Predictability of fat-tailed extremes

Predictability of Fat-tailed Extremes

Tamás Bódai* and Christian Franzke†

We conjecture for a linear stochastic differential equation that the predictability of threshold exceedances (I) *improves* with the event magnitude when the noise is a so-called *correlated additive-multiplicative* (CAM) noise, no matter the nature of the stochastic innovations, and also improves when (II) the noise is purely *additive* obeying a *distribution* that *decays fast*, i.e., not by a power-law, and (III) *deteriorates* only when the *additive* noise *distribution follows a power-law*. The predictability is measured by a summary index of the receiver operating characteristic (ROC) curve. We provide support to our conjecture, to compliment reports in the existing literature on (II), by a set of case studies. Calculations for the prediction skill are conducted in some cases by a direct numerical time-series-data-driven approach, and in other cases by an analytical or semianalytical approach developed here.

Keywords: Fat tail, alpha-stable distribution, Predictability, ROC curve

I. INTRODUCTION

Extreme events in an observable are defined by some authors [1] as states realised by a process which reflect in such values of the observable that are distributed approximately by a *power-law*. It is usually the *tail* of the marginal distribution associated with the observable in question – what we will refer to as a *process distribution* – that follows a power-law. The reason for such a focus on *fat-tailed extremes* [2] only is that other types of events of the same rarity or frequency have a much smaller magnitude. One can compare, for example, a so-called standard Fréchet and a standard exponential random variable (rv) in this regard [3, 4]. In other words, fat-tailed extremes are more dangerous. For that reason their predictability is of great practical interest. Fat-tailed extremes commonly occur in natural or man-made phenomena. Examples include rain fall, relative vorticity, and stock market indices e.g. the DAX (German Stock Index) returns [1, 5–7].

The predictability of fat-tailed extremes have been addressed in the past by several authors, to be discussed next. One group of studies concern the predictability of threshold exceedances in stochastic or deterministic processes based on the so-called receiver operating characteristic (ROC) curve (Sec. II B). These studies treat fat-tailed extremes, but also other ones. One feature of the predictability is shared by the models and observables studied thus far rather surprisingly without exception: *larger magnitude events are better predictable.* This finding remains without an explanation up to now. A collection of these studies is listed in Table I. Separated by horizontal lines we also include studies that do not evaluate the prediction skill [8] or do not measure the predictability based on the ROC curve [9, 10]. Another study [11] that measures predictability by a mean squared error in terms of an ensemble forecast could fit all slots of Table I as its statement on predictability is based only on the process distribution, whether it is produced by a stochastic or a deterministic process. They find that predictability *deteriorates* in the limit of the highest thresholds for *any* of the Extreme Value Distributions. However, even those studies that rely on the ROC curve use different summary statistics or indices of it, such as (find definitions in Sec. II C 2): the 'initial' slope [12]; the area under the curve [13]; or the distance from perfect predictability [14], ROC-D in short; and some [15, 16] does not even evaluate a summary index.

We note that the phrasing "larger magnitude events are better predictable" suggests a nonasymptotic behaviour, and so it is somewhat imprecise. The more precise formulation of the research question is the following. Assume that the asymptotic limit of the prediction skill when increasing the threshold level beyond any limit is nontrivial, i.e., the skill does not vanish, nor does it become maximal, like in a deterministic system with arbitrary amount, precision and accuracy of data, and using ideal precursory variables that fully embed the attractor [14]. In this case the question is whether the limit is approached from above or below. In other words: is the lowest-order term of an asymptotic expansion positive or negative? This meaning is implied implicitly by the approach of [12]. Nevertheless, for convenience we keep using the original phrasing too synonymously with the latter more precise but also lengthier phrasing.

As listed in Table I, we are aware of two studies that examine the predictability of fat-tailed extremes using the ROC curve. The subject of neither of these studies is a process that is defined explicitly by an equation. One of them [12] considers a process that was defined implicitly by a time series that was constructed from a time series of a Gaussian AR(1) process as follows. They generated an auxiliary time series consisting of independent realisations of a symmetrized Pareto rv (Sec. IV B). They then replaced the n'th largest value of the original time series

* Department of Mathematics and Statistics, University of Reading, Reading, UK; T.Bodai@reading.ac.uk; At the time of first submission the author was affiliated to the Meteorological Institute, University of Hamburg, where the research was carried out.

† Meteorological Institute, University of Hamburg, Hamburg, Germany

TABLE I. Studies on the predictability of extremes. Along the horizontal dimension we make distinction regarding the nature of the laws governing the studied process; and along the vertical dimension we distinguish between processes whose marginal or process distribution does or does not feature a fat tail.

	Stochastic	Deterministic
No fat tail	Gaussian AR(1) [12] Atmospheric dynamics [15]	Atmospheric dynamics [15] Lorenz 84 [14] System of coupled FitzHugh-Nagumo units [16] Electronic circuits [8] Geophysical models [9]
Fat tail	Rescaled Gaussian AR(1) [12] Social media [13] Atmospheric dynamics [10]	-

with the n'th largest one from the auxiliary time series. According to the authors such a process has a 'step-wise correlation factor' (with a reference to the parameter of a common AR(1) process) varying in time. They studied this process only numerically. The other study [13] examined fat-tailed time series data to do with the attention received by social media content or scientific publications. For prediction they used the category of the content as a precursor (Sec. II B). Concerning e.g. YouTube videos categories may be: music, sport, cat fails, etc. They found an improving predictability with the event magnitude, and explained it with the difference in the power exponent governing the fat tails of the different distributions conditioned on the category (in our parlance the conditional probability defined by eq. (21)). This is the only theory explaining the said improving predictability that we are aware of. However, it is applicable to a rather specific problem only. A difference in the power exponent seems to be possible for the banal reason that the distinct content categories entail distinct processes.

For the first time regarding the predictability of extremes we will consider in this note an AR(1)-type process that is driven by – instead of a additive Gaussian rv as in case of the 'common' AR(1) – a fat-tailed so-called α-stable or simply stable rv [17]. We will show that the process distribution inherits the power exponent $(1+\alpha)$ of the noise distribution. We find for this process, the first of its kind, that the predictability of threshold exceedances *deteriorates* with increasing threshold level, or, the limit of the ROC-D is approached from below. This is not in contradiction with the finding of [13], because the probability distributions conditioned on different values of the precursor of our choice all inherit the same power exponent.

In search for a theory to explain this finding, we develop an analytical approach (Sec. II C 2) to evaluate the prediction skill as a function of the threshold level. Unfortunately we do not find it a productive way forward in its generic form, and the finding remains unexplained. However, a semianalytical (partially analytical) version of our approach facilitates the evaluation of the prediction skill with a high accuracy, for which reason asymptotic power-laws are confidently detected. Besides, this is numerically much more efficient than the data-driven direct numerical approach (Sec. II C 1).

Furthermore, we are able to disprove a few intuitive propositions as to why the predictability might improve or deteriorate with the event magnitude. It does not improve – when it does – because of a decay of the process distribution (Sec. III), nor does it necessarily deteriorate when the noise distribution is fat-tailed (Sec. V). The existence of the variance of the noise distribution is also not decisive (Sec. IV B). These falsifications can, of course, be facilitated by a pertinent finding in a single case each. But, our case studies, however systematic, cannot imply in a mathematical sense 'positive' universal statements. Nevertheless, they compel us to make a conjecture on the conditions of improving or deteriorating predictability. It is spelled out in terms of a Venn diagram shown in Fig. 1.

II. METHODOLOGY

A. Examined processes

We examine processes that are governed by the following linear stochastic differential equation (in the Itô form):

$$dx = (ax+b)dt + (cx+d)dX, \tag{1}$$

where dX is an infinitesimal increment of an 'input' stochastic process. We can write W in place of X for the Wiener process, which is the integral of a Gaussian white noise process. The process probability density distribution function (PDF) can be found by integrating the Fokker-Planck (FP) equation [18] (for details see eq. (5.13) on page 98 of [19]) as:

$$p(x) = N_0 \frac{2e^{2\frac{ad-bc}{c^2(d+cx)}}}{(d+cx)^{2(1-a/c^2)}}, \tag{2}$$

where N_0 is a normalization constant. Since p as a probability is nonnegative, the lower boundary of the domain is

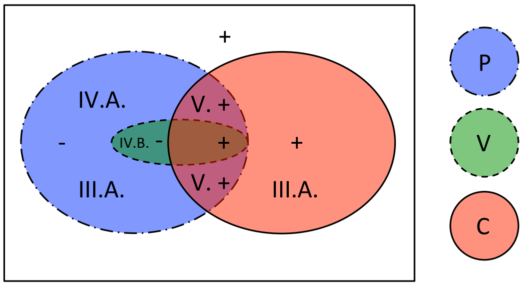

FIG. 1. Schematic depiction of a conjecture on the predictability of threshold exceedances in the linear SDE (1). Round shapes represent sets of models of certain properties identified by capital letter codes, with matching colors or line type of borders, as follows: P – the noise distribution features a power-law tail; V – the variance of the noise distribution does *not* exist; C – the noise is of the CAM-type ($c \neq 0$). The rectangular shape is a representation of all possibilities (except that $b = 0$). In each region defined by a unique combination of these properties we put a + or − sign according to whether the asymptotic limit of the predictability with ever-increasing event magnitude is approached from above or below. (Note that a larger value of the measure of predictability measured by the ROC-D (17) means poorer predictability.) In these regions a code stands for the section number where a particular scenario is examined. The regions in color represent fat-tailed extremes.

