Predictability of threshold exceedances in dynamical systems

Tamás Bódai

Meteorological Institute, University of Hamburg, Hamburg, Germany

Abstract

In a low-order model of the general circulation of the atmosphere we examine the predictability of threshold exceedance events. The likelihood of such binary events is established from long time series of one or more observables of the same system. The prediction skill is measured by a summary index of the ROC curve that relates the hit- and false alarm rates. Our results for the examined system confirm a counterintuitive (and seemingly contrafactual) statement – provided that the bin size for binning time series data is optimized, but not necessarily otherwise – previously formulated for more simple autoregressive stochastic processes, namely, that exceedances of higher thresholds are more predictable; or in other words: rare extremes are more predictable than frequent typical events. We argue that when there is a sufficient amount of data depending on the precision of observation, the skill of data-driven prediction approximates the skill of model-driven prediction, assuming strictly no model errors, and therefore stronger extremes are more predictable also in the latter situation. Furthermore, we show that a quantity commonly regarded as a measure of predictability, the finite-time maximal Lyapunov exponent, does not correspond directly to the ROC-based skill score when they are viewed as functions of the prediction lead time and the threshold level. This points to the fact that even if the Lyapunov exponent as an intrinsic property of the system, measuring the instability of trajectories, determines predictability, it does that in a nontrivial manner.

Keywords: extreme event, peak-over-threshold event, data-driven prediction, precursory structure, prediction skill, ROC curve, Finite-time Lyapunov exponent

1. Introduction

Extreme events have fundamental importance to life, as they are often associated with survival and losses. – Extreme events to do with gains or amusement receive far less attention in general, dissociated from individual events. Rare and large magnitude events of interest arise in physical, technological, social, and other systems. The classical theory of extremes in uncorrelated sequences (or, in sequences in which the auto-correlation is decaying sufficiently fast) [1, 2] has a statistical orientation; it is not- and cannot be concerned with prediction or with uncovering mechanisms that can produce extremes; but it is rather concerned with e.g. expected return times, which can be useful in designing structures of a certain required life time, such as e.g. sea walls [3].

Since Newton revolutionized science, it has become a paradigm that predictions should be based on validated models. These models
describing fluctuating phenomena often take the form of a system of differential equations, also referred to as a dynamical system. Since the work of Lorenz it has become clear that even though some phenomena can be modeled quite accurately, they can be inherently unpredictable because of their extreme sensitivity to initial conditions [4]. Such systems are called chaotic, characterized by positive Lyapunov exponents. This imposes a time horizon on predictions; beyond that only statistical properties can be robustly estimated.

Model-based or model-driven point-wise predictions (MDP), like e.g. a weather forecast, do not distinguish between extremes and typical events – it is not an iterated procedure. However, one can often read that extremes are much harder to predict in this framework. Unfortunately a systematic study of the dependence of some appropriate skill score of predictability of any model on the magnitude of events is still lacking. Inaccuracy of the model may be an important factor determining such a dependence, beside details of its chaotic nature. In contrast, in pure data-driven prediction (DDP), model errors are not present. Instead, beyond errors in measuring the present conditions (as with initial conditions for MDP), the prediction is compromised by the finite size of the data set. That is, the said virtue of DDP can be exploited – when employing it in its pure form – only if enough and good quality data (with a high precision of observation and high signal-to-noise ratio) is available [5].

DDP is gaining increasing prominence nowadays given that data is relatively much more easily accessible than models. This is certainly the case with geophysical phenomena that we are primarily interested in, such as e.g. meteorology. Furthermore, performing predictions based on data can be far less costly than those based on simulating complex models, while the skill may not be much worse [6]. With a view to exploring its full potential, here we advance the analysis of data-driven predictability of binary extreme threshold exceedance events in dynamical systems. Within the framework of our analysis results for MDP can also be obtained, with which the two prediction techniques can be contrasted against one another.

One might expect that the slogan that ‘extremes in comparison with more moderate events are harder to predict’ extends to DDP. In fact, just the opposite has been reported for simple autoregressive processes at least [7], indifferently to whether the probability distribution is exponentially decaying or according to a power-law, and also for some observational data [8]: stronger events are easier to predict. This counter-intuitive statement is based on a kind of skill score that derives from the so-called receiver operating characteristic (ROC) curve [9], which latter takes into account the true positives – meaning that an event is correctly predicted to happen – as well as the false negatives. Concerning rare events, such a skill score is regarded more meaningful than other proper [10] skill scores for probabilistic predictions like the Brier or Ignorance scores. The ROC statistics does not depend on the relative frequency of events, only that the accurate evaluation of the statistics requires a sufficient number of events. We choose to assess predictability in terms of a summary statistics of the ROC curve: the ‘distance’ from the ideal situation of successfully predicting all events without any false alarms. Whether the above statement [7] can be maintained in case of more complex processes has been an open question so far – addressed but not settled with a consensus. Our study is devoted to looking for an answer in case of processes generated by fully deterministic and stochastic versions of a chaotic dynamical system [11], and
provides evidence to the affirmative in those cases.

Recently two studies [12, 13] have been published concerning the predictability of extreme events in dynamical systems with seemingly contradictory results as to whether stronger events are more predictable. Franzke [12] conducted DDP of extreme threshold exceedances in a systematically derived stochastic dynamical system representing climate variability by the resolved (slow) variable(s) and weather variability by noise in place of the unresolved (fast) variable(s) [14]. He maintained the earlier statement [7] in this case, measuring the prediction skill by the ROC statistics, but on the basis of a more selective data set than that is presented in what follows here. On the other hand, Sterk et al. [13] considered a number of dynamical systems of various complexity, and various physical observables. They evaluated finite-time maximal Lyapunov exponents related to trajectories that lead to extremes, and concluded that no generally applicable statements can be made, but the predictability of extremes depends on the system (and so the attractor geometry) and on the observable in question, as well as the prediction lead time. We emphasize that in their study the authors did not take model errors into account. Here, following these authors, we aim to carry out both types of assessment with respect to the same observable of the same model in order to ascertain that it is solely the choice of the measure of prediction skill that can lead to different conclusions. The validity of the latter objective to be reached the way we outlined it above is based on a simple argument that data- and model-driven predictions, provided that no model errors are present, should have the same skill when measured by the same skill score in both cases.

We will examine the dependence of the skill on the prediction lead time in more detail, and find a return of skill at certain points – something that is attributed to low-order chaotic dynamics. We will relate this to the need of optimizing the bin size for binning data.

