Physica D 467 (2024) 134246

Contents lists available at ScienceDirect

Physica D

journal homepage: www.elsevier.com/locate/physd

Limits to extreme event forecasting in chaotic systems

Yuan Yuan *, Adrián Lozano-Durán

Department of Aeronautics and Astronautics, Massachusetts Institute of Technology, Cambridge, MA 02139, United States of America

ARTICLE INFO

Communicated by V.M. Perez-Garcia

Keywords:
Extreme events
Chaotic systems
Forecasting
Information theory

ABSTRACT

Predicting extreme events in chaotic systems, characterized by rare but intensely fluctuating properties, is of great importance due to their impact on the performance and reliability of a wide range of systems. Some examples include weather forecasting, traffic management, power grid operations, and financial market analysis, to name a few. Methods of increasing sophistication have been developed to forecast events in these systems. However, the boundaries that define the maximum accuracy of forecasting tools are still largely unexplored from a theoretical standpoint. Here, we address the question: What is the minimum possible error in the prediction of extreme events in complex, chaotic systems? We derive the minimum probability of error in extreme event forecasting along with its information-theoretic lower and upper bounds. These bounds are universal for a given problem, in that they hold regardless of the modeling approach for extreme event prediction: from traditional linear regressions to sophisticated neural network models. The limits in predictability are obtained from the cost-sensitive Fano's and Hellman's inequalities using the Rényi entropy. The results are also connected to Takens' embedding theorem using the *information can't hurt* inequality. Finally, the probability of error for a forecasting model is decomposed into three sources: uncertainty in the initial conditions, hidden variables, and suboptimal modeling assumptions. The latter allows us to assess whether prediction models are operating near their maximum theoretical performance or if further improvements are possible. The bounds are applied to the prediction of extreme events in the Rössler system and the Kolmogorov flow.

1. Introduction

Extreme events, characterized by rare but intensely fluctuating properties, are ubiquitous in both engineering system and natural phenomena [1]. For instance, turbulent gusts over an aircraft can result in bumpy flights [2], severe weather can disrupt communication systems [3], rare but large cascades in electrical power grids may lead to failures [4], extreme ocean temperature oscillations could impact agriculture and ecosystems [5], rare but significant fluctuations in brain network could cause seizures [6], and sudden increases in traffic flow can trigger network paralysis [7]. In these scenarios, the real-time prediction of extreme events is crucial for enabling proactive measures to avert potential issues [8,9]. By accurately forecasting the extreme states of dynamical systems, we can mitigate adverse effects, reduce downtime, and prevent failures. In this study, we investigate the limits of predictability in extreme event detection using the framework of information theory. The limit obtained is a fundamental property – independent of the modeling approach – that arises from the finite amount of information the observed state contains about the extreme event.

A variety of methods have been employed to predict extreme events in time series of chaotic dynamical systems. Some of the approaches that have proven effective include nonlinear dynamics estimation based on the Koopman operator theory [10] and Takens' embedding theorem [11], along with machine learning techniques, such as support vector machines [12], singular spectrum analysis and the maximum entropy method [13]. Advanced deep learning methods, including autoencoders [14], long short-term memory networks [15], and reservoir computing [16] have also been instrumental to devise forecasting models for chaotic systems with high-dimensional attractors. A discussion on the role of information in the context of model prediction and control for chaotic dynamical systems can be found in Ref. [17].

Despite the significant advancements described above, the inherent nature of chaos continues to impose limits on the accuracy of models for extreme event forecasting. The prediction errors in chaotic dynamical systems stem from three primary sources [18]. First, the model might not accurately represent the physical reality. Second, the observable variables may not capture all the relevant degrees of freedom present in the dynamical system. Third, the initial conditions required for forecasting might not be precisely known.

* Corresponding author.
E-mail address: yuany999@mit.edu (Y. Yuan).

https://doi.org/10.1016/j.physd.2024.134246
Received 26 January 2024; Received in revised form 24 April 2024; Accepted 30 May 2024
Available online 1 June 2024

Improvements in the prediction of extreme events can be achieved either by enhancing models to better represent the physics, gaining access to more observables, or reducing uncertainty in the initial conditions. Eliminating modeling errors is theoretically possible, given the knowledge of a set of governing equations that reflect the underlying dynamics of the observed system. However, accessing variables beyond what is currently observable may be limited by experimental or computational constraints. Additionally, no feasible approach can completely eliminate prediction errors caused by uncertainty in the initial conditions. Even with highly precise measurements, minor errors in the initial state eventually amplify due to chaos, compromising the accuracy of the forecast for long times [19]. Here, our focus is not on developing superior models for extreme event prediction. Instead, we pose the fundamental question: what is the theoretically maximum achievable accuracy in extreme event prediction regardless of the modeling approach and source of error?

2. Formulation

2.1. Modeled extreme event indicator

Consider a chaotic dynamical system completely determined by N time-dependent variables given by the vector $\boldsymbol{Q}(t) = [Q_1(t), Q_2(t), \dots, Q_N(t)]$, where t is the time. We are interested in the extreme values of the variable $Q_E(t)$, which is a function of the components of $\boldsymbol{Q}(t)$. The extreme event indicator $E(t)$ is defined as

$$E(t) = \begin{cases} 1 & \text{if } Q_E(t) \geqslant \eta, \\ 0 & \text{otherwise,} \end{cases} \tag{1}$$

where η is the threshold for extreme event detection. The specific value of the threshold η is dependent on the problem and should be selected based on the definition of extreme event for each particular application. The vector of observable variables is defined as $\hat{\boldsymbol{Q}}(t) = [\hat{Q}_1(t), \hat{Q}_2(t), \dots, \hat{Q}_M(t)]$, which contains the accessible information about the system (i.e., the variables that can be measured or are assumed to be known). The components of $\hat{\boldsymbol{Q}}$ correspond to individual components of $\boldsymbol{Q}$ or functions of them. In practical scenarios, $M \leq N$ and the number of observed variables M is equal or smaller than the number of degrees of freedom of the system N.

We aim to build a predictive model for E. To that end, we define the limited-precision observable containing information from the present time and $p \geq 0$ times in the past:

$$\hat{\boldsymbol{Q}}^- = [\hat{\boldsymbol{Q}}(t), \hat{\boldsymbol{Q}}(t-\delta t_1), \dots, \hat{\boldsymbol{Q}}(t-\delta t_p)] \pm \delta\hat{\boldsymbol{Q}}^-, \tag{2}$$

where $\delta t_i > 0, i = 1, 2, \dots, p$ are the time lags used for prediction, and $\delta\hat{\boldsymbol{Q}}^-$ is the uncertainty in the observations. The latter may arise experimentally from inaccuracies in measurement tools, numerically from round-off errors in $\hat{\boldsymbol{Q}}$ or its discretization, and generally, from any uncertainties in the value of $\hat{\boldsymbol{Q}}$. One could forecast the extreme event indicator in the future $E(t+\delta t)$ after a time horizon $\delta t > 0$ using $\hat{\boldsymbol{Q}}^-$ as the input to the model $\hat{f}$ such that

$$\hat{E}(t+\delta t) = \hat{f}\left(\hat{\boldsymbol{Q}}^-\right), \tag{3}$$

where $\hat{E}$ is the modeled extreme event indicator, which might differ from E. The performance of the model can be evaluated using the probability of error

$$P_e(\hat{\boldsymbol{Q}}^-, \hat{f}) = \text{Probability}(\hat{E} \neq E) = P(\hat{f}(\hat{\boldsymbol{Q}}^-) \neq E). \tag{4}$$