$-d/c$. Also, for a well-posed problem we need $ad - bc < 0$, when $\lim_{x \to -d/c} p(x) = 0$, as it is ∞ otherwise. Fixing $b = 0$ makes no restriction on the qualitative behaviour. Furthermore, it is clear that for $c \neq 0$ this distribution features a fat tail. The noise term $(cx + d)dW$ in eq. (1) is referred to by some authors [6, 10, 20] as the 'correlated additive-multiplicative' (CAM) noise. In this paper we will refer to processes in this form as P1-type processes. The Wiener process W can be replaced by the more generic Lévy process: $dX = dL$. The process PDF in this case can be found by solving the fractional Fokker-Planck (fFP) equation [21].

In the special case when only additive noise is retained, $c = 0$, eq. (1) is known as the Ornstein-Uhlenbeck (OU) equation. It can be shown [22] that by taking the limit $c \to 0$ in eq. (2), $p(x)$ will take the specific form of the normal distribution that is known to be the solution for the OU equation. However, integrating the FP equation with setting $c = 0$ beforehand will lead to a Gaussian process PDF via a more straightforward calculation.

To achieve a fat-tailed process with simple additive noise, the noise itself need to be fat-tailed. We will refer to such processes as P2-type processes. As an alternative way to solving the fFP equation, in order to show how the fat tail of the noise is inherited, first we discretize eq. (1). A discretized form can be obtained by applying some stochastic integrator scheme [23]. A simple choice is the Euler-Maruyama scheme, with which we have:

$$x_n = x_{n-1} + (b + a x_{n-1})\Delta t + (d + c x_{n-1})\sqrt{\Delta t}\xi_{n-1}. \quad (3)$$

With $c = 0$ (which is assumed in the rest of this subsection, unless otherwise said), eq. (3) is known as the auto-regressive model of order 1, AR(1) in short. If ξ_{n-1} are realisations of an α-stable rv, then the x_n's are also α-stable rv's (with the difference that it is a correlated sequence). For example, if $p_\xi(\xi) = p_s(\xi; \alpha, \beta = 0, \gamma_\xi, \delta = 0)$, $0 < \alpha \leq 2$, $0 < \gamma_\xi$ is an unskewed ($\beta = 0$) α-stable distribution, or in short 'stable distribution', then $p(x) = p_s(x; \alpha, \beta = 0, \gamma_x, \delta = 0)$ is also an unskewed stable distribution. The reason for this is that from eq. (3) x_n emerges as a weighted infinite sum of stable variables ξ_n, and that the stable distribution is a so-called infinitely divisible distribution. Equations (1.8) of [17] provide formulae for the parameters of a stable distribution of such a composed rv. From these it follows that the tail behaviour of the noise PDF $p_\xi(\xi)$ is inherited by the process PDF $p(x)$ in that they share the same stability or shape parameter α. Furthermore, the process and noise scale parameters have the following relation:

$$\gamma_x = \gamma_\xi / \sqrt[\alpha]{1 - \varphi^\alpha}, \quad (4)$$

where we introduced $\varphi = 1 + a\Delta t$. For stationarity $|\varphi| < 1$ is required.

In the special case of $\alpha = 2$ the stable distribution is a normal distribution, whose variance exists unlike any other fat-tailed stable distribution ($\alpha < 2$) and is given by the scale parameter γ. Then, clearly, eq. (4) relates the process and noise variances, which is a well-known property of the OU process/Gaussian AR(1). It holds exactly also if ξ_n obeys some other than the normal distribution whose variance exists. Furthermore, if ξ_n is a stable rv, the observed or sample variances s_x and s_ξ of *finite* process and noise time series, respectively, obey not eq. (4) but $s_x = s_\xi / \sqrt{1 - \varphi^2}$.

When $\varphi \lessapprox 1$, the Generalized Central Limit Theorem (GCLT) implies that $p(x) \approx p_s(x; \alpha < 2, \beta = 0, \gamma_x, \delta = 0)$ for any fat-tailed $p_\xi(\xi) \sim \alpha \xi_m^\alpha / \xi^{\alpha+1}$ symmetric around $\xi = 0$, whose variance does not exist, where the relationship between the scale parameter of the asymptotically equivalent classical Pareto distribution (a paradigmatic model for fat-tailed distributions) and γ_x we provide here as:

$$\gamma_x = \frac{\xi_m}{\sqrt[\alpha]{(1 - \varphi^\alpha)\sin(\pi\alpha/2)\Gamma(\alpha + 1)/(\pi\alpha)}}. \quad (5)$$

In the above we used the asymptotic behavior of a stable distribution given by Theorem 1.12 of [17] and eq. (4). Note that we have $\varphi \lessapprox 1$, or more precisely $0 < 1 - \varphi \ll 1$, when $0 < \Delta t \ll -1/a$. The latter means that the time step size needs to be much smaller than the time scale of the linear deterministic part. In fact this is needed also for the discrete-time process (3) to reproduce approximately the process PDF of the continuous-time process (1). The latter is related to the *weak convergence* properties of the stochastic numerical integrator scheme [23]. Note, however, that for e.g. eq. (4) alone to hold no condition on φ is imposed.

Note also that the situation when $\varphi < 1$ significantly can be interpreted in a way that the discrete mapping establishes a connection between states at times further apart, say, some k multiple of Δt, such as: $x_n = \tilde{\varphi}^k x_{n-k} + \sum_{i=1}^{k} \tilde{\varphi}^{i-1} \xi_{n-i}$, provided that $b = c = 0$. Even if $\tilde{\varphi} \lessapprox 1$, $\varphi = \tilde{\varphi}^k < 1$ significantly is possible for a large enough k. This way we can conveniently examine the predictability depending on the prediction lead time. Such a dependence in an SDE with a nonlinear deterministic part has been found nontrivial, featuring a *return of skill* and a reversal of the tendency of the predictability depending the event magnitude [14]. However, we do not expect such effects with a linear deterministic part considered here, and so did not make investigations in this direction.

The lack of process skewness is due to the fact that reversing the signs of x_{n-1} and the symmetrically distributed ξ_{n-1}, the same process equation will result in $-x_n$. Clearly, this is not the case when $c \neq 0$, which is reflected in the generically skewed form under (2).

B. Prediction scheme and skill

Our aim is to predict large excursions of some (scalar) physical observable x, exceeding a chosen threshold level x_*, before that exceedance happens. Figure 2 pictures the situation as the observable is sampled *discretely* in time, $x_n = x(t_n = n\Delta t)$, $n \in \mathbb{Z}$, occasionally exceeding the threshold.

The following methodological description regarding the prediction task closely follows [14, 24]. We introduce a *binary* event variable:

$$\chi_n = \begin{cases} 1, \; x_n > x_* \\ 0, \; x_n < x_* \end{cases} \tag{6}$$

The prediction is based on a precursory structure $\mathbf{x}_n \in \mathbb{R}^M$ of size M, whose different members, observables desirably related to x, may belong to different times, e.g. t_{n-d_m}, preceding the current time t_n, specified by delays $d_m \in \mathbb{Z}$, $m = 1, \ldots, M$. We call $t_n - t_{n-\min(d_m)} > 0$ the prediction lead time. Our binary prediction for χ_n at $t_{n-\min(d_m)}$ is defined as:

$$\hat{\chi}_n = \begin{cases} 1, \; \mathcal{L}(\mathbf{x}_n) > \mathcal{L}_* \\ 0, \; \mathcal{L}(\mathbf{x}_n) < \mathcal{L}_* \end{cases} \tag{7}$$

based on the *likelihood* function:

$$\mathcal{L}(\mathbf{x}) = \mathbb{P}_{\chi|\mathbf{x}}(\chi = 1, \mathbf{x}) = \mathcal{P}(\mathbf{x})/p(\mathbf{x}). \tag{8}$$

In the above $\mathcal{P}(\mathbf{x}) = p_{\mathbf{x}|\chi}(\mathbf{x}, \chi = 1)\mathbb{P}_\chi(\chi = 1)$ is the *posterior* PDF of $\mathbf{x}$, and $p(\mathbf{x})$ is the process PDF in terms of the precursory variables, i.e., the basic PDF generated by the considered process [25]. Refer to the appendix of [26] for an integral formulation of e.g. $\mathcal{P}(\mathbf{x})$ which applies the Heaviside step function as a filter. Note that Eq. (8) expresses Bayes' theorem relating the conditional probabilities: the likelihood and the posterior probability. Our prediction $\hat{\chi}_n$ is controlled by a threshold $\mathcal{L}_* \in [\min(\mathcal{L}), \max(\mathcal{L})]$ of stringency on $\mathcal{L}$. Note that an actual choice is meant to be made as to the applied value of $\mathcal{L}_*$ in practice, for which reason this kind of prediction is not probabilistic, but we call it a *categoric prediction.*

TABLE II. Realised and predicted threshold exceedance event probabilities.