The structure of the paper is as follows. Next we recapitulate the methodology of the applied prediction scheme and the used ROC-based skill score. As an alternative of some sort we also define the finite-time maximal Lyapunov exponent, which is traditionally considered to be a measure of model-driven predictability, the inverse of which being proportional to a time horizon of prediction. The model of our choice, Lorenz’s 1984 model of global atmospheric circulation, simulated to produce time series data for the purpose of assessing the predictability of threshold exceedance events is also briefly described. Subsequently, in Sec. 3, we present our results on the dependence of predictability on several factors, such as: the type and number of observables, the prediction lead time, and the magnitude of extreme events. These results are summarized in a compact table format in Sec. 4 and discussed subsequently. To close our presentation we pose a few open questions for future research into practical aspects of the prediction of extremes, which might potentially have theoretical ramifications.

2. Methodology

2.1. Prediction of threshold exceedances by precursors

Our aim is to predict large excursions of some (scalar) physical observable $x$, exceeding a chosen threshold level $x_*$, before that exceedance happens. Figure 1 pictures the situation as the observable evolves continuously in time, $x = x(t), t \in \mathbb{R}$, occasionally exceeding the threshold. We intend to examine situations when $x(t)$ is generated by a process that can
be described by an ordinary or stochastic differential equation, examples of which for our case study will be briefly described in Sec. 2.3.

**Formal setting.** The following methodological description regarding the prediction task closely follows [9]. We introduce a discrete-time binary event variable:

$$\chi_n = \begin{cases} 1, & x(t_n) > x_s \\ 0, & x(t_n) < x_s \end{cases}$$

(1)

where the times $t_n$, $n \in \mathbb{Z}$, belong to consecutive apexes, that is, local maxima or peaks, of the continuous evolution of $x$ for consecutive values of $n$, and in general they are not equally ‘spaced’. Thereby the continuous-time evolution is discretized. This approach is often referred to as the peak-over-threshold (POT) approach. The prediction is based on a like-wise discrete-time precursory structure $x_n \in \mathbb{R}^M$ 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 $\chi_n$ at $t_{n-\min(d_m)}$ is defined as:

$$\hat{\chi}_n = \begin{cases} 1, & \mathcal{L}(x_n) > \mathcal{L}_s \\ 0, & \mathcal{L}(x_n) < \mathcal{L}_s \end{cases}$$

(2)

based on the likelihood function:

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

(3)

In the above $\mathcal{P}(x) = p_{x|x}(x|\chi = 1)$ is the posterior probability density functions (PDF) of $x$, and $p(x)$ is the ‘parent’ PDF, i.e., the basic PDF generated by the considered process. Note that Eq. (3) expresses Bayes’ theorem relating the conditional probabilities: the likelihood and the posterior probability. Our prediction $\hat{\chi}_n$ is controlled by a threshold $\mathcal{L}_s \in [\min(\mathcal{L}), \max(\mathcal{L})]$ of stringency on $\mathcal{L}$. Note that an actual choice is meant to be made as to the applied value of $\mathcal{L}_s$ in practice, for which reason this kind of prediction is not probabilistic.

Depending on $\mathcal{L}_s$, the rate of true positives, or the hit rate, i.e., the frequency of making...
correct predictions, yields as follows:

\[ H(L_*) = \frac{\int_{x \in \{x : L(x) > L_*\}} dx P(x)}{\int_{\forall x} dx P(x)}. \]  

(4)

Another measure of the overall goodness or skill of prediction is the false alarm rate:

\[ F(L_*) = \frac{\int_{x \in \{x : L(x) > L_*\}} dx [p(x) - P(x)]}{\int_{\forall x} dx [p(x) - P(x)]}. \]  

(5)

Clearly, one can achieve a very good hit rate by reducing the stringency, but in fact [9] always at the price of an increased false alarm rate. Figure 2 shows an example of how the two measures of skill depend on the stringency in terms of a parametric plot or curve \( \{(F(L_*), H(L_*))\} \), which is referred to as the receiver operating characteristic (ROC) curve. With the extremal choices, \( L_* = 0 \) 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, and situated above the diagonal with any skill. In the same diagram another ROC curve is also shown, to be referred to as \( P \)-ROC curve, obtained by replacing in the definitions (2), (4), (5) \( L \) by \( P \) (but not vice-versa!) and \( L_* \) by \( P_* \in [\min(P), \max(P)] \). This is based on the intuitive strategy, expressed by the conditional probability, that one looks at what happens before extreme events. From Eq. (3) one can see that following this strategy the posterior PDF is just the likelihood that such states lead to an extreme event weighted by the relative frequency of those states, whereby the ‘predictive potential of relatively infrequent states is suppressed’. It can be shown [9] that as a result of this the \( P \)-ROC curve will be always wholly underneath the \( L \)-ROC curve, making this intuitive strategy inferior. Furthermore, the \( L \)-ROC curve is always concave, while the \( P \)-ROC curve is not necessarily concave. Besides, in accordance with the above statement on the trade-off situation, \( F_L(L_*) \), \( H_L(L_*) \), \( F_P(P_*) \), \( H_P(P_*) \) are all monotonic functions, and, therefore, so are e.g. \( H_L(F_L) \) and \( H_P(F_P) \).

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 \( L_* \). Clearly this is possible only in case of the deterministic but not the stochastic version of a model, and there are further factors – to be demonstrated in Sec. 3 – that can compromise the prediction skill. In the nonideal situation an optimal \( L_* \) is to be chosen. 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:

\[
D = \min_{\mathcal{L}}(\sqrt{F^2 - (H - 1)^2}),
\]

the distance of the ROC curve from the ideal corner. With no prediction skill at all: \( D = \sqrt{2}/2 \). We note that it is not trivial to interpret what the comparison of \( D \) with a proper skill score of probabilistic prediction means.

**Numerical issues.** Perhaps the most obvious factor that compromises the prediction skill in the data-driven framework is the finite size \( N \) of the data set: \( \{x_n, x_n\}, \ n = 1, \ldots, N \). The distributions \( p(x) \), \( P(x) \), \( L(x) \) will be approximated in our study by histograms \( \{p_b\}, \{P_b\}, \{L_b\}, \ b = 1, \ldots, B \), of a certain uniform bin size \( \Delta x \in \mathbb{R}^M \); different values of \( b \) can be assigned to the different bins by a sensible algorithm. Let us denote by \( \Delta x \) the unique linear bin size applied in all \( M \) dimensions after suitable nondimensionalization. Note that in the coarse-grained situation \( \{L_b\} \) derives from \( \{p_b\} \) and \( \{P_b\} \) much the same way as with the continuous functions according to Eq. (3). With the discrete formulation of Eqs. (4) and (5), 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 \( \{\mathcal{L}_{b}\} = \{L_b\} \). Note that if \( \mathcal{L}_b \) does not-, then neither do \( H_b \) and \( F_b \) change monotonically with increasing \( b \).