Mispredicted extreme events can manifest as either false positives, $P(\hat{E} = 1, E = 0)$, or false negatives, $P(\hat{E} = 0, E = 1)$. However, these two types of errors can bear significantly different consequences. For instance, incorrectly predicting a hurricane (false positive) might be inconvenient but acceptable; however, failing to predict one (false negative) can be catastrophic. To accurately reflect the distinct impact of false positive and negative, we introduce the cost-sensitive probability of error:

$$P_e^c(\hat{\boldsymbol{Q}}^-, \hat{f}) = c^+ P(\hat{E} = 1, E = 0) + c^- P(\hat{E} = 0, E = 1), \tag{5}$$

where $c^+ > 0$ and $c^- > 0$ are the false positive and negative cost weighting factors, respectively. These factors reflect the relative severity of each type of error, and their values are selected according to the specific prediction task. In the case of extreme event prediction, the value of c^- is often larger than c^+. This choice is driven by the understanding that non-extreme events occur more frequently than extreme ones. Prediction models with equal costs ($c^+ = c^-$) are inclined to favor the majority class of non-extreme events. By imposing higher penalties on false negative errors ($c^- > c^+$), we steer the prediction model to focus more on accurately identifying the critical, but less frequent, extreme events. The value of the factors c^- and c^+ is arbitrary, and only their relative magnitude matters. Consequently, c^- and c^+ can be scaled in different manners. To guarantee that the model with the minimum probability of error yields $P_e^c \leq 1/2$, we choose $1/c^+ + 1/c^- = 2$ [see Appendix A for more details].

2.2. Minimum cost-sensitive probability of error

We are interested in the minimum cost-sensitive probability of error given the observable $\hat{\boldsymbol{Q}}^-$ over all possible models $\hat{f}$,

$$P_{e,\min}^c(\hat{\boldsymbol{Q}}^-) = \min_{\hat{f}} P_e^c(\hat{\boldsymbol{Q}}^-, \hat{f}). \tag{6}$$

The minimum cost-sensitive probability of error attainable by any model is [see proof in Appendix A]

$$P_{e,\min}^c(\hat{\boldsymbol{Q}}^-) = \mathbb{E}[I(\hat{\boldsymbol{Q}}^-)] = \sum_{\hat{\boldsymbol{q}}^-} I(\hat{\boldsymbol{Q}}^- = \hat{\boldsymbol{q}}^-) P(\hat{\boldsymbol{Q}}^- = \hat{\boldsymbol{q}}^-), \tag{7}$$

where $\mathbb{E}[I(\hat{\boldsymbol{Q}}^-)]$ is the expectation of I, $\hat{\boldsymbol{q}}^-$ is a particular state (i.e., value) for $\hat{\boldsymbol{Q}}^-$, $P(\hat{\boldsymbol{Q}}^- = \hat{\boldsymbol{q}}^-)$ is the probability of $\hat{\boldsymbol{Q}}^-$ taking the value $\hat{\boldsymbol{q}}^-$, and $I(\hat{\boldsymbol{Q}}^- = \hat{\boldsymbol{q}}^-)$ is the minimum probability of error for the state $\hat{\boldsymbol{Q}}^- = \hat{\boldsymbol{q}}^-$:

$$\begin{aligned} &I(\hat{\boldsymbol{Q}}^- = \hat{\boldsymbol{q}}^-) \\ &= \min\left\{c^- P(E = 1 \mid \hat{\boldsymbol{Q}}^- = \hat{\boldsymbol{q}}^-), c^+\left(1 - P(E = 1 \mid \hat{\boldsymbol{Q}}^- = \hat{\boldsymbol{q}}^-)\right)\right\}, \end{aligned} \tag{8}$$

where $P(E \mid \hat{\boldsymbol{Q}}^- = \hat{\boldsymbol{q}}^-)$ is the probability of E conditioned on $\hat{\boldsymbol{Q}}^- = \hat{\boldsymbol{q}}^-$. The minimum error given by Eq. (7) is the consequence of the unavoidable uncertainty intrinsic to chaotic systems. This uncertainty arises from the lack of knowledge about the variables (e.g., unobserved variables and/or those observed for a limited amount of time) and errors in the initial condition values (e.g. finite precision), which transcend the predictive capabilities of any model.

Eq. (7) provides the precise limit for extreme event forecasting; however, its application to the development, optimization, and evaluation of models for extreme event prediction is challenging due to its non-convex nature. This motivates the derivation of information-theoretic lower and upper bounds for $P_{e,\min}^c(\hat{\boldsymbol{Q}}^-)$ that are more amenable in terms of applications and interpretation. For example, obtaining $P_{e,\min}^c(\hat{\boldsymbol{Q}}^-)$ reliably from Eq. (7) may not be possible in situations where, on the other hand, information-theoretic quantities can be efficiently calculated using estimators [20]. Even when Eq. (7) can be evaluated accurately, its manipulation becomes challenging in the context of model development due to the non-linearity introduced by the $\min(\cdot)$ operator [21]. In such instances, using an information-theoretic formulation of the error facilitates the optimization of model parameters. Information theory can also be employed for feature selection, specifically identifying the input variables that most significantly aid in predicting extreme events [22]. Additionally, the sources of error contributing to $P_{e,\min}^c(\hat{\boldsymbol{Q}}^-)$ are more easily interpreted in terms of information rather than probabilities, since the former adheres to the properties of additivity and the chain rule [23]. In the next section, we derive lower and upper bounds for Eq. (7) using the framework of information theory.

2.3. Information-theoretic bounds for minimum probability of error

The key idea to derive the information-theoretic bounds is that the prediction of extreme events can be intuitively understood as an information transmission process, where information from the current observable state is conveyed to predict the future state [17]. If the forecast is treated as a noisy channel, then the Fano's [24] and Hellman's [25] inequalities provide the foundations for deriving lower and upper bounds on the minimum probability of error in the transmission of discrete messages. We measure the uncertainty in the extreme event indicator E given the information from the observable $\hat{\boldsymbol{Q}}^-$ using the cost-sensitive, conditional Rényi entropy [26,27]

$$H_\alpha^c(E \mid \hat{\boldsymbol{Q}}^-) = \sum_{\hat{\boldsymbol{q}}^-} h_\alpha^c\left(P(E=1 \mid \hat{\boldsymbol{Q}}^- = \hat{\boldsymbol{q}}^-)\right) P(\hat{\boldsymbol{Q}}^- = \hat{\boldsymbol{q}}^-), \tag{9}$$

where

$$h_\alpha^c(p) = h_\alpha\left(\min\{c^- p, c^+(1-p)\}\right)$$

is the cost-sensitive binary Rényi entropy function of order $\alpha > 0$, with

$$h_\alpha(p) = \lim_{\gamma \to \alpha} \frac{1}{1-\gamma} \log_2\left(p^\gamma + (1-p)^\gamma\right).$$