		χ	
		1	0
$\hat{\chi}$	1	$\mathcal{A}$	$\mathcal{B}$
	0	$\mathcal{C}$	$\mathcal{D}$

To quantify the prediction skill the rate of true positives, or the *hit rate*, i.e., the frequency of making correct predictions, which depends on $\mathcal{L}_*$, can be defined [27, 28] as:

$$H(\mathcal{L}_*) = \frac{\mathcal{A}}{\mathcal{A} + \mathcal{C}}, \tag{9}$$

and the *false alarm rate* can be defined as:

$$F(\mathcal{L}_*) = \frac{\mathcal{B}}{\mathcal{B} + \mathcal{D}}. \tag{10}$$

In the above $\mathcal{A}, \mathcal{B}, \mathcal{C}, \mathcal{D}$ are the frequencies or probabilities [29] of situations specified by different combinations of χ and $\hat{\chi}$, which can be collected, following [28], in a table format shown in Table II. These probabilities we express by the following integrals:

$$\mathcal{A}(\mathcal{L}_*) = \int_{\mathbb{R}^M} dV_{\mathbf{x}} \mathcal{P}(\mathbf{x}) \mathcal{H}(\mathcal{L}(\mathbf{x}) - \mathcal{L}_*), \tag{11}$$

$$\mathcal{B}(\mathcal{L}_*) = \int_{\mathbb{R}^M} dV_{\mathbf{x}} (p(\mathbf{x}) - \mathcal{P}(\mathbf{x})) \mathcal{H}(\mathcal{L}(\mathbf{x}) - \mathcal{L}_*), \tag{12}$$

$$\mathcal{C}(\mathcal{L}_*) = \int_{\mathbb{R}^M} dV_{\mathbf{x}} \mathcal{P}(\mathbf{x}) \mathcal{H}(\mathcal{L}_* - \mathcal{L}(\mathbf{x})), \tag{13}$$

$$\mathcal{D}(\mathcal{L}_*) = \int_{\mathbb{R}^M} dV_{\mathbf{x}} (p(\mathbf{x}) - \mathcal{P}(\mathbf{x})) \mathcal{H}(\mathcal{L}_* - \mathcal{L}(\mathbf{x})). \tag{14}$$

In the above $dV_{\mathbf{x}}$ is a volume element in the precursory space; and $\mathcal{H}(\cdot)$ is the Heaviside step function. Upon

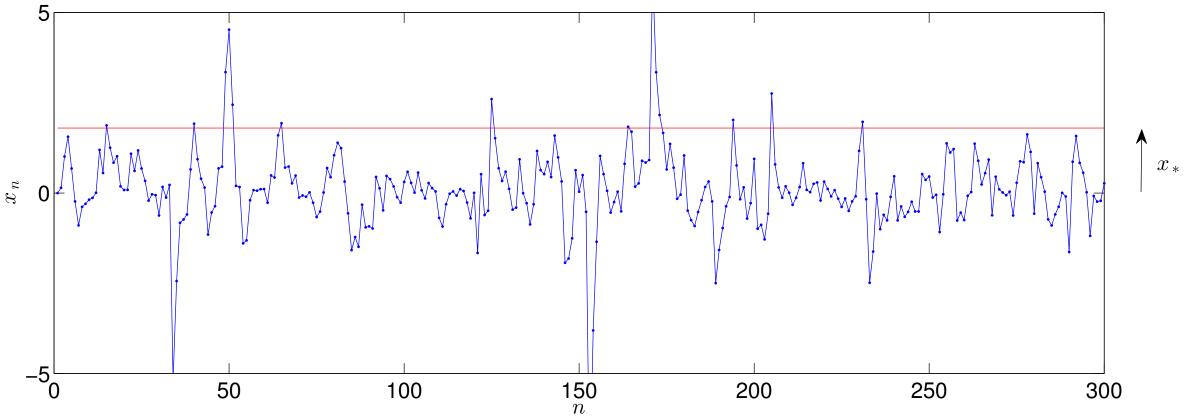

FIG. 2. Illustration of the prediction problem: given a discrete-time irregular time series x_n that every so often overshoot a relatively high threshold x_* (straight horizontal line), we want to predict these overshoots, i.e., extreme events, at the time, say, of the immediately preceding observation. The discrete data points are connected by straight lines to indicate their order in time.

detailing $\mathcal{A}, \mathcal{B}, \mathcal{C}, \mathcal{D}$, the hit and false alarm rates take the following forms:

$$H(\mathcal{L}_*) = \frac{\int_{\mathbb{R}^M} dV_{\mathbf{x}} \mathcal{P}(\mathbf{x}) \mathcal{H}(\mathcal{L}(\mathbf{x}) - \mathcal{L}_*)}{\int_{\mathbb{R}^M} dV_{\mathbf{x}} \mathcal{P}(\mathbf{x})}, \quad (15)$$

$$F(\mathcal{L}_*) = \frac{\int_{\mathbb{R}^M} dV_{\mathbf{x}} [p(\mathbf{x}) - \mathcal{P}(\mathbf{x})] \mathcal{H}(\mathcal{L}(\mathbf{x}) - \mathcal{L}_*)}{1 - \int_{\mathbb{R}^M} dV_{\mathbf{x}} \mathcal{P}(\mathbf{x})}, \quad (16)$$

where only $\mathcal{H}(\mathcal{L}(\mathbf{x}) - \mathcal{L}_*)$ appears.

A parametric plot or curve $\{(F(\mathcal{L}_*), H(\mathcal{L}_*))\}$ is referred to as the *receiver operating characteristic* (ROC) *curve* [30]. With the extremal choices, $\mathcal{L}_* = 0$ and 1, we have $(F = 1, H = 1)$ and $(F = 0, H = 0)$, respectively, i.e., the ROC curve stretches from corner to corner. It is a diagonal straight line with no prediction skill at all (over random predictions $\hat{\chi}$ with $\mathbb{P}(\hat{\chi} = 1) = 1 - \mathcal{L}_*$), and situated above the diagonal with any skill.

The ideal situation when extreme events $(\chi = 1)$ and nonevents $(\chi = 0)$ can be predicted with certainty $(\hat{\chi} = \chi)$ is represented by the $(F = 0, H = 1)$ corner in the ROC diagram. In this case no choice has to be made on the applied stringency $\mathcal{L}_*$. In the nonideal situation an optimal $\mathcal{L}_*$ is to be chosen. This is always the case when the precursory space does not embed the attractor unambiguously, or, when the equations governing the process feature randomness. A unique optimum exists only in terms of a single-objective optimization problem, defined by a scalar-valued cost function. However, in our case the minimization of the false alarm rate and the maximization of the hit rate are both 'valid' objectives. It takes a *specific application* to be possibly able to define a scalar-valued cost function $C(F, H)$. For our *general assessment* of predictability we choose to consider the intuitive measure:

$$D^{opt} = \min_{\mathcal{L}_*}(\sqrt{F^2 + (H-1)^2}), \quad (17)$$

the smallest or optimal distance of the ROC curve from the ideal corner. With no prediction skill at all: $D^{opt} = \sqrt{2}/2$. Other summary statistics for the ROC curve have also been defined, such as the area under the curve [24], or the slope $H'_F(F = 0)$ [12]. Unlike these two, the distance D^{opt} can be associated to actual predictions specified by an actual choice $\mathcal{L}_*^{opt}$ for $\mathcal{L}_*$. We note that it is not trivial to interpret what the comparison of D^{opt} with a proper skill score of probabilistic prediction [24] means.

C. Evaluation of skill

1. Direct numerical approach

Perhaps the most obvious factor that compromises data-driven prediction skill is the finite size N of the data set: $\{x_n, \mathbf{x}_n\}$, $n = 1, \ldots, N$. The distributions $p(\mathbf{x})$, $\mathcal{P}(\mathbf{x})$, $\mathcal{L}(\mathbf{x})$ will be approximated in our study by histograms $\{p_b\}$, $\{\mathcal{P}_b\}$, $\{\mathcal{L}_b\}$, $b = 1, \ldots, B$; different values of b can be assigned to the different bins arbitrarily, it serves only the purpose of identification. Note that in the coarse-grained situation $\{\mathcal{L}_b\}$ derives from $\{p_b\}$ and $\{\mathcal{P}_b\}$ much the same way as with the continuous functions according to Eq. (8). With the discrete formulation of Eqs. (15) and (16), accordingly, the ROC curve turns into (the graph of) a staircase (function), given by a set of discrete data points: $\{(H_b, F_b)\}$, $b = 1, \ldots, B$, belonging to stringency levels $\{\mathcal{L}_{*,b}\} = \{\mathcal{L}_b\}$:

$$H_b = \frac{\sum_{b'=1}^{B} \mathcal{P}_{b'} \tilde{\mathcal{H}}(\mathcal{L}_{b'} - \mathcal{L}_b)}{\sum_{b'=1}^{B} \mathcal{P}_{b'}} \tag{18}$$

$$F_b = \frac{\sum_{b'=1}^{B} (p_{b'} - \mathcal{P}_{b'}) \tilde{\mathcal{H}}(\mathcal{L}_{b'} - \mathcal{L}_b)}{N - \sum_{b'=1}^{B} \mathcal{P}_{b'}} \tag{19}$$

Note first that in the above the bin counts of the histograms $\{p_b\}$ and $\{\mathcal{P}_b\}$ are assumed to be *not* normalized (otherwise one should write 1 in place of N); and e.g. $\{\mathcal{L}_b\}$ and $\{\mathcal{L}_{b'}\}$, $b, b' = 1, \ldots, B$, denote the same set. Second, if $\mathcal{L}_b$ does not', then neither do H_b and F_b change monotonically with increasing b. Third, $\tilde{\mathcal{H}}(\cdot)$ denotes a Heaviside-like function with the only difference that $\tilde{\mathcal{H}}(0) = 0$ (not 1/2).

The above estimation of the measures of skill is *not* conservative [31], which is to do with finite histogram bin counts for $\mathcal{P}(\mathbf{x})$ and associated statistical errors. An approach to fix this problem is the following. The available data is divided equally into 'training' and 'evaluation' data sets. Then, the conservative estimates are defined again by Eqs. (18) and (19), but the different terms appearing in them are associated with different data sets: $\{\mathcal{L}_b\}$ is derived from the training data set, and $\{p_b\}$ and $\{\mathcal{P}_b\}$ are derived from the evaluation data set. Note that the latter requires the use of the same grid forming the bins in case of the training and evaluation data sets.

A further issue to do with small bin sizes when many bins of $\mathcal{P}(\mathbf{x})$ contain a single data point is that the 'ROC staircase' can have an excessively large *last* step. This is so, because bins that contain single data points tend to have empty counterparts mutually between the 'training' and 'evaluation' data sets. This way $D^{opt} > \sqrt{2}/2$ can even be realised.