The above estimation of the measures of skill is not conservative, however, which is to do with small histogram counts 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 discrete versions of Eqs. (4) and (5), 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 \( \{P_b\} \) are derived from the evaluation data set.

A further issue to do with small bin sizes when many bins 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 \( D > \sqrt{2}/2 \) can even be realized.

Too large bin sizes would of course also deteriorate the prediction skill. Therefore, there should be an optimal bin size yielding (locally) minimal \( D \). 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 defining the regular grid.

**Relationship of data- and model-driven predictability.** We note finally that the above description of evaluating predictability applies clearly to the case of data-driven prediction. However, evaluating the model-driven predictability of binary exceedance events measured by the same ROC-based skill score can be done in exactly the same way: by simulating long trajectories of the model and by binning the resulting time series data. This is true more obviously when MDP is thought of in the sense that the PDFs are established preliminary to making any predictions. In this case the ‘archival’ MDP and DDP differ only in that in principle unlimited data is available to establish the PDFs for MDP. But MDP can be thought of also in the sense that the model is only simulated ‘on-demand’, to produce an ensemble forecast from which the likelihood of a threshold exceedance can be established. This case corresponds to the situation when the components of the precursory structure \( x \) of the archival MDP or DDP belong to the
same time instant, representing the *initial conditions* for the on-demand MDP. Note that it is allowed that \( x \) excludes some variables that determine the considered phenomenon; they can be initialized arbitrarily, or possibly as a random sample from a probability distribution. We refer here to the practice of stochastic parametrization of unresolved processes in weather forecast models. As for the definition of the binary event in case of on-demand MDP the following can be taken: in an event a chosen observable exceeds a set threshold in a chosen future time interval, defined by a leading window of width \( \Delta T \) at a lead time \( T \) ahead of the present time. Let us call this an event of *threshold exceedance in an interval* (TEI event in short). In contrast to considering the threshold exceedance of apexes of the time evolution, i.e., POT events, here \( T \) can be set arbitrarily, not restricted to discrete values; accordingly, \( \chi(t) \) and \( x(t) \) are defined in continuous-time.

The bin size \( \Delta x \) would in this case represent the precision of measuring initial conditions. In order to well-approximate the skill of MDP for a given measurement precision, the time series has to be long enough so that the likelihood for all bins, even those which cover a relatively small portion of the invariant measure of the attractor, is well-approximated on the first place. That is, the ratio of the prediction skill of DDP and that of MDP (with no model errors), which is always smaller than unity, depends on \( N \Delta x \).

2.2. Finite-time Lyapunov exponent

A well-known measure of predictability is the positive maximal Lyapunov exponent (MLE), which approximates the average rate of the exponential separation of very close trajectories on a chaotic attractor [4]. Sterk et al. [13] evaluated the finite-time version of this measure to assess the predictability, with some lead time \( T \), of extreme events. We consider here ‘apriori’ known nonautonomous dynamical systems \( \dot{y} = f(y, t) \) in a \( d \)-dimensional phase space with generic initial condition \( y_0 = y(y_0, t = t_0, t_0) \in \mathbb{R}^d \), where \( y(\cdot, \cdot, \cdot) \) denotes the two-time evolution operator.

The spectrum of finite-time \( T \) Lyapunov exponents (FTLE) \( \lambda_i^{(T)} \) can be defined in a *pullback sense* [15] as follows:

\[
\lambda_i^{(T)} = \lim_{t_0 \to -\infty} \frac{\ln(s_i^{1/2}(t, t_0)) - \ln(s_i^{1/2}(t - T, t_0))}{T} \quad (7)
\]

\( i = 1, \ldots, d^* \leq d \), where \( s_i(t, t_0) : \det(Y(t, t_0) \cdot Y^T(t, t_0) - sI) = 0 \) are the singular values of the deformation gradient, \( Y = \partial y / \partial y_0 \), governed by the variational equation:

\[
\dot{Y} = \frac{\partial f}{\partial y} \bigg|_{y(y_0, t, t_0)} \cdot Y. \quad (8)
\]

The initial condition is not arbitrary but implied as \( Y(y_0, t_0, t_0) = I \). Note that the LEs are recovered as: \( \lambda_i = \lim_{T \to \infty} \lambda_i^{(T)} \). We will omit the index \( i \) to denote the MLE simply by \( \lambda \) or \( \lambda^{(T)} \). Clearly, by \( \lambda^{(T)}(t) \) the predictability from the present time \( t - T \) of a trajectory at the future time \( t \) is defined.

A summary statistics for this measure of predictability can be defined, generalizing the proposal of Sterk et al. [13], by an ensemble average over parts of the pullback or *snapshot attractor* [16] that realize extreme events in terms of some physical observable \( x(y) \):

\[
\langle \lambda^{(T)} \rangle = \int_{x(y(t)) > x^*} dy \lambda^{(T)}; \quad (9)
\]

or over parts of the snapshot attractor collecting the ensemble of trajectories that would cross the threshold in a leading window of time of width \( \Delta T \) at time \( t \) (TEI events):
The average FTLE \( \langle \lambda^{(T)} \rangle \) is dissimilar to \( D \) in that it is not calculated from predicted data, but rather it expresses an intrinsic property of the system that determines predictability. Nevertheless, we will compare figures obtained for \( \langle \lambda^{(T)} \rangle \) by (10) and \( D \), at least in case of the autonomous dynamics when the attractor is time-invariant.

### 2.3. The model climate

We carry out the assessment of the predictability of extremes in both ways as described in Secs. 2.1 and 2.2, respectively, in a model of geophysical relevance, which constitutes a nonlinear dynamical system featuring complex chaotic deterministic dynamics, yet, is simple enough – involving just three scalar prognostic variables – to yield data relatively inexpensively and to allow for a more tangible demonstration of some aspects of predictability. Our choice of a model, Lorenz’s model of global atmospheric circulation (L84) with standard parameter settings reads as follows [11, 17, 16]:

\[
\begin{align*}
\dot{x} &= -y^2 - z^2 - x/4 + F/4, \\
\dot{y} &= xy - 4xz - y + 1, \\
\dot{z} &= xz + 4xy - z.
\end{align*}
\]

(11)

The model describes – in a very coarse manner [18] – the meridional heat transport via eddies, represented by principal mode amplitudes \( y \) and \( z \), given rise by the baroclinic instability of the midlatitude jet, represented by its average speed \( x \). The instability occurs for appropriate conditions defined by the large scale meridional temperature gradient, represented in the model by \( F \), due to differential heating between the equator and the poles. The equations are nondimensionalized with respect to time by the average damping time of eddies, being about 5 days.

