity to the signal increases its entropy and (because entropy is insensitive to a bias) the closer the step is to the middle of the series the larger is the added entropy. We assume that the locations (but not the heights) of step discontinuities are known based on a pre-analysis. Thus, the residual time series can be broken into multiple partial series, one between every two consecutive discontinuities. Each partial time series is assigned a probability, P_i:

$$P_i = T_i/T \quad \text{(S10)}$$

where T_i is the length to the partial series and T is the total length in years. For example, two discontinuities at quarter and middle the series divide it to three partial series, two are of length $T/4$ and the third is of length $T/2$. Thus, we assign probability $P_i = (T/4)/T = 0.25$ to the first two and $P_i =$

$(T/2)/T = 0.5$ to the third. The entropy of each partial series, $\mathbb{H}_i, i = 1,2,\cdots,d+1$, where d is the number of discontinuities, is computed by (9) and the total entropy of the entire series is computed by:

$$\mathbb{H} = \sum_{i=1}^{d+1} P_i \cdot \mathbb{H}_i \quad \text{(S11)}$$

5) Entropy can be used for decorrelating colored noise in two steps. First, a nonparametric average entropy, $\mathbb{H}$, is computed from the residuals using (9) (and (12) if necessary) as an approximation of the entropy rate. This entropy rate accounts for all noises and all correlations within the series. Now, an "effective", uncorrelated (iid) Gaussian RV which has that same entropy is sought. This is achieved by substituting its Gaussian pdf, $f(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}}$, in (5) and replacing the base 2 by natural logarithm:

$$H = \int_{-\infty}^{\infty} f(x) \cdot \ln(\frac{1}{f(x)}) \cdot dx = -\int_{-\infty}^{\infty} f(x) \cdot \ln(f(x)) \cdot dx = -\int f(x) \cdot \{ln\frac{1}{\sqrt{2\pi\sigma^2}} - \frac{(x-\mu)^2}{2\sigma^2}\} \cdot dx$$

$$= \frac{1}{2}\ln(2\pi\sigma^2) + \frac{1}{2\sigma^2}\int (x-\mu)^2 \cdot f(x) \cdot dx = \frac{1}{2}\ln(2\pi\sigma^2) + \frac{\sigma^2}{2\sigma^2} = \frac{1}{2}\ln(2\pi e\sigma^2) \quad \text{nats} \quad \text{(S12)}$$

To obtain the entropy in bits, the natural logarithm in (S12) is replaced by a base 2 logarithm. Thus, the average entropy, $\mathbb{H}$, can be converted to the SD, $\sigma_{\delta p}$, of an iid Gaussian position noise by:

$$\mathbb{H} = \frac{1}{2} log_2(2\pi e\sigma_{\delta P}^2) = 2.0471 + log_2(\sigma_{\delta P}) \quad \text{bits} \quad \text{(S13)}$$

$$\Longrightarrow \quad \sigma_{\delta P} = \frac{1}{\sqrt{2\pi e}} 2^{\mathbb{H}} \equiv 2^{(\mathbb{H} - 2.0471)} \quad \text{mm} \quad \text{(S14)}$$

where the notation $\sigma_{\delta p}$ is used to emphasize that this SD is that of a position change relative to P_0 and not of an absolute position. Computing the SD of a representative residual position by (S14) is equivalent to decorrelating the residual position time series.

S3) Velocity estimation by regression with a priori error covariance matrix

We use four types of simulated noise in this study: (1) A stationary mixture of noise with 2.7 mm white (Gaussian) noise (WN) and 5 mm power-law (PL) noise of spectral index $k = -0.9$; (2) A slightly non-stationary noise with 2.7 mm WN and 5 mm PL noise of $k = -1.1$; (3) A slightly non-stationary, similar to type (1) but with the addition of random walk (RW) (i.e., $k = -2.0$) of step size of *s=0.1 mm*; (4) A significantly non-stationary similar to type (3) except that the step size of RW is *s=0.5 mm*.