Too large bin sizes would of course also deteriorate the prediction skill. Therefore, there should be an *optimal* bin size yielding (locally) minimal D^{opt}. Our numerical experience shows that there is always, for any given prediction lead time or threshold level x_*, a unique (globally) optimal *uniform* bin size. Beside a regular grid one can also use e.g. an irregular grid such that the same number N/B of data points fall in each bin. In this case there is an optimal choice for N/B, and so for B.

We adopt the latter approach. The irregular high dimensional grid (when $M > 1$) can be created iteratively, treating one dimension of the precursory variable space at a time. We note that the grid, i.e., the configuration of bins, depends on the order of precursory variables considered in the iterative process. We do not claim that any of these grids is the optimal grid. That is a question to be investigated elsewhere.

2. *Analytical and Semianalytical approaches*

To be able to describe a 'mechanism' that produces some properties of $D^{opt}(x_*)$, it would be desirable to construct this functional relationship analytically. Next we outline the formulae for this, considering processes governed by eq. (3), when the aim is to predict threshold exceedances of x_n based on a single precursory variable $\mathbf{x}_n = x_{n-1}$. Equations (8)-(16) indicate that we need two things:

(i): The likelihood function $\mathcal{L}(\mathbf{x}; x_*)$, and

(ii): the process PDF $p(\mathbf{x})$.

The process PDF we have already discussed in Sec. II. The likelihood function, on the other hand, we can write for a scalar precursory variable as:

$$\mathcal{L}(x_{n-1}; x_*) = \int_{x_*}^{\infty} dx_n p_{x_n|x_{n-1}}(x_n, x_{n-1}), \tag{20}$$

in which the conditional probability can be written as:

$$\begin{aligned} &p_{x_n|x_{n-1}}(x_n, x_{n-1}) = \\ &\quad N_0(x_{n-1}; \Delta t) p_\xi(\xi_{n-1}(x_n, x_{n-1}; \Delta t)). \end{aligned} \tag{21}$$

upon expressing the noise variable $\xi_{n-1} = \xi_{n-1}(x_n, x_{n-1}; \Delta t)$ – whose distribution $p_\xi(\cdot)$ we know – from the process equation (3) as:

$$\begin{aligned} &\xi_{n-1}(x_n, x_{n-1}; \Delta t) = \\ &\quad \frac{x_n - x_{n-1} - (b + a x_{n-1})\Delta t}{(d + c x_{n-1})\sqrt{\Delta t}}. \end{aligned} \tag{22}$$

In eq. (21) the normalization factor derives from the following basic property of a conditional probability:

$$\begin{aligned} &\int_{-\infty}^{\infty} dx_n p_{x_n|x_{n-1}}(x_n, x_{n-1}) = \\ &\quad \frac{N_0(x_{n-1}; \Delta t)}{(d + c x_{n-1})\sqrt{\Delta t}} = 1. \end{aligned} \tag{23}$$

The likelihood function can in fact be formulated, thanks to the factorization of additive terms of $\xi_{n-1}(x_n, x_{n-1}; \Delta t)$ wrt. x_n and x_{n-1}, by the noise cumulative probability distribution function (CDF) as:

$$\mathcal{L}(x_{n-1}; x_*) = 1 - F_\xi(\xi_{n-1}(x_*, x_{n-1})), \tag{24}$$

in which the normalization factor N_0 does not appear.

The numerator of the expression for the hit rate under (15), for example, can be written as:

$$\mathcal{A}(x_*, \mathcal{L}_*) = \int_{x_L(x_*, \mathcal{L}_*)}^{\infty} dx_{n-1} \mathcal{L}(x_{n-1}; x_*) p(x_{n-1}), \tag{25}$$

where the lower limit of integration is found by solving

$$\mathcal{L}(x_L; x_*) = \mathcal{L}_* \tag{26}$$

for x_L. Then the shortest distance D^{opt} of the ROC curve from perfect predictability in terms of an optimal $\mathcal{L}_*^{opt}$ can be obtained by solving the equation:

$$D'(\mathcal{L}_*^{opt}) = 0. \tag{27}$$

The alternative summary index the area under the ROC curve can be obtained by the integral:

$$\int_0^1 d\mathcal{L}_* H(\mathcal{L}_*) F'(\mathcal{L}_*); \tag{28}$$

or the 'initial' slope of the ROC curve is:

$$S = \lim_{\mathcal{L}_* \to 1} H'(\mathcal{L}_*)/F'(\mathcal{L}_*). \tag{29}$$

We are not aware of a model of the form (3) for which $D^{opt}(x_*)$ can be obtained analytically, even approximately relying on the GCLT, avoiding the use of stable noise variables whose PDF is in general not in an analytic form. We have been able to obtain $F(\mathcal{L}_*)$ and $H(\mathcal{L}_*)$ analytically in some cases, but with any attempt using the software package Mathematica we failed to carry out either (27) or (28). However, even if it was possible, it is doubtful how much it could aid further studies, as the expressions already for $F(\mathcal{L}_*)$ and $H(\mathcal{L}_*)$ are extremely complicated. In this respect it is not much help that the slope S (29) is a more simple quantity than D^{opt} or AUC in that the latter require in addition either solving a nonlinear equation or carrying out an integral. A further problem with S is that it could be less informative too. For example, if the limiting ROC curve as $x_* \to \infty$ is in the linear form of $H = c_1 + c_2 F$, $c_1, c_2 \in [0, 1]$, then the limit of S does not exist, and so it could not indicate a nontrivial, less than perfect predictability as specified by $c_1 \neq 1$, $c_2 \neq 0$. This is in fact the case with unskewed stable or symmetrized Pareto noise distributions (and $c = 0$), as introduced in Sec. IV.

Even if a fully analytical treatment is not possible, the direct numerical approach based on (long) time series data (Sec. II C 1) is not the only alternative. We can still make use of eqs. (20)-(27) etc. by evaluating them numerically. We will refer to this in the following as a *semianalytical approach*. With this approach we can evaluate D^{opt} at a desired x_*. It turns out that a choice of x_* has its limits regarding the accuracy of calculations due to finite machine precision. What is important is that a sufficiently large range of x_* is considered in which the asymptotic behaviour shows up, and if it is a scaling behaviour, then an accurate estimation of the scaling exponent be possible.

In our numerics D^{opt} (and $\mathcal{L}_*^{opt}$) is found by Matlab's `fminbnd` (an approximate solution of eq. (27)) for which $D(x_*, \mathcal{L}_*)$ is calculated by numerical integration of expressions like (25) using Matlab's `integral` for any $\mathcal{L}_*$ required by `fminbnd` and a readily fixed x_*. Equation (26) is solved numerically by Matlab's `fzero`. We note that the procedure is greatly expedited by using eq. (24) avoiding the numerical integration of the conditional probability (21) as dictated by (20). This is particularly useful when considering stable rv's of α for which the stable distribution does not have an analytic form (as in Sec. IV) but it has to be numerically computed by e.g. Matlab's `makedist`.

III. PROCESS DISTRIBUTION NOT DECISIVE

It could be thought that (i) a decaying $D^{opt}(x_*)$, as has been found in all the situations ever examined that we are aware of and reported on in Sec. I, is due to a decaying process distribution $p(x)$, or, (ii) a reducing frequency of events [32]. Or (iii) it could be thought that an increasing $D^{opt}(x_*)$ is due to some other property of the process distribution alone again, e.g., a fat tail (Frechét-type extreme value distribution [3, 4]). We will demonstrate in this section that, in relation with (iii), $D^{opt}(x_*)$ can be decreasing even if the process distribution is fat-tailed (Sec. III A), and, in relation with (i), it can be increasing even if the process distribution is decaying (Sec. III A), or, in relation with (ii), if the frequency of events is increasing (Sec. III B).

A. Two processes with identical process distributions

First we give the examples of two processes which feature the same process distribution, yet $D^{opt}(x_*)$ approaches its limit value from above in one case and from below in the other. This will indicate that something else than the process distribution, precipitated in some likelihood function, is also at work [33]. One of the processes is a P2-type process, as introduced in Sec. II A, driven by a stable rv of $p_\xi(\xi) = p_s(\xi; \alpha = 1/2, \beta = 1, \gamma_\xi, \delta_\xi)$. The special case of $\alpha = 1/2$ is called a Lévy distribution, one of two stable distributions that take analytic forms expressible by elementary functions:

$$\begin{gathered} p_L(x; \mu, \gamma) = p_s(x; \alpha = 1/2, \beta = 1, \gamma, \delta = \gamma + \mu) = \\ \frac{\gamma}{2\pi} \frac{e^{-\frac{\gamma}{2(x-\mu)}}}{(x-\mu)^{3/2}}, \end{gathered} \tag{30}$$

$\mu < x < \infty$. Notice that the skewness parameter of this stable distribution takes on its largest possible value, $\beta = 1$, so that the distribution looses the power-law tail on the left and becomes even bounded on that side. Such a noise produces a Lévy P2-type process distribution $p_x(x) = p_s(x; \alpha = 1/2, \beta = 1, \gamma_x = \gamma_\xi / \sqrt[\alpha]{1 - \varphi^\alpha}, \delta_x = \delta_\xi/(1-\varphi))$, as can be derived from eqs. (1.8) of [17].

We point out that the (analytic) process PDF (2) with appropriate parameters is in the form of the Lévy distribution (30). The appropriate parameters we find to be: $a = 1/4$, $b = 0$, $c = 1$, $d = -1$, with which $\mu = 1$, $\gamma = 1$. Therefore, in order to match the P1 and P2 Lévy process distributions, for P2 with, say, $\varphi = 1/2$ we need noise parameters: $\gamma_\xi = (1 - 1/\sqrt{2})^2$, $\delta_\xi = 1/2$.