We will examine the autonomous dynamics in perpetual winter conditions realized by, say, \( F = 8 \), since this gives rise to chaotic dynamics, which is nontrivial from the point of view of predictability. We will consider also nonautonomous dynamics by introducing some driving or time-dependent forcing to the L84 system in the form of \( F(t) = F_0 + A \tilde{x}(t) \), where the fluctuating process (in a mathematical sense) \( \tilde{x}(t) \) can be seen as an unresolved, i.e., physically not modeled, (physical) process. However, from the point of view of data-driven predictability, as described in Sec. 2.1, autonomous and nonautonomous systems are not distinguished – the driving mimics additional degrees of freedom of the system. Assuming a time scale separation between the resolved and unresolved processes one can apply an uncorrelated white noise (WN) driving: \( \tilde{x}(t) = \xi, \int_{-\infty}^{t} dt \xi = W_t \), where \( W_t \) is a Wiener process. Or, we can represent additional degrees of freedom of comparable time scales to that of the resolved dynamics, \( \tau_{L84} \approx 4 \), by a continuous smooth chaotic process, such as e.g. the first component of the classical Lorenz equations (L63): \( \{ \dot{x} = \tau^{-1} \sigma (\tilde{y} - \tilde{x}), \dot{y} = \tau^{-1} (\rho \tilde{x} - \tilde{y} - \tilde{x} \tilde{z}), \dot{z} = \tau^{-1} (-\beta \tilde{z} + \tilde{x} \tilde{y}) \} \), in which we introduced a time scale parameter \( \tau \) so that, with the common choice for a chaotic solution: \( \tau = 10 \), \( \rho = 28, \beta = 8/3, 0.7 \tau / \tau_{L84} \) is of \( o(1) \). Note that \( \tau_{L63} = 0.7 \) is the time scale of L63 with \( \tau = 1 \) defined by the crossover frequency in its power spectrum. More specifically, we define a new time scale \( \tau' \) to be unity when \( \tau \) is such that the L63-driving exerts maximal response of the driven L84 in terms of extremal behavior, measured e.g. by either the kurtosis.
or a high quantile of the distribution of \( x \), and will consider the ‘resonant’ scenario of \( \tau' = 1 \). Based on our finding \[16\] \( \tau' = 0.4\tau \). Note that \( \tau' \approx 2.3 > 1 \) when \( \tau_{L63} = \tau_{L84} \). Beside the resonant scenario we will also consider one when the driving is much faster than the model climate: \( \tau' = 1/4 \). We use a coupling strength \( A = 0.025 \) in case of the L63-driving, and an appropriate choice of \( A \) in case of the WN-driving that gives the same variance of the driving.

3. Results

3.1. Makeup of the precursory structure: the type and number of observables

The most simple case of a discrete-time precursory structure is that of the previous peak value of the observable whose threshold exceedances are to be predicted: \( x_n = x_{n-1} \). As an example we take the first component of L84 as for an observable (denoted identically to our generic observable by \( x \)). The two panels of Fig. 3 show the posterior PDF and the likelihood function, respectively, represented by adequate histograms. The cases of the autonomous, i.e., undriven, and L63-driven L84 are shown in one diagram side-by-side. These histograms and other numerical results in this paper are constructed from sets of about \( 5 \times 10^6 \) discrete data points each, resulting from appropriately long simulations. We simulate the autonomous and L63-driven L84 using Matlab’s ode45, which integrator chooses the time step size \( h \) adaptively, and we employ the explicit order 1.5 strong scheme described in \[16\] to integrate the WN-driven L84 (results following below) with fixed \( h = 0.01 \). As expected, the driving smooths out features of both distributions, however, contrary to expectations: e.g. the likelihood (b) can be even enhanced by driving (see for example \( 1.2 < x_{n-1} < 1.4 \)). Furthermore we point out that the relationship given by Eq. (3) is manifested in the more broad structures of the distribution of the likelihood as compared to that of the posterior probability density. As already mentioned in Sec. 2.1, this broadening ought to be reflected in the relative positions of the respective ROC curves. In fact the pair of ROC curves in Fig. 2 belong to the present prediction scenario considering the undriven L84.

It is an intuitive expectation that the predictability can be improved by relying on more information by means of extending the precursory structure. The most simple step in this direction – in the realistic context when only one variable is available or practical to observe – is that beside the previous data point we monitor also the one just before that, i.e., \( x_n = (x_{n-1}, x_{n-2}) \). The distributions for the same scenarios as considered before in Fig. 3 are displayed in Fig. 4. The bivariate distributions are visualized by color plots; and beside the distributions, on the left we also display scatter plots of data points (but for a better visibility of features we plot fewer points than those that the histograms are constructed from). In the case of the autonomous L84 the scatter plot reveals a fractal pattern. This could lead one to think that there is a one-to-one or unique relationship between subsequent pairs of \( (x_{n-1}, x_{n-2}) \), which is also called a mapping or map \[4\]. This can be confirmed by looking at the distribution of the likelihood, which takes on the maximum value of unity wherever the scatter plot exhibits fractality. – Because of the uniqueness, we have a deterministic rule to predict the next value, and so we can tell with certainty whether it will exceed the threshold. In regions where a lack of clear fractality is observed, e.g. around \( (x_{n-1}, x_{n-2}) = (1.4, 1.4) \), \( \mathcal{L} < 1 \) consistently. The exhibited pattern of the scatter plot can give the intuition that the lack of uniqueness is a result of ‘looking at’ a curved surface liv-
ing in 3D ‘from a poor angle’ so that the 2D view of some parts of the surface is obstructed by other parts of it. In other words, the surface looks folded. In fact, Takens’ embedding theorem [19] states that an attractor of Hausdorff dimension $D_0$ can (always) be embedded by $M > 2D_0$ number of delay variables. For us this means an unfolded appearance. In our case $D_0 \approx 1.6$ [16], and so for uniqueness we need maximum $M = 4$. This does not ‘encourage’ us that we can have an unobstructed 2D view, although neither does it say that we cannot have. In fact, in our case we can have such a view, to be described next.