Eq. (9) quantifies the additional information required to determine the outcome of E given the information in $\hat{\boldsymbol{Q}}^-$ accounting for the weighting factors c^- and c^+. It is useful to interpret $H_\alpha^c(E \mid \hat{\boldsymbol{Q}}^-)$ as the uncertainty in E after conducting the 'measurement' of $\hat{\boldsymbol{Q}}^-$. If E and $\hat{\boldsymbol{Q}}^-$ are independent random variables, then $H_\alpha^c(E \mid \hat{\boldsymbol{Q}}^-) = H_\alpha^c(E)$, i.e., knowing $\hat{\boldsymbol{Q}}^-$ does not reduce the uncertainty in E. In this case, $\hat{\boldsymbol{Q}}^-$ is not a useful observable for forecasting E. Conversely, if knowing the observable $\hat{\boldsymbol{Q}}^-$ provides the knowledge to completely determine E, then $H_\alpha^c(E \mid \hat{\boldsymbol{Q}}^-) = 0$, i.e., there is no uncertainty in E given $\hat{\boldsymbol{Q}}^-$, and $\hat{\boldsymbol{Q}}^-$ can potentially be used to predict E with no error. The order α determines the extent to which different probabilities influence the uncertainty, with larger values of α giving greater weight to higher probabilities. For $c^+ = c^- = \alpha = 1$, $H_1^c(E \mid \hat{\boldsymbol{Q}}^-)$ is equal to the classic conditional Shannon entropy [28], which is a concave function in the conditional distribution, making it well-suited for optimization tasks.

The minimum probability of error can be lower and upper bounded as a function of the cost-sensitive conditional Rényi entropy [see proof in Appendix A]

$$P_{e,\min,\mathrm{LB}}^c(\hat{\boldsymbol{Q}}^-) \le P_{e,\min}^c(\hat{\boldsymbol{Q}}^-) \le P_{e,\min,\mathrm{UB}}^c(\hat{\boldsymbol{Q}}^-), \tag{10}$$

where the lower and upper bounds are

$$\begin{aligned} P_{e,\min,\mathrm{LB}}^c(\hat{\boldsymbol{Q}}^-) &= h_\alpha^{-1}\left(H_\alpha^c(E \mid \hat{\boldsymbol{Q}}^-)\right) \\ P_{e,\min,\mathrm{UB}}^c(\hat{\boldsymbol{Q}}^-) &= \min\left\{\frac{1}{2} H_\alpha^c(E \mid \hat{\boldsymbol{Q}}^-), C\right\}, \end{aligned} \tag{11}$$

and $C = \min\left\{c^- P(E=1), c^+(1-P(E=1))\right\}$. Eq. (10) is valid for $0 < \alpha \le 2$, with the tightest bounds achieved for $\alpha = 2$, i.e., the conditional quadratic entropy $H_2^c(E \mid \hat{\boldsymbol{Q}}^-)$. Nonetheless, maintaining the more general formulation with α is beneficial, as it establishes a relationship between error and information within the context of different entropies. It is worth noting that the minimum error in Eq. (7) and the information-theoretic bounds in Eq. (10) hold for any value of the thresholding parameter η defining the cutoff for extreme events. Furthermore, Eqs. (7) and (10) are generally applicable to the prediction of any binary events, whether they are extreme or not.

A corollary from the conditional entropy inequality (a.k.a. *information can't hurt*) [23] is that incorporating additional time lags into the vector of observables can decrease (but never increase) the minimum probability of error [see proof in Appendix B]

$$\begin{aligned} P_{e,\min,\mathrm{LB}}^c(\hat{\boldsymbol{Q}}^{-_l}) &\le P_{e,\min,\mathrm{LB}}^c(\hat{\boldsymbol{Q}}^{-_p}), \text{ for } l > p, \\ P_{e,\min,\mathrm{UB}}^c(\hat{\boldsymbol{Q}}^{-_l}) &\le P_{e,\min,\mathrm{UB}}^c(\hat{\boldsymbol{Q}}^{-_p}), \text{ for } l > p, \end{aligned} \tag{12}$$

where l and p denote the number of time lags in $\hat{\boldsymbol{Q}}^{-_l}$ and $\hat{\boldsymbol{Q}}^{-_p}$, respectively, i.e., $\hat{\boldsymbol{Q}}^{-_l} = [\hat{\boldsymbol{Q}}(t), \hat{\boldsymbol{Q}}(t-\delta t_1), \ldots, \hat{\boldsymbol{Q}}(t-\delta t_l)] \pm \delta\hat{\boldsymbol{Q}}^{-_l}$ (and similarly for $\hat{\boldsymbol{Q}}^{-_p}$). The inequality presented in Eq. (12) is particularly useful in scenarios where not every variable of the system is directly observable. In such instances, it is still possible to lower the minimum probability of error by employing additional time-lagged measurements of the observed variables. This result can be connected to Takens' embedding theorem [29], whereby the dynamics of a dynamical system can be captured by embedding a sequence of past observations into a higher-dimensional space. The latter is consistent with the decrease in the probability of error from Eq. (12). Takens' embedding theorem also states that the delayed-embedding phase space of $\hat{\boldsymbol{Q}}^{-_l}$ is topologically equivalent to the original phase space of the full dynamical system $\boldsymbol{Q}$ for a non-degenerate, noise-free observable with $l > 2d_A$, where d_A is the dimension of the attractor. From an information-theoretic viewpoint, this implies that $\hat{\boldsymbol{Q}}^{-_l}$ provides the same information as the state vector of the full system $\boldsymbol{Q}$ when there is no uncertainty in the observable ($\delta\hat{\boldsymbol{Q}}^{-_l} = \boldsymbol{0}$) and $l > 2d_A$. Under those conditions, $P_{e,\min}^c(\hat{\boldsymbol{Q}}^{-_l}) = P_{e,\min}^c(\boldsymbol{Q}) = 0$ and exact predictions are possible.

The minimum probability of error from Eq. (7) and the information-theoretic bounds from Eq. (10) are derived assuming the discretization of the continuous variable $\hat{\boldsymbol{Q}}$. This reflects the inherent uncertainty in the initial conditions, which is captured by the term $|\delta\hat{\boldsymbol{Q}}^-| > 0$. Nonetheless, the same inequality holds when $\hat{\boldsymbol{Q}}$ is assumed to be a continuous variable by replacing the H_α^c with its continuous extension [see Appendix C]. In that case, there is no uncertainty in the value of $\hat{\boldsymbol{Q}}$, and we can take $\delta\hat{\boldsymbol{Q}}^- = \boldsymbol{0}$. It can also be shown that the right-hand side of Eq. (7) and the bounds in Eq. (11) converge to their continuous counterparts when the partition size used to discretize $\hat{\boldsymbol{Q}}$ is refined towards zero. Here, we focus on the discrete case, as in practical applications there is always some degree of uncertainty in $\hat{\boldsymbol{Q}}$.