For the described P1 and P2-type processes we see $D^{opt}(x_*)$ and the corresponding $\mathcal{L}_*^{opt}(x_*)$ diagrams in Figs. 3 and 4, respectively. In the same diagrams direct numerical (Sec. II C 1) and semianalytical results (Sec. II C 2) are plotted. They match rather closely. The different diagrams of $D^{opt}(x_*)$ are as forecast in the beginning of this section. We note that in both cases $\mathcal{L}_*^{opt}(x_*)$ are

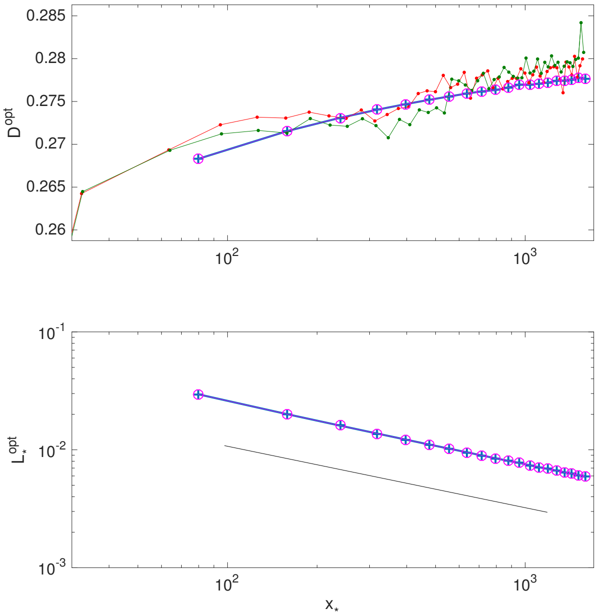

FIG. 3. Predictability of extremes in a P2-type process (Sec. II A) using two different fat-tailed noise variables of power exponent 3/2 both (see main text for details). One noise variable is of a Lévy-type (blue and red curves with plus sing + and dot • markers, respectively), and the other one is of a matching (see main text for details) Pareto-type (magenta and green curves with circle ○ and dot • markers, respectively). The calculations were done both by a direct numerical method (Sec. II C 1, red and green curves, $N = 10^6$, $B = 2 \cdot 10^4$) and by a semianalytical method (Sec. II C 2 blue and magenta curves). A straight line of slope 1/2 is included for reference.

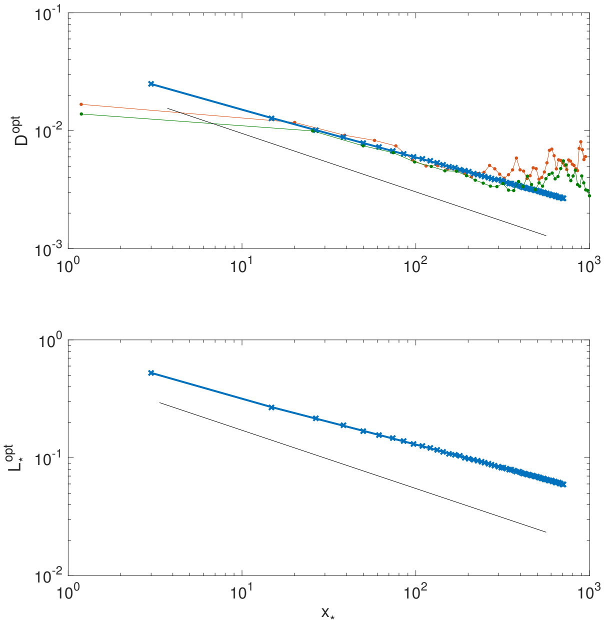

FIG. 4. Predictability of extremes in a P1-type process (Sec. II A) using two different noise variables of equal variance. One noise variable is a Gaussian one (blue and red curves with cross × and dot • markers, respectively), and the other one is uniformly distributed (green curve with dot • markers). The calculations were done both by a direct numerical method (Sec. II C 1, red and green curves, $N = 10^6$, $B = 2 \cdot 10^4$, $\Delta t = 10^{-2}$) and by a semianalytical method (Sec. II C 2, blue curve). Straight lines of slope 1/2 are included in both panels for reference.

vanishing according to power-laws. The power exponent in Figs. 3 is measured to be about 1/2 and in Fig. 4 somewhat less, which suggests a connection with the stability parameter $\alpha = 1/2$, possibly an exact one in the former case. Because of the good quality of the power-laws for the semianalytical results, we believe that the results for $D^{opt}(x_*)$ in the examined range of x_* are accurate, even if this can break down for larger x_*'s due to finite machine precision. In case of the P2-type process, a power-law asymptotic behaviour for $D^{opt}(x_*)$ cannot be established, because we do not know the presumably nontrivial asymptotic limit $\lim\limits_{x_*\to\infty} D^{opt}(x_*)$.

In Fig. 3 we plotted results using – beside a stable rv – a Pareto rv too. The PDF of it can be given by a (classical) Pareto distribution that is shifted 'horizontally'. This can be put in terms of the Generalized Pareto Distribution which, beside a shape parameter k and scale parameter σ, sports a location parameter μ too: $p_\xi = p_{GP}(\xi; k = 1/\alpha, \sigma = \xi_m/\alpha, \mu)$, where we used

$$\xi_m = \gamma_x \sqrt[\alpha]{(1-\varphi^\alpha)(1+\beta)\sin(\pi\alpha/2)\Gamma(\alpha+1)/(\pi\alpha)}. \tag{31}$$

Furthermore, in order to approximate the above described Lévy process distribution most closely, we have found empirically that the location parameter should be $\mu \approx 0.52$. With this approximation the semianalytical $D^{opt}(x_*)$ curve in Fig. 3 seems indistinguishable from that for the exact model. This is to indicate that $D^{opt}(x_*)$ could in principle be obtained analytically, with a very good approximation at least. However, our attempt to do this have not been fruitful. As we referred to this in Sec. II C 2, in this particular example the analytical treatment stumbled at the last hurdle, trying to solve eq. (27).

Upon this experience our idea was to use, instead of a continuous precursory variable, a discrete one. With this the ROC curve is not continuous but discrete, staircase-like, and so no equation like (27) arises. For convenience, we took the extreme case of a discrete precursory variable: a binary one. Specifically:

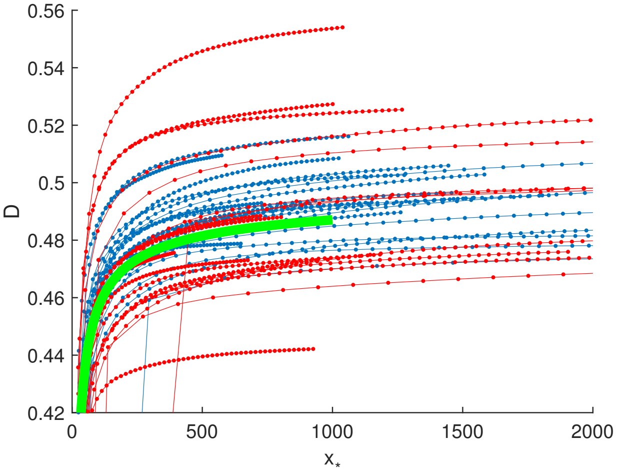

FIG. 5. Predictability of extremes in a P1-type process (Sec. II A) using a uniformly distributed noise variable and a binary precursory variable (see main text for details). The calculations were done both by a direct numerical method (Sec. II C 1, red and blue curves with dot • markers, $N = 10^6$, $B = 2 \cdot 10^4$, $\Delta t = 10^{-2}$) and by a fully analytical method (Sec. II C 2, thick solid green curve). The direct numerical results have been repeated 2×20 times. For the red curves the median defining the binary precursor (32) is estimated from data, and for the blue curves the true value of it is used. Discrete data points are connected by straight lines

$$\mathbf{x}_n = \begin{cases} 2, & x_{n-1} > x_{50\%} \\ 1, & x_{n-1} < x_{50\%} \end{cases} \qquad (32)$$

where $x_{50\%}$ denotes the median of the process PDF. The Appendix details how the distance D is calculated analytically for a binary precursor. We were able to complete this calculation for the P1-type process specified in this subsection above. The resulting expressions are, however, so complicated that it does not seem to be productive to write them out here; we have not gained an insight by them. Nevertheless, it is worthwhile to see the graph of $D(x_*)$ which we display in Fig. 5. Its tendency is the opposite of that with the continuous precursor. The lesson is thus that the much simpler precursor performs much poorer, with even the qualitative behaviour altered. Beside a lack of insight, the analytic treatment fails to have an added value in this regard too. We note that the analytic result is approximate, because we replaced the Gaussian rv by a uniform rv of unit variance. However, the approximation should be good, as is indicated by the closely matching red and green curves in Fig. 4.

We would like to emphasize that while just above for the P1-type process the predictability was deteriorating for increasing thresholds because of the 'weakness' of the precursor, further above for the P2-type process the deteriorating predictability (Fig 3) is down to a more fundamental reason, yet unknown. Also, ours is the first report on a process for which deteriorating predictability is found.

B. Predicting threshold nonexceedances

When we predict threshold *non*exceedances, by raising the threshold x_* we have more and more events. One could expect that the predictability is improving, or, changes in the opposite direction to that when predicting threshold exceedances. We show here that it is not the case: $D^{opt}(x_*)$ is the exact same for both exceedances and nonexceedances. Therefore, the monotonicity of $D^{opt}(x_*)$ is not due the monotonicity of the process CDF, as 'hypothesized' above under (ii).