The a priori covariance matrix for any of these four noise types can be described as (Williams 2003):

$$C_x = (2.7\ mm)^2 \cdot I_N + (5.0\ mm)^2 \cdot R \cdot R^T + s^2 \cdot S \cdot S^T \quad \text{(S15)}$$

where I_N is the identity matrix of size *N* (the total number of observations), and:

$$R = \begin{bmatrix} \psi_0 & 0 & 0 & 0 & \cdots & 0 \\ \psi_1 & \psi_0 & 0 & 0 & \cdots & 0 \\ \psi_2 & \psi_1 & \psi_0 & 0 & \cdots & 0 \\ \psi_3 & \psi_2 & \psi_1 & \psi_0 & \cdots & 0 \\ & & \vdots & & & \\ \psi_N & \psi_{N-1} & \psi_{N-2} & \cdots & & \psi_0 \end{bmatrix} \quad \text{(S16)}$$

where:

$$\psi_n = \frac{\Gamma(n - \kappa/2)}{n!\ \Gamma(-\kappa/2)} \qquad n = 0,1,2,\cdots,N \quad \text{(S17)}$$

and where κ is the spectral index of the power-law noise ($\kappa = -0.9$ for type 1, 3 and 4 and $\kappa = -1.1$ for type 2) and Γ is the gamma function. For large *N* (*>100*), equation (S17) reduces to:

$$\psi_n \cong \frac{n^{(-\kappa/2\,-\,1)}}{\Gamma(-\kappa/2)} \quad . \tag{S18}$$

The random walk step size, s, in equation S10 is zero for type 1 and 2, $0.1\,mm$ for type 3 and $0.5\,mm$ for type 4, and S is the square-root of the covariance matrix of the random walk, computed very similarly to R of equation (S16) except that $\kappa = -2.0$.

Williams (2003) reported that the triangular square-root matrices R and S should be scaled by $\Delta t^{(-\kappa/4)}$ before computing the covariance matrix in (S15), where Δt is the sampling interval (1 day in our simulations). This scaling, however, distorted the estimated value of the variance of unit weight (i.e., the $\hat{\sigma}^2 = e^T P e/(n-2)$ where e is the regression residual vector). While without the scaling, the variance of unit weight in all 100 simulations was very close to a unit as it should be when the noise covariance is accurately known, it was significantly different than a unit when the scaling (using $\Delta t = 1/365$ years) was done. We therefore ignored the scaling (i.e., considered Δt as 1.0 (day) and not as $1/365$ (years)).

S4) Why is the M-Spacings entropy a good approximation for the "entropy rate"?

The results of section 4.1 suggest that the approximation (8) and the choice of (9) lead to the correct results. In other words, replacing the entropy rate by a nonparametric, M-Spacings, average entropy, computed from the variations within the residual time series, is a good approximation. In this section, we explain how we arrived at that approximation and further clarify its numerical advantages.

The classical way to proceed using the formal definitions (6) and (7) is by assuming that the stochastic process which generates the residual time series is a multivariate normal with a joint pdf:

$$f(X) = \frac{1}{(\sqrt{2\pi})^N |C_X|^{1/2}}\, e^{-\frac{1}{2}(X-\mu)^T C_X^{-1}(X-\mu)} \tag{S19}$$

where $X = [X_1, X_2, \cdots, X_N]$ is the series, μ is the mean, C_X is the covariance matrix of the series, and $|C_X|$ is its determinant. This is a commonly used assumption and a reasonable one thanks to the central limit theorem. By substituting (S19) in (6), it could be shown that the joint entropy of this multivariate Gaussian process is given by (Cover and Thomas 2006, Theorem 8.4.1):

$$H(X) = \frac{1}{2} log|2\pi e C_X| = log\{|2\pi e C_X|^{1/2}\} \quad \text{bits} \tag{S20}$$

where $|2\pi e C_X|$ is the determinant of the matrix $2\pi e C_X$.