2.4. Sources contributing to the probability of error

The probability of error in the extreme event forecaster can be decomposed into three sources: $P_e^c(\hat{\boldsymbol{Q}}^-, \hat{f}) = P_{e,I}^c + P_{e,O}^c + P_{e,M}^c$ given by

$$\begin{aligned} P_{e,I}^c &= P_{e,\min}^c(\boldsymbol{Q}^-), \\ P_{e,O}^c &= P_{e,\min}^c(\hat{\boldsymbol{Q}}^-) - P_{e,\min}^c(\boldsymbol{Q}^-), \\ P_{e,M}^c &= P_e^c(\hat{\boldsymbol{Q}}^-, \hat{f}) - P_{e,\min}^c(\hat{\boldsymbol{Q}}^-), \end{aligned} \tag{13}$$

where the vector $\boldsymbol{Q}^- = \boldsymbol{Q} \pm \delta\boldsymbol{Q}$ contains all the degrees of freedom governing the system $\boldsymbol{Q}$ (i.e., absolute observability) but with finite precision $\delta\boldsymbol{Q}$. The interpretation of each term in Eq. (13) is as follows:

- $P_{e,I}^c$ represents the probability of error solely arising from uncertainty in the initial conditions. This is because $\boldsymbol{Q}^-$ contains all the degrees of freedom of the system, which are sufficient to integrate the system forward in time. However, the process is subject to the initial uncertainty $\delta\boldsymbol{Q}$ such that a higher $\delta\boldsymbol{Q}$ might result in a higher $P_{e,I}^c$. The magnitude of $\delta\boldsymbol{Q}$ varies depending on the specific problem. If $\boldsymbol{Q}^-$ is known with infinite precision (i.e., $\delta\boldsymbol{Q} = \boldsymbol{0}$), then $P_{e,I}^c$ equals zero.
- $P_{e,O}^c$ denotes the probability of error caused by missing information from unobserved variables. This error originates from the fact that $\hat{\boldsymbol{Q}}^-$ contains less information than $\boldsymbol{Q}^-$. As discussed in Section 2.3, the inclusion of multiple time lags in $\hat{\boldsymbol{Q}}^-$ can compensate for the lack of observed variables [29]. However, $P_{e,O}^c$ will still be an important contributor to the total probability of error in those situations where the number of degrees of freedom is much larger than the number of observed variables, $N \gg M$.
- $P_{e,M}^c$ is the probability of error attributable to a suboptimal model. Values of $P_{e,M}^c > 0$ imply that $\hat{f}$ is not efficiently exploiting the information available in $\hat{\boldsymbol{Q}}^-$. In those cases, the model is not operating at its theoretical maximum performance, and further improvements are possible. Conversely, $P_{e,M}^c = 0$ implies that $\hat{f}$ is the best-performing model given the observed variables and uncertainties in the initial conditions.

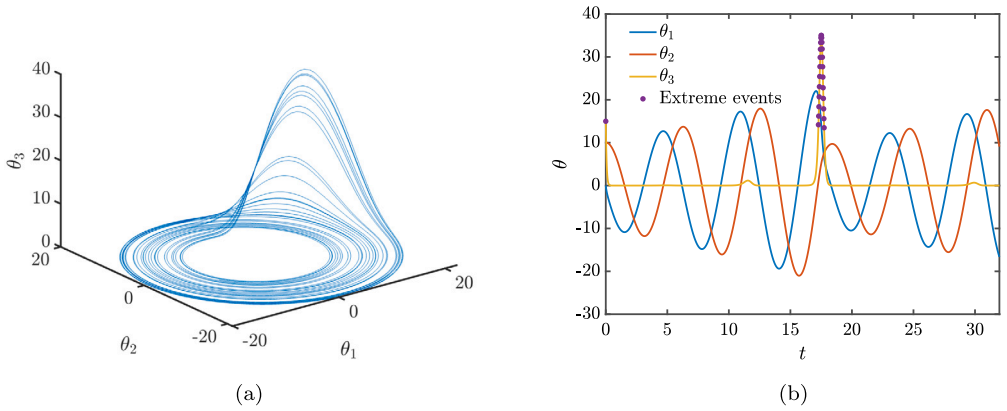

Fig. 1. (a) Trajectory of the Rössler system. (b) Extraction of time series of θ_1, θ_2, θ_3 and extreme events in the Rössler system. Although not shown, the whole time-span of the signals is 10,000 time units.

In the following, we demonstrate the application of our results in two distinct scenarios: the Rössler system and the Kolmogorov flow. The Rössler system offers a simple case for studying extreme events in a chaotic system where all variables can be observed. We use this case to illustrate the classification of errors from Eq. (13). On the other hand, the Kolmogorov flow, characterized by complex, multi-scale interactions among numerous degrees of freedom, represents the dynamics of extreme events found in more realistic systems. This case is used to demonstrate the effect of cost-sensitive analysis.

3. Applications

3.1. Rössler system

The Rössler system with state variables $\boldsymbol{Q} = [\theta_1, \theta_2, \theta_3]$ is governed by the ordinary differential equation:

$$\begin{aligned} \frac{\mathrm{d}\theta_1}{\mathrm{d}t} &= -\theta_2 - \theta_3, \\ \frac{\mathrm{d}\theta_2}{\mathrm{d}t} &= \theta_1 + a\theta_2, \\ \frac{\mathrm{d}\theta_3}{\mathrm{d}t} &= b + \theta_3(\theta_1 - c), \end{aligned} \tag{14}$$

with parameters $a = 0.1, b = 0.1$, and $c = 14$. We investigate extreme events in θ_3, which exhibits rare excursions of intense magnitude. Fig. 1(a) shows the trajectory of the Rössler system in the three-dimensional phase space. The extreme event indicator is defined as

$$E(t) = \begin{cases} 1 & \text{if } \theta_3(t) \geqslant \bar{\theta}_3 + 3\sigma_{\theta_3}, \\ 0 & \text{otherwise,} \end{cases} \tag{15}$$

where $\bar{\theta}_3$ and σ_{θ_3} are the mean and standard deviation of θ_3 over time. The threshold is set to $\eta = \bar{\theta}_3 + 3\sigma_{\theta_3}$, but the conclusions drawn in this section apply to other values of η. Results for a higher threshold can be found in Appendix D. Fig. 1(b) contains a fragment of the time history of $\boldsymbol{Q}$ and the extreme event indicator E.

We investigate the case with balanced risk $c^+ = c^- = 1$ and define the normalized probability of error as $\bar{P}_e^c = P_e^c/C$. This normalization is such that $\bar{P}_{e,\min}^c \to 1$ for $\delta t \to \infty$ in practical applications. Fig. 2 shows the normalized minimum probability of error as a function of time-horizon for extreme event prediction δt using Eq. (7). Three scenarios are considered.