The table of realised and predicted nonexceedance probabilities is given in Table III. It derives from Table II observing that $\chi^{\downarrow} = 1 - \chi^{\uparrow}$ and $\hat{\chi}^{\downarrow} = 1 - \hat{\chi}^{\uparrow}$. The downward directed arrow in e.g. $\chi^{\downarrow}$ expresses that we have an event when the data point is below the threshold. Also, what was e.g. χ in Table II or definition (6) we denoted here more distinctively as $\chi^{\uparrow}$. From the truth table we have the hit and false alarm rates, similarly to eqs. (9) and (10), as:

$$H^{\downarrow} = \frac{\mathcal{D}}{\mathcal{D}+\mathcal{B}} = 1 - F^{\uparrow}, \qquad (33)$$

$$F^{\downarrow} = \frac{\mathcal{C}}{\mathcal{A}+\mathcal{C}} = 1 - H^{\uparrow} \qquad (34)$$

It is more convenient to consider the 'miss rate' $M = 1 - H$ instead of the hit rate H. With that we have

$$M^{\downarrow} = F^{\uparrow}, \qquad (35)$$

$$F^{\downarrow} = M^{\uparrow}. \qquad (36)$$

And the definition of the distance from perfect predictability $(F, M) = (0, 0)$ also takes a more simple form:

$$D = \sqrt{F^2 + M^2}. \qquad (37)$$

That is, the miss and false alarm rates both are to be minimized. This reveals its similarity to the Brier skill score [24]. But the main point is that since in eq. (37) F and M play 'symmetric roles' in determining D, swapping these quantities when we predict nonexceedances instead of exceedances, as expressed by eqs. (35) and (36), leaves D unchanged.

IV. PROPERTIES OF THE ADDITIVE NOISE DISTRIBUTION

A. The influence of the power exponent

Our interest here is the dependence of predictability on the power exponent characterizing the tail of the pro-

TABLE III. Realised and predicted threshold nonexceedance event probabilities.

		$\chi^{\downarrow}$	
		1	0
$\hat{\chi}^{\downarrow}$	1	$\mathcal{D}$	$\mathcal{C}$
	0	$\mathcal{B}$	$\mathcal{A}$

cess PDF. It is given by the stability parameter α (more precisely: $\alpha + 1$) when we use a stable rv in a P2-type process. Specifically, we consider unskewed stable distributions $p_\xi(\xi) = p_s(\xi; \alpha, \beta = 0, \gamma = 1, \delta = 0)$ and $\varphi = 1/2$. In our parameter study we use the range of values $0.7 + 0.1k$, $k = 0, 1, \ldots, 13$ for α. We carry out computations following our semianalytical approach (Sec. II C 2). For smaller values of α than 0.7 the procedure tends to break down. The largest value is 2 yielding a Gaussian distribution. For smaller values the peaked middle part of the PDF are also Gaussian-like, but only the tail follows a power-law. The crossover between these two regimes shifts to the right for values closer and closer to 2. We display the graphs of $D^{opt}(x_*)$, one for each sample value of α, in Fig. 6 (a). Our main observations are the following. First, all curves approximate their respective horizontal asymptotes, situated at some presumably nontrivial levels, from below. This is similar to the result shown in Fig. 3, presumably for the same (yet unknown) reason. Second, the asymptotic predictability (the elevation of the asymptote) is the better, the smaller α is. Third, For some critical value of α a nonmonoticity of $D^{opt}(x_*)$ develops. It only vanishes for $\alpha = 2$ when the power-law tail vanishes shifting infinitely to the right. For that value we recover the results reported in [12] for the common Gaussian AR(1).

We also display in Fig. 6 (b) the optimal choice of the threshold $\mathcal{L}_*^{opt}(x_*)$ on the likelihood. Since for each parameter value it vanishes, it reveals the power-laws that govern the respective tails. The slopes in a log-log diagram show that the power exponent is inherited from the noise rv, although it is not $\alpha + 1$ just α. This needs further research which is beyond the scope of this paper. In any case, this fact makes us believe that the semianalytical calculations result in accurate D^{opt} estimates too. One might guess that $D^{opt}(x_*)$ decays to its limit value governed by a power-law with the same exponent. On this ground one might try to fit such a functional form to the data and estimate the elevation of the asymptote. We have not done this exercise, but we suspect that the accuracy of such estimates depends on α.

We note that $D^{opt}(x_*)$ are even functions and $\mathcal{L}_*^{opt}(x_*) - 1/2$ are odd functions (but these are not visible in the log-log diagram). We can explain the 'symmetry' of $D^{opt}(x_*)$ by pointing out, on one hand, that D^{opt} is the exact same for threshold exceedances and nonexceedances (Sec. III B). On the other hand, the P2-type process equation is satisfied by x_n, x_{n-1}, ξ_{n-1} upon reversing their signs, and $p_\xi(\xi)$ is symmetric.

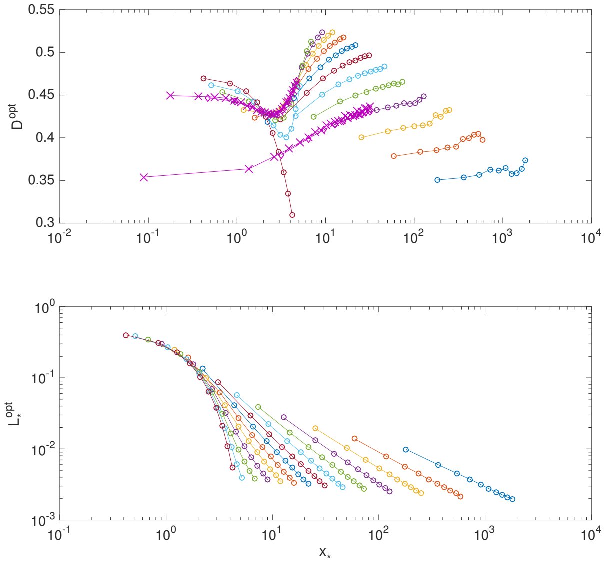

FIG. 6. Predictability of extremes in a P2-type process (Sec. II A) using stable noise variables (see main text for details). Sample values for α are $0.7 + 0.1k$, $k = 0, \ldots, 13$. The matching curves of $D^{opt}(x_*)$ and $\mathcal{L}_*^{opt}(x_*)$ are color coded with matching colors. The $\mathcal{L}_*^{opt}(x_*)$ curves follow an order from right to left for increasing values of α. Purple $D^{opt}(x_*)$ curves with circle ○ markers are extended by evaluating D^{opt} using the same seminalytical technique (Sec. II C 2) for more sample values of x_* (diamond ◊ markers in magenta); and D^{opt} is evaluated also based on long time series data (Sec. II C 1, $N = 10^6$, $B = 2 \cdot 10^3$) for a series of sample values of x_* (cross × markers in magenta). The close match of results by the two independent methods gives confidence in these results.

To reveal the symmetry of $D^{opt}(x_*)$ with direct numerical computation, B needs to be carefully chosen (results not shown), something that we do not yet fully understand.

B. What if the variance exists?

We do now a similar exercise as in Sec. IV A but instead of unskewed stable noise variables we use rv's obeying symmetrized Pareto distributions of the form:

$$p_{sP}(\xi; \alpha, \tilde{\xi}_m) = \frac{\alpha \tilde{\xi}_m^{\alpha}}{(|\xi| + \tilde{\xi}_m)^{\alpha+1}}. \tag{38}$$

As the expression tells, this distribution is created by 'mirroring' a (classical) Pareto distribution to $\xi = 0$ and 'pulling' both sides to $\xi = 0$ thereby 'making them meet'. It is straightforward to show that the asymptotic behaviour of the right tail of the symmetrized Pareto distribution can be described by a Pareto distribution whose scale parameter (as appearing in eq. (5)) is:

$$\xi_m = \tilde{\xi}_m / \sqrt[\alpha]{2}. \tag{39}$$

For a series of sample values of α, fixing $\tilde{\xi}_m = 1$, we calculated by our direct numerical approach the curves $D^{opt}(x_*)$. Beyond those considered in Sec. IV A, we include values up to 3, for which the variances of the noise and so the process PDFs exist. The results are displayed in Fig. 7. The important difference as compared with a stable noise variable is that no nonmonotonicity develops, and $D^{opt}(x_*)$ is increasing for all values of α examined. This includes values for which the variance of the distribution exists. Therefore, we can conclude that the existence of the variance or the lack of it is not decisive wrt. the monotonicity of $D^{opt}(x_*)$, namely, whether predictability improves or not, respectively, with the event magnitude. The difference between the results using stable and symmetrized Pareto rv's for $\alpha = 2$ is particularly striking, which should have to do with the fact that with a symmetrized Pareto noise the power-law tail does not 'get infinitely marginalized'. This is an indication that for deteriorating predictability what matters only is the power-law tail of the additive noise.

For two sample values, $\alpha = 1$ (circle $\circ$ marker) and 1.7 (diamond $\Diamond$ marker), similar results have been generated applying our semianalytical approach. The process distribution was approximated by a stable distribution based on the GCLT. The scale parameter of the stable distribution was obtained by eq. (5) in which for ξ_m we put the value obtained from eq. (39). While we see a convergence for the smaller value of α, which already gives a confidence in the results, the convergence is not 'in sight' for the larger value. This is clearly because in the latter case the fact that $\varphi < 1$ significantly is felt more, and so the GCLT implies for a symmetrized Pareto noise variable a poorer approximation of the process distribution and hence D^{opt} at the same x_*. Also, we notice that the curve in question in Fig. 7 is monotonic while the corresponding one in Fig. 6 was not. We speculate that it might be because the power-law scaling of the noise distribution (that determines the likelihood function via eq. (24)) prevails for all values of x, not just asymptotically like in case of the stable distribution.