For daily noise series with *N*=1095, 1825, 2555, 3650, 5000, 6205, 7300, 8395 and 10000 days, and given the simulated noise as described in section 4.1, we formed the covariance matrices, C_X, of the noise according to equations (S15-S18) of section S3 (Williams 2003). For every *N*, we scaled this covariance matrix by the constant $2\pi e$ and computed a singular value decomposition of the result, which lead to a vector of all its *N* singular (or eigen) values, $\Lambda = (\lambda_1, \lambda_2, \cdots, \lambda_N)$. Needless to say, this operation is computationally expensive. The determinant of this scaled covariance matrix is then computed by multiplying its eigen values, but to avoid numerical issues, the joint entropy is computed by:

$$H(X) = \sum_{i=1}^{N} log_2 \sqrt{\lambda_i} \quad \text{(S21)}$$

and the results are shown in Fig S10. Notice the linear relation between the joint entropy, $H(X)$, and the number of observations, N. This functional relation, as estimated by an Excel spreadsheet, is:

$$H(X) = 4.5841N - 0.1489 \quad \text{(S22)}$$

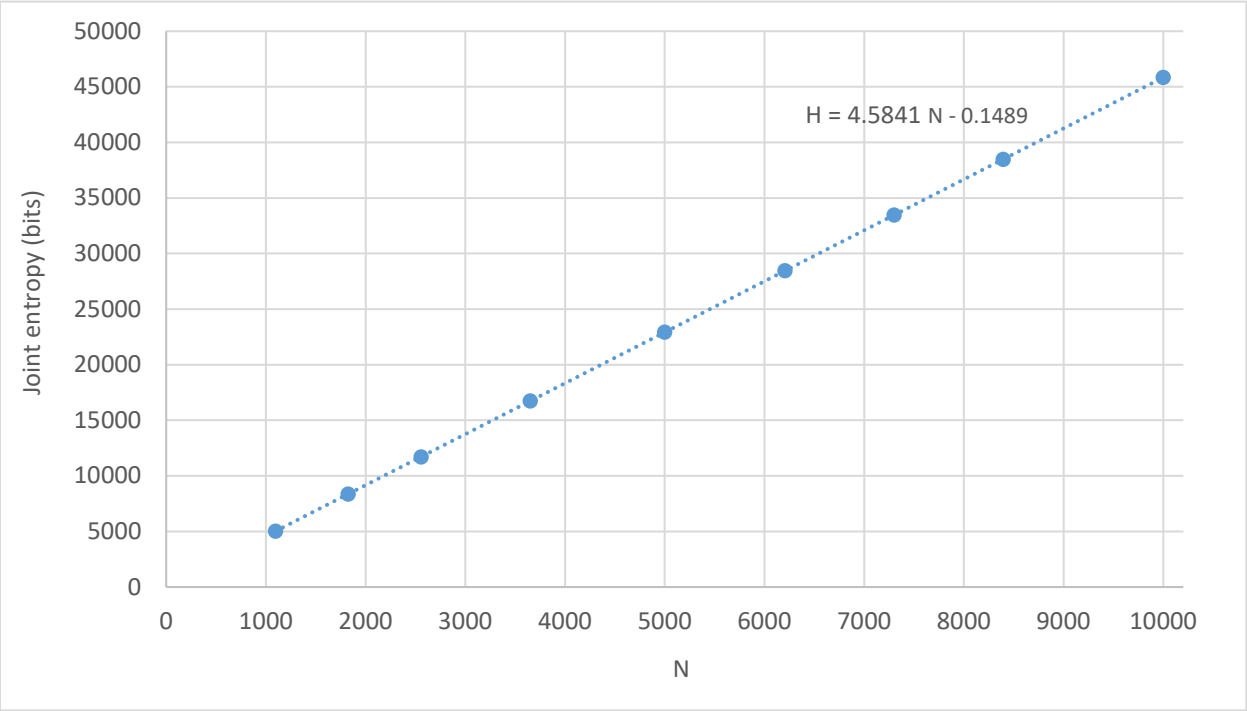

Fig S10: Joint entropy for the noise type described in section 4.1, based on the assumption that the stochastic process which generates the position time series is multivariate normal.