- In the first case, we assume that the only observable variable is $\hat{Q}_1^- = \theta_3(t) \pm \delta\theta_3(t)$, where the uncertainty in the initial condition is set to $\delta\theta_3 \leq 0.05\sigma_{\theta_3}$. Here, the uncertainty $\delta\theta_3$ is not introduced by perturbing the equations of the system. Instead, the uncertainty is incorporated in a non-intrusive manner when calculating the probability $P(\theta_3)$ by discretizing θ_3 into bins of size $2\delta\theta_3$. This is equivalent to assuming that the solution passing through θ_3 cannot be distinguished from another trajectory, also contained within the attractor of the system, at a distance from θ_3 equal to or less than $\delta\theta_3$. The associated minimum probability of error, $\bar{P}_{e,\min}^c(\hat{Q}_1^-)$, is represented by the solid line in Fig. 2(a).
- In the second scenario, the observable includes two time lags in addition to the present time: $\hat{\boldsymbol{Q}}_2^- = [\theta_3(t), \theta_3(t-\delta t), \theta_3(t-2\delta t)] \pm \delta\theta_3$, where the uncertainty $\delta\theta_3$ is again set to $0.05\sigma_{\theta_3}$ for all time lags. The minimum probability of error, $\bar{P}_{e,\min}^c(\hat{\boldsymbol{Q}}_2^-)$, is depicted by the solid line in Fig. 2(b). The difference between the two minimal errors $\bar{P}_{e,\min}^c(\hat{\boldsymbol{Q}}_2^-)$ and $\bar{P}_{e,\min}^c(\hat{Q}_1^-)$ from Fig. 2(a) serves as a measure of the improvement in predictive accuracy gained by incorporating observations from two additional times in θ_3.
- In the third scenario, it is assumed that the knowledge of the full state is available at the present time with finite precision, i.e., $\boldsymbol{Q}^- = [\theta_1(t), \theta_2(t), \theta_3(t)] \pm \delta\boldsymbol{\theta}$, where $\delta\boldsymbol{\theta} \leq 0.05[\sigma_1, \sigma_2, \sigma_3]$ with σ_i the standard deviation of θ_i. The minimum probability of error, $\bar{P}_{e,\min}^c(\boldsymbol{Q}^-)$, is indicated by the dashed line in Fig. 2(a) and 2(b). Errors arising from uncertainty in initial conditions are quantified by $\bar{P}_{e,I}^c = \bar{P}_{e,\min}^c(\boldsymbol{Q}^-)$ (highlighted by the purple shaded region in Fig. 2). The discrepancy between $\bar{P}_{e,\min}^c(\hat{Q}_1^-)$ and $\bar{P}_{e,\min}^c(\boldsymbol{Q}^-)$ in Fig. 2(a), and between $\bar{P}_{e,\min}^c(\hat{\boldsymbol{Q}}_2^-)$ and $\bar{P}_{e,\min}^c(\boldsymbol{Q}^-)$ in Fig. 2(b), allows us to quantify the errors resulting from the lack of knowledge of θ_1 and θ_2 (i.e., $\bar{P}_{e,O}^c$, indicated by the yellow shaded region in Fig. 2).

The region beneath each curve, $\bar{P}_{e,\min}^c(\hat{Q}_1^-)$, $\bar{P}_{e,\min}^c(\hat{\boldsymbol{Q}}_2^-)$, and $\bar{P}_{e,\min}^c(\boldsymbol{Q}^-)$, corresponds to models that are unattainable given the observable $\hat{Q}_1^-$, $\hat{\boldsymbol{Q}}_2^-$, and $\boldsymbol{Q}^-$, respectively, whereas the region above represents models that are suboptimal. Over time, all cases converge to $\bar{P}_e^c \to 1$ given the chaotic nature of the system. This convergence is slower for $\bar{P}_{e,\min}^c(\boldsymbol{Q}^-)$, as errors are only due to uncertainties in the initial condition.

To illustrate the errors from an actual predictive model, we trained decision tree models, $\hat{f}^{DT}$, to predict E using either $\hat{Q}_1^-$ or $\hat{\boldsymbol{Q}}_2^-$ as input. Different decision tree models are trained to forecast E at each δt. The maximum number of branch node splits is 8, and each leaf contains at least 10 observations. The results are also included in Fig. 2(a) and (b). Additional details about the confusion matrix for the decision tree model can be found in Appendix D. The normalized probability of error for the decision tree models, $\bar{P}_e^c(\hat{Q}_1^-, \hat{f}^{DT})$ and $\bar{P}_e^c(\hat{\boldsymbol{Q}}_2^-, \hat{f}^{DT})$, enables the quantification of the model error $\bar{P}_{e,M}^c$ for the specific case. The results in Fig. 2(a) show that the model error $\bar{P}_e^c(\hat{Q}_1^-, \hat{f}^{DT})$ closely approaches the minimum error given by $\bar{P}_{e,\min}^c(\hat{Q}_1^-)$, indicating that the model is operating near its maximum theoretical performance. On the other hand, the results in Fig. 2(b) reveal a gap between the model error $\bar{P}_e^c(\hat{\boldsymbol{Q}}_2^-, \hat{f}^{DT})$ and the minimum theoretical error $\bar{P}_{e,\min}^c(\hat{\boldsymbol{Q}}_2^-)$ for $\delta t < 0.4$.

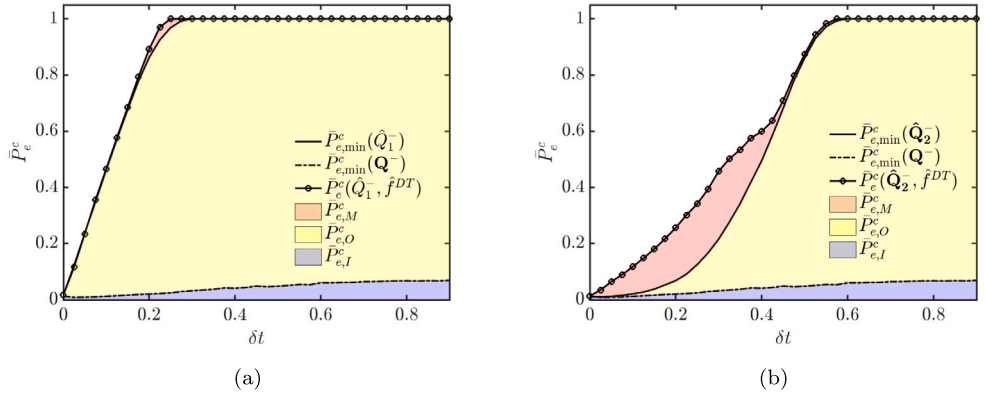

Fig. 2. Normalized probability of error for extreme events prediction in the Rössler system for the threshold $\eta = \bar{\theta}_3 + 3\sigma_{\theta_3}$ using as observable (a) $\hat{\boldsymbol{Q}}_1^- = \theta_3(t) \pm \delta\theta_3$ and (b) $\hat{\boldsymbol{Q}}_2^- = [\theta_3(t), \theta_3(t-\delta t), \theta_3(t-2\delta t)] \pm \delta\theta_3$. $\bar{P}^c_{e,\min}(\hat{Q}_1^-)$, $\bar{P}^c_{e,\min}(\hat{\boldsymbol{Q}}_2^-)$, and $\bar{P}^c_{e,\min}(\boldsymbol{Q}^-)$ are the minimum probability of error using the observable $\hat{Q}_1^-$, $\hat{\boldsymbol{Q}}_2^-$, and $\boldsymbol{Q}^- = [\theta_1(t), \theta_2(t), \theta_3(t)] \pm \delta\boldsymbol{\theta}$, respectively. $\bar{P}^c_{e,I}$ (purple) is the error due to uncertainty in the initial conditions; $\bar{P}^c_{e,O}$ (yellow) is the error caused by unobserved variables; $\bar{P}^c_{e,M}$ (red) is the error due to suboptimal model.

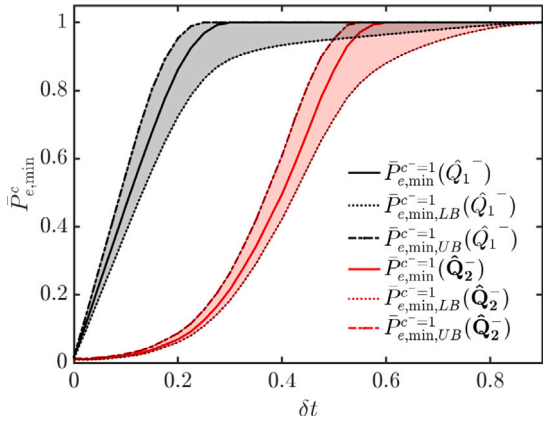

Fig. 3. Normalized information-theoretic upper and lower bounds of the minimum probability of error for extreme events prediction in the Rössler system for the threshold $\eta = \bar{\theta}_3 + 3\sigma_{\theta_3}$ and observables $\hat{Q}_1^-$ and $\hat{\boldsymbol{Q}}_2^-$. The solid line is $\bar{P}^c_{e,\min}$ and the shaded area represents the region confined within $\bar{P}_{e,\min,\mathrm{LB}}$ and $\bar{P}_{e,\min,\mathrm{UB}}$ obtained for the second-order, cost-sensitive conditional Rényi entropy.