V. NOISE DISTRIBUTION NOT DECISIVE

Let us consider a P1-type counterpart of the P2-type process from Sec. IV B driven by a symmetrized Pareto noise of tail index $\alpha = 3$. It is specified by $a = -1$, $b = 0$, $c = 1$, $d = 1$ in (2), yielding the process distribution:

$$p(x) = N_0 \frac{2e^{-\frac{2}{1+x}}}{(1+x)^4}. \tag{40}$$

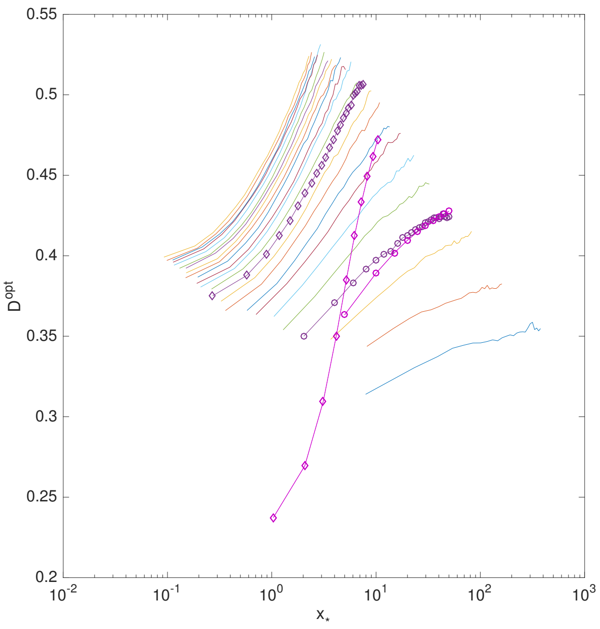

FIG. 7. Predictability of extremes in a P2-type process (Sec. II A) using symmetrized Pareto noise variables (see main text for details). The calculations have been carried out by a direct numerical approach (Sec. II C 1, $N = 10^6$, $B = 2 \cdot 10^3$). Sample values for α are $0.7 + 0.1k$, $k = 0, \ldots, 23$. The curves follow an order from right to left for increasing values of α. For two sample values, 1 (circle $\circ$ marker) and 1.7 (diamond $\Diamond$ marker), similar results are generated applying our semianalytical approach (Sec. II C 2).

We calculated the predictability for this both by the semianalytical and direct numerical approaches, whose results are displayed in Fig. 8. They match very well; and show an improving predictability with event magnitude. This suggests that this behaviour might be robust against variations of the power exponent. It may be down only to the CAM nature of the noise. However, we can ask the question whether this behaviour 'survives' the replacement of the Gaussian noise by a fat-tailed stable noise. This is tested by our last exercise.

We used stable noise variables of a series of stability parameter values, such as: $\alpha = 1 + 0.1k$, $k = 0, \ldots, 10$, and also 1.95. The calculations are based on time series data of equal size of $N = 10^6$ data points for each sample values. The reason for this is that we do not solve the fFP eqation, and without that we do not know the process distribution. Hence, we cannot employ our semianalytical technique. Each noise time series has a unit *observed* variance. The results are displayed in Fig. 9. We can see that the decaying nature of $D^{opt}(x_*)$ survives. At the highest thresholds examined we believe that these are only statistical errors that mask the monotonic decay. We back this up by examining, in two cases, two

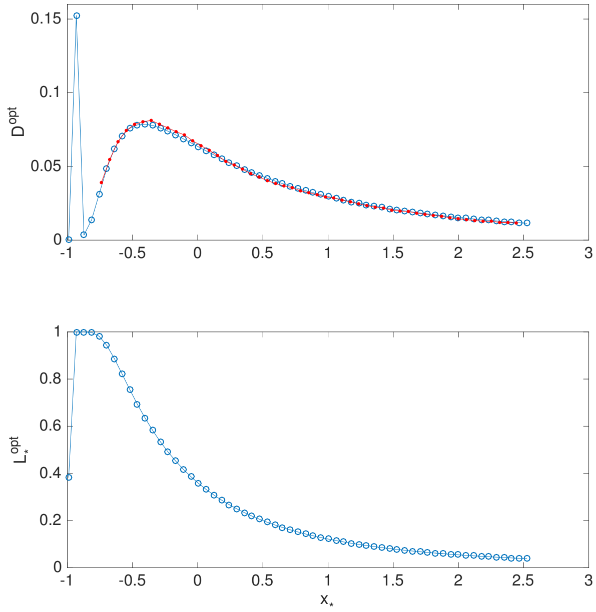

FIG. 8. Predictability of extremes in a P1-type process (Sec. II A) using a Gaussian noise variable (see main text for details). Calculations are done both by the semianalytical (Sec. II C 2, blue line with circle ○ markers) and direct numerical approaches (Sec. II C 1, red line with dot • markers, $N = 10^6$, $B = 2 \cdot 10^4$, $\Delta t = 10^{-2}$). Outliers in both diagrams indicate that in that regime the algorithm has difficulties dealing with fast-varying functions.

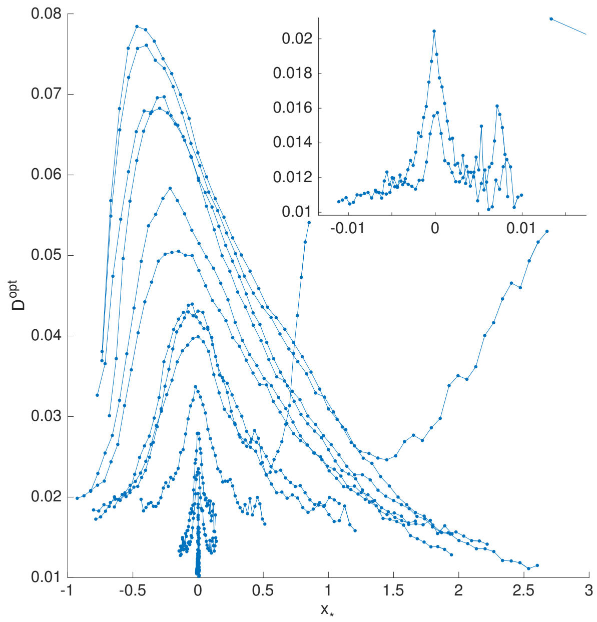

FIG. 9. Predictability of extremes in the same P1-type process as in Fig. 8 but using stable noise variables (see main text for details). Calculations are done only by the direct numerical approach (Sec. II C 1, $N = 10^6$, $\Delta t = 10^{-2}$). Sample values of the stability parameter values used are: $\alpha = 1 + 0.1k$, $k = 0, \ldots, 10$, and also 1.95. These values of α positively correlate with the peak height of the corresponding curves. We chose $B = 2 \cdot 10^4$, but for $\alpha = 1.6$ and 2 we also used $B = 2 \cdot 10^3$. With a time series of length $N = 10^6$ these give respectively $N/B = 50$ and 500 number of data points in each bin. A blowup is needed to show features of some curves.

different choices of the number B of bins (N unchanged) each case. With smaller B, which should result in larger statistical errors than the larger B chosen (see the figure caption), the tail of $D^{opt}(x_*)$ 'picks up' erroneously sooner than with the larger B chosen.

VI. CONCLUSIONS

We conducted a set of case studies that disprove categorically a number of intuitive propositions on the predictability of threshold exceedances in a linear stochastic differential equation (SDE). Neither any *single* property of the noise distribution considered nor any single property of the process distribution considered decides alone whether the predictability is improving or deteriorating with the event magnitude. But we conjecture that it is determined by the noise only, as presented visually in Fig. 1: if the noise is additive, then predictability is down to the noise distribution having a fat tail or not; and noise presented in a correlated additive-multiplicative manner to the SDE [34] determines *alone* the predictability to be improving.

The authors of [10] have a similar conclusion regarding the dominance of this property, but their conclusion is opposing to ours in that they find the predictability to be always worse with CAM noise. There is no contradiction, however, as these authors used a different measure of predictability, an anomaly correlation. It should be noted that these authors did not concern the predictability of extremes only. A similar situation arose with the verdicts of [9, 15] on the predictability of extremes. In [14] the two different measures of predictability used in [9, 15] were evaluated for the *same* model and observable demonstrating that the use of different measures can lead to different conclusions. We can conclude, therefore, that predictability has to be evaluated either comprehensively in terms of many different measures of prediction skill, or, some thought should be given as to what measures should not be used or rather preferred concerning extremes.

Finally, we would like to point out that Table I collecting studies on the predictability of extremes has an empty slot. An asymptotic theory for extremes in *deterministic* dynamical systems excludes the possibility of fat tails concerning so-called physical observables; as the chaotic attractor of a *dissipative* dynamical system is always compact, the extreme value distribution is always bounded, therefore, it is of the Weibull-type [4, 35]. Nevertheless, even deterministic dynamical systems can exhibit distributions of observables governed *approximately* by a power-law over *finite* ranges. Models of atmospheric dynamics provide examples. The finite range power-law scaling is faithfully reproduced by systematically derived reduced stochastic models, thanks to nonlinearities in their deterministic part [20] – unlike the linear eq. (1) which exhibits an asymptotic fat tail as expressed by (2). It remains to be seen what is the predictability of extremes in such models, whether it is like our finding for the CAM noise (Fig. 1), or the nonlinearities can bring about qualitatively different behaviour. Also, does the predictability of these models faithfully represent that of the real atmosphere? Such a study appears relatively straightforward owing to the possibility of employing the data-driven approach of evaluating predictability (Sec. II C 1). And finally, it would be desirable, possibly even practical, to underpin the emergence of fat-tailed Fréchet-type extremes and their predictability by a theory based on dynamical systems theory, possibly relating parameters of extreme value distributions, such as a positive shape parameter, as well as measures of predictability with dynamical characteristic numbers. This can be envisaged as a nonasymptotic extension of the currently existing theory [4, 35].