Let us bear in mind that the discrete $x_n$ data belong to apexes of the continuous $x(t)$. In these points $\dot{x} = 0$. We can use this fact in conjunction with the first component of the equations of L84 (11) to determine that the apexes are situated on the surface: $y^2 + z^2 = -x/4 + aF$. This can be viewed as a Poincaré surface of intersection that defines a slice of the attractor – the Poincaré section or PS in short [4]. For any fixed $x$ we recognize the equation of a circle. That is, the surface itself is locally conical, which approximation applies well to the chaotic attractor with $F = 8$ extending between about $[-0.5,2.5]$ wrt. $x$, as seen in Fig. 1. Such a surface can be rectified on the plane spanned by the azimuthal angle

$$\phi = \arctan(y/z),$$

periodic in e.g. $[0, 2\pi]$, and $x$. Therefore, there exists a unique mapping between subsequent pairs of $(x_n, \phi_n)$. This allows for an unfolded view of the PS, and so the prediction of the next apex with certainty. Accordingly, as seen in Fig. 4, the scatter plot exhibits fractality everywhere, and $L = 1$ (or 0) also everywhere. This certainty is compromised in the numerics only by the effect of coarse-graining, when $\{L_b\}$ may be less than unity due to the finite data set size.

By introducing a driving as defined in Sec. 2.3, the dimensionality of the problem increases. Therefore, the same precursory structure of only two quantities is inevitably insuf-
sufficient for predictions with certainty. The scatter plot becomes area-filling, and we will have distributions of the likelihood that take on all values between 0 and 1. However, as long as the forcing strength is moderate, the new features tend to develop through a smearing of the original ones. Evidently, this applies partially in our case.

ROC curves that derive from the distributions seen in Figs. 3 and 4 are shown in Fig. 5. $\mathcal{L}$-ROC curves are always above the $\mathcal{P}$-ROC curves, and the corresponding ROC curves for undriven and driven versions of L84 have the same relation consistently. In particular, for the scenario of the undriven L84 when $x_n = (x_{n-1}, \phi_{n-1})$, the ROC curves approach very near the ideal corner of certainty.

For the various scenarios considered the distances $D$ (6) from the corner, as a summary measure, is provided in Table 1. Besides the scenarios treated in Fig. 5, data is provided in the table also for several other scenarios as follows. First off, we used white noise driving too. Unexpectedly, the predictability with a single delay variable is better for this driven case than the undriven one. This is unchanged even if we apply a smoothing to the time series. Note that $x$ of the WN-driven L84 is a red noise-like nonsmooth process. An improvement of predictability by smoothing is achieved only with larger precursory structures. This effect could be due to a stabilization of the trajectory by noise. However, instead of $x$ considering another observable, namely, the total cyclonic activity in the model,

$$r = \sqrt{y^2 + z^2},$$

no stabilization is prompted; see Table 2. Note that the process of $r$, contrary to that of $x$, is smooth. Therefore, the more likely cause of the unexpected effect is that the geometry of the attractor is altered by noise in a favorable manner, when the PS ‘looks’ less folded ‘in view of $x$’ (but not $r$).

Beside a fast ($\tau' = 1/4$) L63-driving we also considered a slower one with matching time scales of driving and model climate, i.e., $\tau' = 1$. The rationale for this is that we expect that the delay variables, with delay times determined dominantly by the main system, would be able to pick up more information on the driving of a longer decorrelation time. However, while this mechanism should be at work, an improvement of predictability is not registered, but on the contrary – also when using one more delay variable: $x_n = (x_{n-1}, x_{n-2}, x_{n-3})$, or considering observable $r$. The likely cause of this is that, as a counter-effect, the trajectory is destabilized by the driv-
Figure 4: Scatter plots (left) and distributions of the likelihood $L$ (middle) and posterior probability density $P$ (right). In each diagram the color scale ranges from dark blue for 0 to dark red for the maximal density value, which is unity for $L$ for the presented scenarios, but various different values for $P$. In the four rows from top to bottom results are presented for the scenarios specified in Table 1 with the following numbers framed in boxes: 7, 5, 11, 9.
ing more so with \( \tau' = 1 \) than \( 1/4 \).

Beside – but not independent of – the issue of the choice of the observable, we can make an interesting observation to do with the size of the precursory structure too. Extending the two delay variables with a third one (all of the same type) did improve the predictability. However, it was still not as good as with the shorter precursory structure involving the azimuthal angle \( \phi \). This is consistent with Takens’ embedding theorem, as mentioned earlier.

### 3.2. Variable prediction lead time

Let us emphasize that the effect of the destabilization of trajectories, mentioned above, can influence data-driven predictability only in the practical sense of having a finite trajectory length, i.e., finite data set size. Because of the latter, a coarse-graining is inevitable in constructing the histograms. When establishing a correspondence of DDP of POT events with MDP in terms of the Poincaré mapping (not the original flow), the coarse-graining can be translated into terms of errors in measuring initial conditions/precursory observables\(^1\), of size bounded from above by the histogram bin size. In case of a chaotic trajectory the error in tracing the trajectory grows exponentially fast (at least while the error is still small).

This effect will have an impact on the predictability in a way that the further ahead we need to forecast, the less successful we are able to do that. In the situation with an ideal precursory structure as discussed above, the likelihood was evaluated to be nil or unity, or that with a very good approximation, because the prediction lead time was limited by the typical time scale of the system, given that from one apex we intended to predict the next one. We could evaluate the dependence of predictability on the lead time by looking further than the next apex to predict. However, instead of this exercise we prefer to map out the predictability as a continuous function of the prediction lead time, instead of its discrete advances, which are also not known ‘apriori’, i.e., before integrating the system. That is, next we examine the predictability of not POT but TEI events.

In practice \( T, \Delta T, \) and \( t \) can take values that are integer multiples of the trajectory sampling time, which latter is chosen small anyway in order to secure good accuracy of tracing out trajectories by numerical integration of (11). In fact, for this exercise we use the classical Runge-Kutta algorithm/stochastic integrator mentioned above in case of the undriven/WN-driven L84 with fixed \( h = 0.01 \), and we save the state in every 5th time step (to make sure that there will never be two trajectory points subsequent in time in one bin); furthermore we apply \( \Delta T = 2 \times 5 \times h \). As for the precursory structure we will take the triplet of the system state variables \( (x,y,z) \). Therefore, the threshold \( x_\ast \), the interval length \( \Delta T \), and the dynamics itself, determines an event volume in phase space. We then generate data points for the tri-variate histogram \( \{ P_b \} \) by identifying trajectory sample points in the event volume and trace them backward in time by \( T \). Here we choose a bin size arbitrarily, and by ‘predictability’ we mean that conditional to the fixed bin size, not the best possible predictability – given a fixed data set – as a result of some optimization.