Table S1: A comparison of entropy rate and the corresponding velocity uncertainty: nonparametric M-Spacings versus a multivariate normal assumption

N	T (years)	Vasicek (Eq. 8, 9 and 16)		Multivariate Normal (Eq. S19, S20, 14 and 16)	
		Entropy rate	σ_V (mm/yr)	Entropy rate	σ_V (mm/yr)
1095	3	4.3914	1.693	4.5841	1.934
1825	5	4.4356	1.047	4.5841	1.161
2555	7	4.4664	0.764	4.5841	0.829
3650	10	4.4894	0.544	4.5841	0.580
5000	13.7	4.5200	0.405	4.5841	0.424
6205	17	4.5486	0.333	4.5841	0.341
7300	20	4.5623	0.286	4.5841	0.290
8395	23	4.5755	0.251	4.5841	0.252
10000	27.4	4.5927	0.213	4.5841	0.212

Substituting the expression (S22) of the joint entropy in (7) results in an entropy rate of 4.5841 bits. Table S1 lists the entropy rate values computed based on this expensive, determinant-based, multivariate normal assumption and its corresponding velocity uncertainty based on (14) and (16). It also lists the entropy rate based on the much more efficient M-Spacing approximation (8) and (9), and its corresponding velocity uncertainty. The differences in velocity uncertainty between the two are clearly negligible except perhaps for the very short time series, justifying the approximation we made by using (8) and (9).

S5) NGL time series used

We downloaded the position time series from http://geodesy.unr.edu/gps_timeseries/tenv/IGS14 (Blewitt et al, 2018).

The 171 linear and continuous NGL time series (Blewitt et al, 2018) are those of stations:
AB09 AB25 AB28 AB33 AB41 AB44 AB45 AB46 AC71 AL81 ALCA ALEB ALRE ARDQ AST1 AST2 ATQK AZUP AZWA BJAB BJKA BJNA BJNI BJPA BJSA BRFT BRW1 CBRG CCGN CCGS CCMK CHES CHL1 CHME CHR1 CN14 CN19 COBK COFC CONO CTI4 DCSO FLMK FLWE FTS1 GAAY GOGA HILO HNLC HTCC IACE IDNL IDSS INSY ISBS JCT1 JTNT KA14 KAR8 KWJ1 LANM LOLO LS01 LS08 LTHM LUMT MACM MAMI MAUI MC01 MC04 MDAI MDSI MEBA MERO MIDI MIPR MOMF MOWS MRC1 MTCB MTCU MTGW MTHM MTLG MTLO MTMI MTOP MTPJ MTZM NCKI NCRB NCSO NCSQ NCZO NDAS NDDI NDEL NDSK NDST NEGO NEOR NJBR NOMT NYHM NYIL NYMC OHRS ORS1 ORS2 OSPA P020 P025 P089 P100 P372 P739 P804 P805 P817 PAFM PAJP PATT PSC1 RIC1 SAV2 SCEB SCGT SCHA SCJR SCLC SDSF SDWE SE01 SIDC SMRT STB2 STBT SUP1 SUR1 TLDO TN23 TN25 TN31 TN35 TN36 TN44 TN49 TNCU TXBW TXMT TXRP UTMN VALY VARI VAST VAWI VDOT VITH VTBE VTEB VTHA VTSA VTUV WASR WIFH WLAX WMEL WSMN WVMO WVS6

The 55 NGL visibly nonlinear but continuous time series are:
AC37 AC72 ALDS ALFA ALLA ARGS AZCO BRI2 CHT1 CIC1 ECSD EFAY FLE5 FLE6 GDAC GODN INAS INAX INBR INES INLN INMD INMO INPD INRN INTC INVI INWB INWL INWN INWR KYDH KYTC KYTD MOAL MOBE MOBF MOBO MOHS MOPL MOSG OHCB P049 RG09 TCUN TN31 TXL1 VODG WHP1 WICH WICR WIHU WIKR WISU WLNC

The 50 discontinuous but linear NGL time series are:
ABQ6 ACU5 ALHC ANP5 ARP7 BAY5 BIL5 CACC CAE1 CARM CCV5 CHA1 COF1 COSG DET6 DRV6 ENG5 FAI1 FMTS GAAE GACU GWN5 HAC6 HAG6 HDF6 HRN6 HTV5 ILUC KAR4 KEW6 KOK5 KYTF KYW5 MC05 MCD5 MCD6 MCN5 MLF5 MNDT NCG5 NEB2 OMH5 PAAP PNB5 POR8 PPT5 RED5 SAV5 TN1B TND6