This indicates that the model is suboptimal, and models with improved performance are possible. For both cases, as δt increases, $\bar{P}^c_{e,O}$ rapidly becomes the predominant source of error. Conversely, $\bar{P}^c_{e,I}$ is minor compared to $\bar{P}^c_{e,O}$. Hence, the analysis also shows that missing variables have a greater impact on the accuracy of the forecast compared to suboptimal modeling and uncertainty in the initial conditions.

Finally, we compare the exact minimum probability of error from Eq. (7) with the information-theoretic bounds from Eq. (11) for the second-order, cost-sensitive conditional Rényi entropy. The results are presented in Fig. 3 using either $\hat{Q}_1^-$ or $\hat{\boldsymbol{Q}}_2^-$ as observables. In both cases, the bounds provide a narrow region within which $\bar{P}^c_{e,\min}$ must be confined. In situations where directly obtaining $\bar{P}^c_{e,\min}$ is challenging, the region defined by the upper and lower bounds can be used to demarcate the theoretical zone of near-optimal operation for a model. If the error falls within this zone, the model can be considered as possibly operating near its best theoretical performance.

3.2. Kolmogorov flow

Next, we evaluate the cost-sensitive error bounds for forecasting intense energy dissipation events in a turbulent flow [30]. The case considered is the Kolmogorov flow: a high-dimensional, chaotic dynamical system described by the two-dimensional Navier–Stokes equations and driven by monochromatic body forcing [31]:

$$\frac{\partial u_i}{\partial t} = -\frac{\partial (u_i u_j)}{\partial x_j} - \frac{\partial \Pi}{\partial x_i} + \frac{1}{Re}\frac{\partial^2 u_i}{\partial x_k \partial x_k} + f_i, \quad i = 1, 2, \tag{16}$$

$$\frac{\partial u_i}{\partial x_i} = 0, \tag{17}$$

where repeated indices imply summation, $u_i(x_1, x_2, t)$ is the ith velocity component, Π is the pressure, x_1 and x_2 are the spatial coordinates, and f_i is the forcing term with $f_1 = \sin(4x_2)$ and $f_2 = 0$. The velocity vector is denoted as $\boldsymbol{u} = [u_1, u_2]$. The flow setup is characterized by the Reynolds number $Re = 100$ for which the flow exhibits intermittent bursts of dissipation events [32]. For our analysis, we use data from Farazmand and Sapsis [30], obtained by numerically resolving all the scales of the problem in a doubly periodic box with side 2π and 256^2 spatial Fourier modes. Fig. 4(a) shows the velocity amplitude $|\boldsymbol{u}(x_1, x_2, t)|$ at a given time.

Our focus is on the prediction of extreme events characterized by fluctuations in the mean dissipation rate of kinetic energy $D(t) = \langle 2S_{ij}S_{ij}/Re \rangle$, where $S_{ij} = 1/2(\partial u_i/\partial x_j + \partial u_j/\partial x_i)$ is the rate-of-strain tensor, and $\langle \cdot \rangle$ denotes average in space. The extreme event indicator is

$$E(t) = \begin{cases} 1 & \text{if } D(t) \geqslant \bar{D} + 1.5\sigma_D, \\ 0 & \text{otherwise,} \end{cases} \tag{18}$$

where $\bar{D}$ and σ_D are the mean and standard deviation of D over time. Predictions of extreme events using a higher threshold can be found in Appendix B. The observable chosen is the magnitude of the spatial Fourier mode of $\boldsymbol{u}$ corresponding to the wavenumber $[1, 0]$, which is denoted by $\hat{Q} = |\hat{u}_{1,0}|$. The latter is one of the preferred observables for predicting extreme dissipation events in the Kolmogorov flow, as it has been shown to correlate with the growth of D [30]. An excerpt of $\hat{Q}$ and D extracted from the full time history is presented in Fig. 4(b). The time is non-dimensionalized by $t_e = (Re\bar{D})^{-1/2}$.

Three cases are investigated with increasing cost for false positives: $c^- = 1$, 1.5, and 2. The first case ($c^- = 1$) penalizes false positives and false negatives equally. The other two cases assign a higher penalty to false negatives, such that the cost of failing to predict an extreme event is two times (for $c^- = 1.5$) or three times (for $c^- = 2$) the cost of incorrectly predicting a non-extreme event. We denote the minimum probability of error for each case as $P^{c^-=1}_{e,\min}(\hat{Q}^-)$, $P^{c^-=1.5}_{e,\min}(\hat{Q}^-)$, and $P^{c^-=2}_{e,\min}(\hat{Q}^-)$, respectively. Fig. 5 shows the results as a function of δt. The minimum cost-sensitive probability of error is normalized as $\bar{P}^c_{e,\min} = P^c_{e,\min}/C$ such that $\bar{P}^c_{e,\min} \to 1$ for $\delta t \to \infty$.

The results from Fig. 5 show that increasing c^- reduces the normalized minimum probability of error, making the predictions less challenging. This trend is particular to the Kolmogorov flow and the

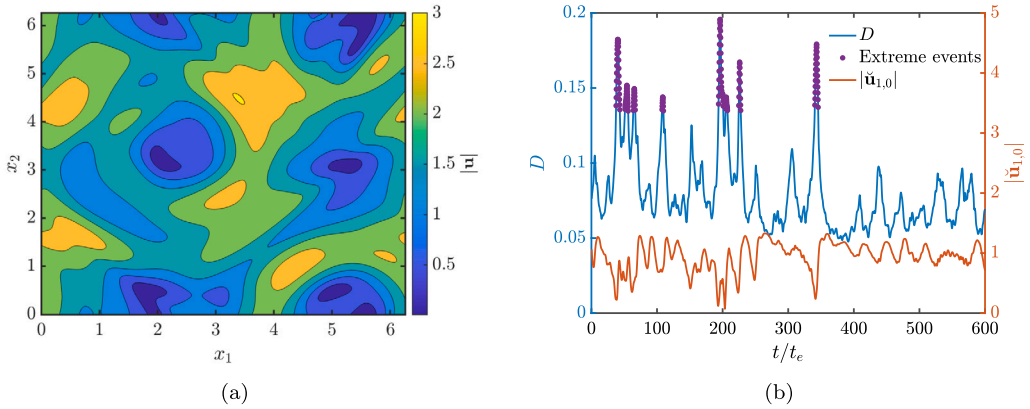

Fig. 4. (a) The instantaneous velocity amplitude $|\boldsymbol{u}|$ for the Kolmogorov flow. (b) Time history of $D(t)$, $|\check{\boldsymbol{u}}_{1,0}|(t)$, and $E(t)$ for $\eta = \bar{D} + 1.5\sigma_D$.