The results of evaluating the predictability in terms of the distance \( D \) for the undriven and the WN-driven L84 are displayed in Fig. 6. We can make a number of observations. First, the driven system is less predictable, as

\(^1\)For this point it does not matter whether the Poincaré mapping can be constructed analytically to facilitate the on-demand MDP of POT events. In fact, it is not possible in general even for the most simple chaotic flows. In that case only archival MDP of POT events is possible, whose skill, nevertheless, should be the same as that of the hypothetical on-demand MDP.
Table 1: Summary measure for the ROC-statistics: the ‘distance’ $D$ (6) from the ideal case of all events correctly predicted without making any false alarms. The number of significant digits were determined based on only two independent realizations. The figures in round brackets were obtained by a smoothing of the nonsmooth white noise-driven time series over a moving window of width of a nondimensional time unit. The numbers framed by boxes indicate the correspondence with results shown in Fig. 5.

| Precursory structure | Undriven L84 | L63-driven *, $\tau' = 1$ | L63-driven *, $\tau' = 1/4$ | WN-driven * |
|----------------------|--------------|--------------------------|----------------------------|-------------|
| $x_{n-1}$            | $0.414 \ 0.359$ | $0.480 \ 0.409$ | $0.490 \ 0.410$ | $0.390 \ 0.357$ |
|                      | 4          | 3                       | 2                         | 1           |
| $(x_{n-1}, x_{n-2})$| $0.062 \ 0.027$ | $0.278 \ 0.217$ | $0.265 \ 0.197$ | $0.527 \ 0.320$ |
|                      | 8          | 7                       | 6                         | 5           |
| $(x_{n-1}, \phi_{n-1})$ | $0.0108 \ 0.0054$ | $0.1803 \ 0.1116$ | $0.1840 \ 0.1114$ | $0.2930 \ 0.2587$ |
|                      | 12         | 11                      | 10                        | 9           |
| $(x_{n-1}, x_{n-2}, x_{n-3})$ | $0.0177 \ 0.0087$ | $0.2456 \ 0.1434$ | $0.2257 \ 0.1350$ | $0.4917 \ 0.2828$ |
|                      | 0.062     | 0.025                    | 0.1667 \ 0.0802 | 0.1433 \ 0.0660 | 0.2829 \ 0.1440 |
|                      | 12        | 11                      | 10                        | 9           |

Table 2: Same as in Table 1 but with observable $r$ of cyclonic activity. The number of significant digits is taken to be the same as in case of observable $x$, i.e., not based on a number of independent realizations.

| Precursory structure | Undriven L84 | L63-driven *, $\tau' = 1$ | L63-driven *, $\tau' = 1/4$ | WN-driven * |
|----------------------|--------------|--------------------------|----------------------------|-------------|
| $r_{n-1}$            | $0.475 \ 0.329$ | $0.513 \ 0.368$ | $0.503 \ 0.354$ | $0.599 \ 0.420$ |
|                      | 4          | 3                       | 2                         | 1           |
| $(r_{n-1}, r_{n-2})$| $0.054 \ 0.025$ | $0.192 \ 0.123$ | $0.171 \ 0.098$ | $0.288 \ 0.181$ |
|                      | 12         | 11                      | 10                        | 9           |
| $(r_{n-1}, r_{n-2}, r_{n-3})$ | $0.0150 \ 0.0069$ | $0.1667 \ 0.0802$ | $0.1433 \ 0.0660$ | $0.2829 \ 0.1440$ |
|                      | 0.0150     | 0.0069                    | 0.1667 \ 0.0802 | 0.1433 \ 0.0660 | 0.2829 \ 0.1440 |
expected. Second, the predictability is declining in both cases with increasing prediction lead time, as it should, given the chaotic dynamics and the finite data set size. We have checked that beyond the range of $T$ shown the curves approach the $\sqrt{2}/2$ asymptote, belonging to a straight diagonal ROC curve, meaning no prediction skill at all. This is also expected. Third, on finer scales $D(T)$ is not monotonic in either case, unlike in the well-known case of an auto-regressive AR(1) process of order one, studied regarding data-driven predictability by Hallerberg and Kantz [7]. In the case of L84 the deterministic or autonomous part of its equations results in a chaotic dynamics which is much more complex than the linear bias term of the AR(1). In particular, the bias term of AR(1) need to have a stabilizing effect on the process in order to have a bounded dynamics, while, although on the compact chaotic attractor of L84 trajectories are bounded, they are unstable in a long-term average sense measured by a positive Lyapunov exponent. This instability can also deteriorate predictability, beside a stochastic part if any. On short-terms, however, the deterministic trajectory can experience stable periods, which periods are associated with the return of skill admitted by the plateaus, or negative slopes even, of $D(T)$.

3.3. Variable threshold level

In comparison with the AR(1) process, a further matter of interest is the dependence of predictability on the threshold level. The counterintuitive finding in case of AR(1) was reported by Hallerberg and Kantz [7], namely, that stronger extremes – indifferently to the distribution that the process realizes – are more predictable. The obvious question to ask is, then, whether the latter holds also in case of processes with a more complex deterministic dynamics subjected to stochastic forcing.

Figure 6: Predictability as a function of the prediction lead time. The curve on top (blue) is obtained for the WN-driven L84, and the other one (green) for the undriven L84. Histograms were constructed with 40 bins in all dimensions in the respective ranges where data points are present.

Figure 7 (a) and (c) show the data-driven predictability as a function the prediction lead time as well as the threshold level for the cases of the undriven and WN-driven L84, respectively. Firstly, the nonmonotonic nature of $D(T)$ is prevalent on any fixed threshold level. Secondly, we observe that while for some fixed prediction lead times stronger events are more predictable, i.e., $D(x_*)$ is a decreasing function, it is just the opposite for some other $T$’s. That is, the above statement for AR(1) does not seem to hold in general for more complex dynamical systems.

We assess predictability, with the aim of a comparison, also in terms of the average finite-time Lyapunov exponent, as defined in Sec. 2.2. Note that the definition of the FTLE allows for it to be evaluated only in systems without model uncertainties involved, that is, only in case of the autonomous/undriven L84 here. (In nonautonomous systems where the time-dependence/driving is apriori known, the FTLE can be evaluated, of course.) The results are shown in Fig. 7 (b). Note that a monotonically increasing feature wrt. $T$, as
in the corresponding diagram in panel (a), is missing from this diagram. The error growth for individual trajectories scales exponentially with $T\lambda(T)$, but this would clearly not establish a correspondence either. We will pick up this line of the discussion later. For any fixed $x_\ast$, $\langle \lambda(T) \rangle$ is nonmonotonic, similarly to $D(T)$. However, the ‘waves are straightened out’ for larger $T$’s, as the long-time average is approximated better and better. On the other hand, the threshold-dependence $\langle \lambda(T)(x_\ast) \rangle$ does not seem to be in line with $D(x_\ast)$ at all.