ACKNOWLEDGMENTS

TB was supported financially from the EU ERC grant for project NAMASTE (under grant no. 257106) owned by Valerio Lucarini, and the EU Horizon 2020 grant for project CRESCENDO (under grant no. 641816). CF was supported financially by the German Science Foundation (DFG) through the Cluster of Excellence 'CliSAP (EXC177) and the Collaborative Research Centre TRR 181 'Energy Transfer in Atmosphere and Ocean' and DFG grant FR3515/3-1.

APPENDIX: PREDICTION SKILL WITH A DISCRETE PRECURSOR

Let us consider a discretization of the precursory variable $\mathbf{x}$ by the binning of its space such that each bin covers the same probability of occurrence. This can be expressed for a *scalar* precursor x as:

$$x_b = b, \text{ when } x_{100\frac{b-1}{B}\%} < x < x_{100\frac{b}{B}\%}, \qquad (41)$$

$b = 1, \dots, B$, in terms of regular quantiles satisfying $F(x_{100\frac{b}{B}\%}) = b/B$. (The assignment $x_b = b$ is clearly just symbolic.) This is the same how we proposed to evaluate the prediction skill from time series data in Sec. II C 1. For each bin we can calculate the average likelihood:

$$\mathcal{L}_b(x_*) = \int_{x_{100\frac{b-1}{B}\%}}^{x_{100\frac{b}{B}\%}} d\mathbf{x}\mathcal{L}(x, x_*)p(x)/B. \qquad (42)$$

The posterior probabilities are then $\mathcal{P}_b(x_*) = \mathcal{L}_b(x_*)/B$. With these we can make use of the discrete formulations (18) and (19) and express the hit and false alarm rates as:

$$H_b = \frac{\langle \mathcal{L}_{b'}\tilde{\mathcal{H}}(\mathcal{L}_{b'} - \mathcal{L}_b)\rangle_{b'}}{\langle \mathcal{L}_{b'}\rangle_{b'}}, \qquad (43)$$

$$F_b = \frac{\langle (1 - \mathcal{L}_{b'})\tilde{\mathcal{H}}(\mathcal{L}_{b'} - \mathcal{L}_b)\rangle_{b'}}{1 - \langle \mathcal{L}_{b'}\rangle_{b'}}, \qquad (44)$$

where $\langle \cdot \rangle_{b'}$ denotes averaging wrt. index $b' = 1, \dots, B$. In the generic situation when the p_b's are not necessarily equal, the formulation is still straightforward via averaging the likelihood in bins and applying (18) and (19). Find an alternative formulation in [13].

Considering a binary precursor, $\mathcal{L}_2 > \mathcal{L}_1$ when $x_* > x_{50\%}$, and we can write

$$H_2 = \frac{\mathcal{L}_2}{\mathcal{L}_1 + \mathcal{L}_2}, \qquad (45)$$

$$F_2 = \frac{1 - \mathcal{L}_2}{2 - (\mathcal{L}_1 + \mathcal{L}_2)}, \qquad (46)$$

which is the only nontrivial 'step' of the ROC 'staircase'. Hence, no nontrivial optimization, i.e., selection of the maximal value, is needed to determine the prediction skill in terms of D.

[1] C. Franzke, Chaos (2017).

[2] The expression is a shorthand for the preceding definition. In our treatment we are concerned with situations when the *asymptotic* behaviour follows a power-law, and we disregard the fact, like in [36], that in physical systems the scaling range is limited by, say, the finite system size [37] or other nonlinearities [20].

[3] S. Coles, *An Introduction to Statistical Modeling of Extreme Values* (Springer, 2001).

[4] T. Bódai, in *Nonlinear and Stochastic Climate Dynam-*

ics, edited by C. Franzke and T. O'Kane (Cambridge University Press, Cambridge, 2017).

[5] Y. Malevergne and D. Sornette, *Extreme financial risks: From dependence to risk management* (Springer Science & Business Media, 2006).

[6] P. D. Sardeshmukh and P. Sura, Journal of Climate **22**, 1193 (2009).

[7] P. Sura, Atmospheric Research **101**, 1 (2011).

[8] H. L. D. de S. Cavalcante, M. Oriá, D. Sornette, E. Ott, and D. J. Gauthier, Phys. Rev. Lett. **111**, 198701 (2013).

[9] A. E. Sterk, M. P. Holland, P. Rabassa, H. W. Broer, and R. Vitolo, Nonlinear Processes in Geophysics **19**, 529 (2012).

[10] P. Sura, M. Newman, C. Penland, and P. Sardeshmukh, Journal of the Atmospheric Sciences **62**, 1391 (2005).

[11] A. E. Sterk, D. B. Stephenson, M. P. Holland, and K. R. Mylne, Quarterly Journal of the Royal Meteorological Society **142**, 58 (2016).

[12] S. Hallerberg and H. Kantz, Nonlinear Processes in Geophysics **15**, 321 (2008).

[13] J. M. Miotto and E. G. Altmann, PLoS ONE **9**, e111506 (2014).

[14] T. Bódai, Physica D: Nonlinear Phenomena **313**, 37 (2015).

[15] C. Franzke, Phys. Rev. E **85**, 031134 (2012).

[16] S. Bialonski, G. Ansmann, and H. Kantz, Phys. Rev. E **92**, 042910 (2015).

[17] J. P. Nolan, *Stable Distributions – Models for Heavy Tailed Data* (Birkhauser, Boston, 2015) in progress, Chapter 1 online at academic.2.american.edu/ jpnolan.

[18] "The Fokker-Planck equation associated with the SDE in the Itô form $dx = h_I(x,t)dt + g(x,t)dW^*$ is $\partial_t\rho(x,t) = -\partial_x[D^{(1)}(x,t)\rho(x,t)] + \partial_{xx}[D^{(2)}(x,t)\rho(x,t)]$, where $D^{(1)}(x,t) = h_I(x,t)$, $D^{(2)}(x,t) = g^2(x,t)$, and in our case $h_I(x) = ax + b$, $g(x) = (cx + d)/\sqrt{2}$, because $\sqrt{2}dW = dW^*$.".

[19] H. Risken, *The Fokker-Planck equation* (Springer, 1996).

[20] A. J. Majda, C. Franzke, and D. Crommelin, Proceedings of the National Academy of Sciences **106**, 3649 (2009), http://www.pnas.org/content/106/10/3649.full.pdf+html.

[21] D. Schertzer, M. Larchevque, J. Duan, V. V. Yanovsky, and S. Lovejoy, Journal of Mathematical Physics **42**, 200 (2001).

[22] The limit function of the Generalized extreme value distribution [3] as its shape parameter $\xi \to 0$ can be shown in a similar way to be the Gumbel extreme value distribution, using the following definition of the exponential function: $\lim_{n\to\infty}(1 + x/n)^n = \exp(x)$. In our case, furthermore, we have to bear in mind that the normalization factor N_0 can feature appropriately vanishing factors that 'balance' diverging factors of the rest of $p(x)$. A nondiverging factor of the numerator and one that cancels a factor of the denominator can be found both as first-order terms of respective Taylor expansions.

[23] P. E. Kloeden and E. Platen, *Numerical Solution of Stochastic Differential Equations* (Springer, 1995).

[24] S. Hallerberg, J. Bröcker, and H. Kantz, in *Nonlinear Time Series Analysis in the Geosciences*, Vol. 112, edited by R. V. Donner and S. M. Barbosa (Springer, 2008).

[25] The probability density $p_{\mathbf{x}|\chi}(\mathbf{x}, \chi = 1)$ of $\mathbf{x}$ conditioned on some realised value of χ is usually denoted more simply as $p(\mathbf{x}|\chi = 1)$, but we want to emphasize that we consider a function of two variables. Also, it would create ambiguity if the symbol p without a subscript was to be reused to denote another function, and we prefer to reserve p for the process PDF.

[26] S. Hallerberg and H. Kantz, Phys. Rev. E **77**, 011108 (2008).

[27] C. Marzban, Weather and Forecasting **13**, 753 (1998).

[28] C. A. T. Ferro, Weather and Forecasting **22**, 1089 (2007).

[29] And for that reason $\mathcal{A} + \mathcal{B} + \mathcal{C} + \mathcal{D} = 1$.

[30] J. P. Egan, *Signal Detection Theory and ROC Analysis* (Academic Press, 1975).

[31] By 'conservative estimation of the skill' we mean that an overestimation of skill (e.g. D^{opt} estimated to be smaller than the true value) is excessively unlikely. This entails an appropriate sign of the bias of the estimator, and a standard deviation of the estimator much smaller than the modulus of the bias.

[32] Proposition (ii) is disproved already by Fig. 6 of [14] as the monotonicity of $D^{opt}(x_*)$ is disrupted at small scales at locations where the monotonicity of $p(x)$ is disrupted, which clearly cannot disrupt the monotonicity of the CDF. This observation could give rise to proposition (i).

[33] As the definition of the likelihood function given by eqs. (22) and (24) suggests, two factors can be at play: the dynamics and the noise distribution. In Sec. V it is indicated that the later does not make a difference.

[34] It is called a 'CAM noise' in the literature [6, 10, 20], although one can argue that the multiplicative nature of the noise is a deterministic effect.

[35] V. Lucarini, D. Faranda, A. C. G. M. M. de Freitas, J. M. M. de Freitas, M. Holland, T. Kuna, M. Nicol, M. Todd, and S. Vaienti, *Extremes and Recurrence in Dynamical Systems* (Wiley, 2016).

[36] G. A. Gottwald and I. Melbourne, Proceedings of the National Academy of Sciences **110**, 8411 (2013).

[37] A. Garber and H. Kantz, The European Physical Journal B **67**, 437 (2009).