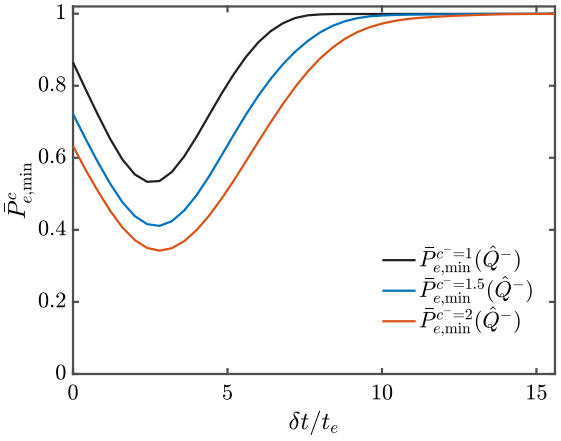

Fig. 5. Normalized, cost-sensitive, minimum probability of error for extreme events prediction in the Kolmogorov flow for $\eta = \bar{D} + 1.5\sigma_D$ and $c^- = 1$, 1.5, and 2.

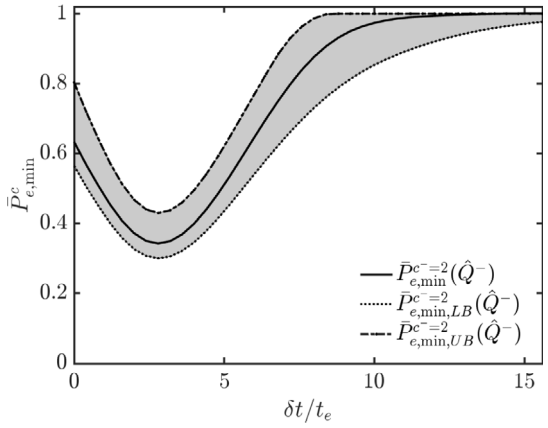

Fig. 6. Normalized, cost-sensitive, information-theoretic upper and lower bounds of the minimum probability of error for extreme events prediction in the Kolmogorov flow for $\eta = \bar{D}+1.5\sigma_D$. The solid line is $\bar{P}^c_{e,\min}$ and the shaded area represents the region confined within $\bar{P}_{e,\min,\mathrm{LB}}$ and $\bar{P}_{e,\min,\mathrm{UB}}$ obtained for the cost-sensitive, second-order, conditional Rényi entropy.

chosen variables, and other systems can exhibit different behavior. Fig. 5 also illustrates two interesting characteristics of $\bar{P}^c_{e,\min}(\hat{Q}^-)$ for the three c^- values considered. First, $\bar{P}^c_{e,\min}(\hat{Q}^-)$ is not initially zero when $\delta t = 0$. This situation arises because D was not considered an observed variable, leading to uncertainty in E even at $\delta t = 0$. The second interesting observation is that $\bar{P}^c_{e,\min}$ reaches its lowest value at $\delta t_{\min} = 2.8t_e$. This observation can be understood by noting that in this system, energy is transferred among different scales due to nonlinear interactions until it is ultimately dissipated. This process occurs on a timescale comparable to $\delta t_{\min}$ [30], which explains the effectiveness of $|\check{\boldsymbol{u}}_{1,0}|$ in predicting extreme dissipation events at that time lag. For times beyond this point, $|\check{\boldsymbol{u}}_{1,0}|$ becomes increasingly less effective due to the chaoticity of turbulence.

Finally, we compare the normalized information-theoretic upper and lower bounds of the minimum probability of error when $c^- = 2$. The results, presented in Fig. 6, show that the bounds accurately reflect the trend observed for $\bar{P}^{c^-=2}_{e,\min}$: there is a non-zero minimum probability of error at $\delta t = 0$, which initially decreases before eventually increasing towards one.

4. Conclusions

In this study, we have derived the minimum cost-sensitive probability of error in extreme event forecasting (Eq. (7)) along with its information-theoretic lower and upper bounds (Eq. (11)). The bounds are rooted in the cost-sensitive Fano's and Hellman's inequalities for the Rényi entropy. Furthermore, the minimum probability of error and its bounds are applicable to scenarios involving both balanced and unbalanced risks. The results are also connected to Takens' embedding theorem using the *information can't hurt* inequality, which shows that incorporating additional time lags into the vector of observables can decrease (but never increase) the minimum probability of error. The probability of error for a forecasting model was also decomposed into three sources: uncertainty in the initial conditions, hidden variables, and suboptimal modeling assumptions (Eq. (13)).

We have demonstrated the application of these bounds to determine the limits of extreme event prediction in two cases related to fluid dynamics: the Rössler system and the Kolmogorov flow. These applications illustrate the utility of the minimum probability of error and its lower and upper bounds as tools for investigating the intrinsic limitations of forecasting extreme events in chaotic systems. We have shown that the minimum error and its bounds maintain their validity irrespective of the chosen modeling method and play a crucial role in assessing whether models are functioning near their theoretical optimum. Future efforts will be devoted to understanding the limits of predictability for extreme events such as dissipation, wall-shear stress, and wall-pressure in turbulent flows at high Reynolds numbers, which are relevant for advancing the field of external aerodynamics. Nonetheless, the method presented here is generally applicable to problems in other fields such as economics, biology, and finance, among others.

Finally, the use of Eqs. (7) and (11) extends beyond the extreme event predictions presented here; they are broadly applicable to any

binary classification of events, whether they are categorized as extreme or non-extreme. Eqs. (7) and (11) also lay the foundation for future extensions to forecasting continuous in time signals by employing generalized versions of Fano's and Hellman's inequalities.

CRediT authorship contribution statement

Yuan Yuan: Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Investigation, Conceptualization. **Adrián Lozano-Durán:** Writing – review & editing, Writing – original draft, Project administration, Methodology, Investigation, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Yuan Yuan reports financial support was provided by Massachusetts Institute of Technology. Adrian Lozano Duran reports financial support was provided by Massachusetts Institute of Technology. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This work was supported by the National Science Foundation, United States under Grant No. 2140775 and MISTI Global Seed Funds, United States.

Appendix A. Proof of information-theoretic bounds

The minimum cost-sensitive probability of error achievable by any model based on the cost-sensitive uncertainty in E conditioned to the observable $\hat{\boldsymbol{Q}}^-$ is

$$P^c_{e,\min}(\hat{\boldsymbol{Q}}^-) = \mathbb{E}[I(\hat{\boldsymbol{Q}}^-)], \tag{A.1}$$

and it is lower and upper bounded as

$$h_\alpha^{-1}\left(H^c_\alpha(E \mid \hat{\boldsymbol{Q}}^-)\right) \le P^c_{e,\min}(\hat{\boldsymbol{Q}}^-) \le \min\left\{\frac{1}{2}H^c_\alpha(E \mid \hat{\boldsymbol{Q}}^-), C\right\}, \tag{A.2}$$

where $h_\alpha(p)$ is a concave function for $p \in [0, 1/2]$ when $0 < \alpha \le 2$, and $C = \min\left\{c^- P(E=1), c^+(1-P(E=1))\right\}$.