Let us turn back to the problem of data-driven prediction. Figure 7 (c) of the noisy L84 admits values of $D > \sqrt{2}/2$, which should be erroneous. In fact the reason for this error is that the corresponding ROC curves or staircases (not shown) do not extend to the corner (1,1), or more precisely, they feature an excessively large last step – for the reason stated in the end of Sec. 2.1. This is so because the back-traced data points from the event volume are spread out in relatively large domains of the phase space due to the relatively large prediction lead times and strong instabilities (on average) of the trajectories. As also mentioned already, there is a unique optimal (uniform) bin size yielding a minimal $D$. We determined numerically this optimum for each sample combination of $(T,x_\ast)$ separately. This was done using a simple algorithm of maximum finding detailed in the Appendix, which is suitable for treating nonsmooth functions of one variable. The result of this for the undriven and WN-driven L84 can be seen in Fig. 8 (a) and (c), respectively. The surprising outcome with the bin size optimization is that stronger events are generally better predictable, reinstating the rule found by Hallerberg and Kantz [7] for the simple stochastic process of AR(1). Only in the WN-driven case do we see an anomaly for very high thresholds, and at this point it is not clear whether this is just a numerical effect. A further observation is that the nonmonotonic $T$-dependence is suppressed/gone almost completely for the undriven/WN-driven L84.

The optimal number of bins are shown in Fig. 9. Comparing these diagrams with the corresponding ones in Fig. 7 one can notice that larger values of $D$ correspond to fewer and so larger bins. The reason for this is that in these situations the trajectories are more unstable and therefore they scatter in a larger volume, which ‘asks for’ increasing the bin size in order to have a better estimate of the likelihood in those bins.

Finally, we note that model-driven predictability cannot involve such optimization; the bin size is determined by the precision of observation only. Figure 7 does not represent model-driven predictability, however, because in many bins there is an insufficient number of points for the evaluation of the likelihood. For a given data set size the likelihood can be well-approximated in most bins with a bin size larger than the optimal one for DDP. For the data set size in our analysis we evaluate the model-driven predictability as a function of the prediction lead time $T$ and the threshold level $x_\ast$ for a ‘hypothetical’ observational precision that derives from approximately the largest optimal bin size in the considered ranges of $T$ and $x_\ast$, taken to be 25 bins in all dimensions. The interesting result is that also the model-driven predictability is the better the stronger the extremes. This statement fails only in case of the WN-driven L84 in a relatively small regime of the strongest extremes when $T > 5$, likely because the bin size is not larger than the optimal one.

4. Summary and discussion

We examined the predictability of threshold exceedance ‘extreme’ events in a simple
Figure 7: Predictability in terms of $D$ (left) and $\langle \chi(T) \rangle$ (right) as functions of the prediction lead time $T$ and threshold level $x^*$. Results are shown for the undriven (a)-(b) and the WN-driven L84 (c). Histograms in association with (a) and (c) were constructed with 100 bins in all dimensions in the respective ranges where data points are present. Notice the different ranges of $D$ shown.
Figure 8: Diagrams of the kind as in Fig. 7 but (a), (c) having the histogram bin size [uniform wrt. one histogram but different for each combination of $(T, x_*)$] optimized, and (b), (d) with 25 bins in all dimensions. Notice the different ranges of $D$ shown.
but chaotic continuous-time dynamical system or ‘flow’. Given the nature of the problem, namely, that extremes are rare, we chose an arguably [9] appropriate skill score for assessing predictability: the ROC statistics, more specifically, a distance measure $D$ form the ideal situation of having all events successfully predicted without any false alarms. We examined the dependence of predictability on various factors; see Table 3 for a detailed summary of our results.

Our results are illustrative of the interdependence of the following factors: the size of the precursory structure, the choice of observables, and intrinsic properties of the system. Therefore, their respective effects are hard to examine in isolation. Nevertheless, fairly simple arguments can be given for the mechanisms through which they take their respective effects.

Intrinsic properties include the stability of trajectories, which can be quantified by the finite-time $T$ Lyapunov exponent $\lambda^{(T)}$. We compared the overall predictability measured by $D$ to $\langle \lambda^{(T)} \rangle$ as functions of the prediction lead time $T$ and threshold level $x^*$, but did not find a close correspondence. This motivates the development of theory, as future work, to characterize the (nontrivial) involvement of the stability of trajectories in determining predictability. A question in this respect is that whether the Lyapunov exponents can provide all the information, or, a nonlocal description of the deformation of phase space volumes is needed.

Furthermore, we have demonstrated in an arbitrarily chosen chaotic dynamical system that stronger extremes of an arbitrarily chosen observable are better predictable, maintaining an earlier statement made for autoregressive processes [7]. We argued that this applies to both data- and model-driven predictability, since the former, based on a sufficient amount

Figure 9: Optimal number of bins. Panels (a) and (b) correspond to (a) and (c) of Fig. 8. The colorbar applies to both diagrams.
of data, well-approximates the latter, assuming no model errors and some finite precision of the observations, the measure of which is set to be the bin size for binning time series data in the data-driven prediction scheme.

When a given amount of data is available, a unique optimum for the bin size of a regular grid has been found to exist resulting in the best data-driven predictability. Advanced algorithms to bin data will be the objective of future research. In this regard it is envisaged that a theory linking the Lyapunov exponent(s) etc. with the prediction skill could indicate the optimal grid in a straightforward manner, rather than having to find this grid by conducting a costly general iterative optimization procedure. Furthermore, it is often the case that the data set is very limited for pure data-driven prediction, while some model, even if inaccurate at the current stage of its development, is known. In this case it would be most beneficial if a combined data- and (archival or on-demand) model-driven prediction technique could exploit fully the assets of data and model at hand. We will concentrate efforts to develop such techniques.

The said counterintuitive property of model-driven predictability seems to be at odds with the ubiquitous claim that extremes are harder to predict than typical events. The most likely cause of this in our opinion is the inaccuracy of models used for prediction in any practical application. Other possibilities include: the choice of the prediction skill score that might also affects the assessment, or, whether pointwise or binary predictions are made. A systematic analysis of these factors will be the subject of a separate study.

Acknowledgments

Most of this work was carried out at the Max Planck Institute for the Physics of Complex Systems, Dresden, Germany; the scholarship awarded by the Max Planck Society is gratefully acknowledged. The work was supported also by the NAMASTE project owned by Valerio Lucarini (under the ERC grant No. 257106). The author is indebted for useful discussions with Holger Kantz, Tamás Tél, Christian Franzke.

Appendix: Algorithm for finding the approximate global maximum of a nonsmooth function of one variable

The ROC skill score of predictability $D$ is a discontinuous function of the linear bin size $\Delta x$ due to the finite data set size $N$. Conversely, if there was infinite data available, it would be a continuous function. We will assume here that with finite $N$, $D$ features a single minimum. If a function is discontinuous or nonsmooth, the Newton-Raphson algorithm that relies on the derivative cannot be applied to find a global minimum.

Instead of the bin size, we will specify the number of bins (along a single dimension) in the domain where data points are found. Our experience is that, given the data available as specified in Sec. 2.3 and the ranges of $T$ and $x^*$ desired to be explored, the optimal number of bins is between, say, 6 and 200. One could evaluate $D$ for all intermediate integers to find the one that gives the smallest $D$. However, one can do better than that. The following algorithm is applicable to smooth functions $f(x)$ possessing a single maximum, but also to discontinuous/nonsmooth or discrete approximations of such functions, provided that the root-mean-square error of approximation is relatively small (loosely speaking: smaller than the ‘elevation of the maximum’). To start with, we define five equally spaced values of the independent variable $x \in \mathbb{R}$ defined by the choice for the smallest and largest values: $x_{i,j}$,
Table 3: Summary of the different factors/choices that determine predictability and the mechanism through which they do that. Except for the bin size, all other factors apply to data- as well as model-driven predictability.

| Factors                                      | Effect/mechanism                                                                 |
|----------------------------------------------|---------------------------------------------------------------------------------|
| Size of precursory structure                 | Larger size →                                                                  |
|                                               | + Better unfolding                                                              |
|                                               | + More information captured                                                     |
|                                               | ± Introduces new observable                                                    |
| Choice of observables                         | ± Better unfolding – depending on attractor geometry                            |
|                                               | (Ideally: skill is limited only by coarse-graining)                             |
|                                               | ± Alters the involvement of intrinsic characteristics                           |
| Intrinsic characteristics:                   | (i) Stronger instabilities →                                                    |
| (i) Instability of trajectories               | − Poorer predictability,                                                       |
| (ii) Attractor geometry                       | ± but involvement altered by the choice of observable                         |
| (ii) Attractor geometry – depending on the    | ± can result in better unfolding                                               |
| choice of observable –                       |                                                                                  |
| Likelihood vs Posterior PDF                   | + Always better                                                                 |
| Prediction lead time $T$                      | Longer $T$ →                                                                    |
|                                               | ± Prediction gets worse, but possibly with a nonmonotonic rate                  |
|                                               | due to the deterministic dynamics (return of skill)                            |
| Bin size $\Delta x$                          | ± An optimum exists                                                            |
| Event size $x_*$                              | Larger events are                                                              |
|                                               | + more predictable with optimal bin size,                                       |
|                                               | ± but not necessarily without’                                                 |
References

[1] R. A. Fisher, L. H. C. Tippett, Limiting forms of the frequency distribution of the largest and smallest member of a sample, Proc. Cambridge Phil. Soc. 24 (1928) 180–190.

[2] B. V. Gnedenko, Sur la distribution limite du terme maximum d’une serie aleatoire, Annals of Mathematics 44 (1943) 423–453.

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

[4] T. Tél, M. Gruiz, Chaotic Dynamics, Cambridge University Press, Cambridge, 2006.

[5] M. I. Bogachev, A. Bunde, On the predictability of extreme events in records with linear and nonlinear long-range memory: Efficiency and noise robustness, Physica A: Statistical Mechanics and its Applications 390 (12) (2011) 2240 – 2250.

doi: http://dx.doi.org/10.1016/j.physa.2011.02.024.

URL http://www.sciencedirect.com/science/article/pii/S0378437111001506

[6] S. Siegert, J. Bröcker, H. Kantz, Skill of data based predictions versus dynamical models – case study on extreme temperature anomalies, arXiv:1312.4323.

[7] S. Hallerberg, H. Kantz, How does the quality of a prediction depend on the magnitude of the events under study?, Nonlinear Processes in Geophysics 15 (2) (2008) 321–331. doi:10.5194/npg-15-321-2008.

URL http://www.nonlin-processes-geophys.net/15/321/2008/

[8] J. M. Miotto, E. G. Altmann, Predictability of extreme events in social media, arXiv:1403.3616.

[9] S. Hallerberg, J. Bröcker, H. Kantz, Prediction of extreme events, in: R. V. Donner, S. M. Barbosa (Eds.), Nonlinear Time Series Analysis in the Geosciences, Vol. 112, Springer, 2008.

[10] J. Bröcker, L. A. Smith, Scoring probabilistic forecasts: the importance of being proper, Wea. Forecasting 22 (2007) 382–388.

[11] E. N. Lorenz, Irregularity: a fundamental property of the atmosphere, Tellus 36A (1984) 98–110.

[12] C. Franzke, Predictability of extreme events in a nonlinear stochastic-dynamical model, Phys. Rev. E 85 (2012) 031134. doi:10.1103/PhysRevE.85.031134.

URL http://link.aps.org/doi/10.1103/PhysRevE.85.031134

[13] A. E. Sterk, M. P. Holland, P. Rabassa, H. W. Broer, R. Vitolo, Predictability of extreme values in geophysical models, Nonlinear Processes in Geophysics 19 (5) (2012) 529–539. doi:10.5194/npg-19-529-2012.

URL http://www.nonlin-processes-geophys.net/19/529/2012/

[14] A. J. Majda, C. Franzke, D. Crommelin, Normal forms for reduced stochastic climate models, Proceedings of the National Academy of Sciences 106 (10) (2009) 3649–3653. arXiv: http://www.pnas.org/content/106/10/3649.full.pdf+html, doi:10.1073/pnas.0900173106.

URL http://www.pnas.org/content/106/10/3649.abstract

[15] T. Bódai, E. G. Altmann, A. Endler, Stochastic perturbations in open chaotic systems: Random
versus noisy maps, Phys. Rev. E 87 (2013) 042902. doi:10.1103/PhysRevE.87.042902.
URL http://link.aps.org/doi/10.1103/PhysRevE.87.042902

[16] T. Bódai, G. Károlyi, T. Tél, Driving a conceptual model climate by different processes: Snapshot attractors and extreme events, Phys. Rev. E 87 (2013) 022822. doi:10.1103/PhysRevE.87.022822.
URL http://link.aps.org/doi/10.1103/PhysRevE.87.022822

[17] A. Provenzale, N. J. Balmford, Chaos and structures in geophysics and astrophysics (Woods Hole lecture notes) (1999).
URL http://www.whoi.edu/fileserver.do?id=21476&pt=10&p=17353

[18] L. van Veen, Baroclinic flow and the lorenz-84 model, International Journal of Bifurcation and Chaos 13 (08) (2003) 2117–2139. doi:10.1142/S0218127403007904.

[19] F. Takens, Detecting strange attractors in turbulence, in: A. Rand, L.-S. Young (Eds.), Dynamical Systems and Turbulence, Vol. 898, Springer-Verlag, 2008, pp. 366–381.