Proof. The overall cost-sensitive probability of error is a weighted sum of the probability of error for each specific state of $\hat{\boldsymbol{Q}}^-$,

$$\begin{aligned} P^c_e(\hat{\boldsymbol{Q}}^-, \hat{f}) &= c^- P(\hat{E}=0, E=1) + c^+ P(\hat{E}=1, E=0) \\ &= \sum_{\hat{\boldsymbol{q}}^-} \Big(c^- P(\hat{E}=0, E=1 \mid \hat{\boldsymbol{Q}}^- = \hat{\boldsymbol{q}}^-) \\ &\quad + c^+ P(\hat{E}=1, E=0 \mid \hat{\boldsymbol{Q}}^- = \hat{\boldsymbol{q}}^-) \Big) P(\hat{\boldsymbol{Q}}^- = \hat{\boldsymbol{q}}^-), \end{aligned} \tag{A.3}$$

where the factors c^- and c^+ are scaled as $1/c^- + 1/c^+ = 2$ to ensure that

$$\begin{aligned} P^c_{e,\min}(\hat{\boldsymbol{Q}}^-) &\le \begin{cases} c^- P(E=1), & \text{if } \hat{E}=0 \\ c^+ P(E=0), & \text{if } \hat{E}=1 \end{cases} \\ &\le \min\left\{c^- P(E=1), c^+(1-P(E=1))\right\} \\ &= C \le \frac{c^- c^+}{c^+ + c^-} = \frac{1}{2}. \end{aligned} \tag{A.4}$$

This convention was adopted to avoid $P^c_{e,\min}(\hat{\boldsymbol{Q}}^-) > 1/2$, as in those situations, the model with the minimum probability of error could be obtained by flipping the model with $P^c_{e,\min}(\hat{\boldsymbol{Q}}^-) > 1/2$ to the one with a probability of $1 - P^c_{e,\min}(\hat{\boldsymbol{Q}}^-) \le 1/2$. For each specific state of $\hat{\boldsymbol{Q}}^-$, the minimum probability of error is determined by

$$\begin{aligned} &c^- P(\hat{E}=0, E=1 \mid \hat{\boldsymbol{Q}}^- = \hat{\boldsymbol{q}}^-) + c^+ P(\hat{E}=1, E=0 \mid \hat{\boldsymbol{Q}}^- = \hat{\boldsymbol{q}}^-) \\ &= \begin{cases} c^- P(\hat{E}=0, E=1 \mid \hat{\boldsymbol{Q}}^- = \hat{\boldsymbol{q}}^-), & \text{if } \hat{E}(\hat{\boldsymbol{q}}^-) = 0 \\ c^+ P(\hat{E}=1, E=0 \mid \hat{\boldsymbol{Q}}^- = \hat{\boldsymbol{q}}^-), & \text{if } \hat{E}(\hat{\boldsymbol{q}}^-) = 1 \end{cases} \\ &= \begin{cases} c^- P(E=1 \mid \hat{\boldsymbol{Q}}^- = \hat{\boldsymbol{q}}^-) - c^- \underbrace{P(\hat{E}=1, E=1 \mid \hat{\boldsymbol{Q}}^- = \hat{\boldsymbol{q}}^-)}_{0}, \\ \quad \text{if } \hat{E}(\hat{\boldsymbol{Q}}^- = \hat{\boldsymbol{q}}^-) = 0 \\ c^+ P(E=0 \mid \hat{\boldsymbol{Q}}^- = \hat{\boldsymbol{q}}^-) - c^+ \underbrace{P(\hat{E}=0, E=0 \mid \hat{\boldsymbol{Q}}^- = \hat{\boldsymbol{q}}^-)}_{0}, \\ \quad \text{if } \hat{E}(\hat{\boldsymbol{Q}}^- = \hat{\boldsymbol{q}}^-) = 1 \end{cases} \\ &\ge \min\left\{c^- P(E=1 \mid \hat{\boldsymbol{Q}}^- = \hat{\boldsymbol{q}}^-), c^+ P(E=0 \mid \hat{\boldsymbol{Q}}^- = \hat{\boldsymbol{q}}^-)\right\} \\ &= \min\left\{c^- P(E=1 \mid \hat{\boldsymbol{Q}}^- = \hat{\boldsymbol{q}}^-), c^+\left(1 - P(E=1 \mid \hat{\boldsymbol{Q}}^- = \hat{\boldsymbol{q}}^-)\right)\right\}. \end{aligned} \tag{A.5}$$

Let us define the minimum probability of error at each state as

$$\begin{aligned} &I(\hat{\boldsymbol{Q}}^- = \hat{\boldsymbol{q}}^-) \\ &= \min\left\{c^- P(E=1 \mid \hat{\boldsymbol{Q}}^- = \hat{\boldsymbol{q}}^-), c^+\left(1 - P(E=1 \mid \hat{\boldsymbol{Q}}^- = \hat{\boldsymbol{q}}^-)\right)\right\}. \end{aligned} \tag{A.6}$$

Applying Eqs. (A.3) and (A.5) we have,

$$P^c_{e,\min}(\hat{\boldsymbol{Q}}^-) = \sum_{\hat{\boldsymbol{q}}^-} I(\hat{\boldsymbol{Q}}^- = \hat{\boldsymbol{q}}^-) P(\hat{\boldsymbol{Q}}^- = \hat{\boldsymbol{q}}^-) = \mathbb{E}[I(\hat{\boldsymbol{Q}}^-)] \le P^c_e(\hat{\boldsymbol{Q}}^-, \hat{f}), \tag{A.7}$$

which is the Bayes error rate typically discussed in statistical classification [33] but applied here in the context of extreme event prediction.

Given the concave function

$$h_\alpha(p) = \lim_{\gamma \to \alpha} \frac{1}{1-\gamma} \log_2\left(p^\gamma + (1-p)^\gamma\right)$$

for $p \in [0, 0.5]$[34] when $\alpha \in (0, 2]$, the Jensen's inequality for the random variable $I(\hat{\boldsymbol{Q}}^-)$ results in

$$h_\alpha\left(P^c_{e,\min}(\hat{\boldsymbol{Q}}^-)\right) = h_\alpha\left(\mathbb{E}[I(\hat{\boldsymbol{Q}}^-)]\right) \ge \mathbb{E}\left[h_\alpha\left(I(\hat{\boldsymbol{Q}}^-)\right)\right]. \tag{A.8}$$

The right hand side of Eq. (A.8) is

$$\begin{aligned} &\mathbb{E}\left[h_\alpha\left(I(\hat{\boldsymbol{Q}}^-)\right)\right] = \sum_{\hat{\boldsymbol{q}}^-} h_\alpha\left(I(\hat{\boldsymbol{Q}}^- = \hat{\boldsymbol{q}}^-)\right) P(\hat{\boldsymbol{Q}}^- = \hat{\boldsymbol{q}}^-) \\ &= \sum_{\hat{\boldsymbol{q}}^-} h_\alpha\left(\min\left\{c^- P(E=1 \mid \hat{\boldsymbol{Q}}^- = \hat{\boldsymbol{q}}^-), c^+(1 - P(E=1 \mid \hat{\boldsymbol{Q}}^- = \hat{\boldsymbol{q}}^-))\right\}\right) \\ &\quad P(\hat{\boldsymbol{Q}}^- = \hat{\boldsymbol{q}}^-) \\ &= \sum_{\hat{\boldsymbol{q}}^-} h^c_\alpha\left(P(E=1 \mid \hat{\boldsymbol{Q}}^- = \hat{\boldsymbol{q}}^-)\right) P(\hat{\boldsymbol{Q}}^- = \hat{\boldsymbol{q}}^-) = H^c_\alpha(E \mid \hat{\boldsymbol{Q}}^-), \end{aligned} \tag{A.9}$$

where $H^c_\alpha(E \mid \hat{\boldsymbol{Q}}^-)$ is defined as the cost-sensitive conditional Rényi entropy, and the cost-sensitive binary entropy function is given by

$$h^c_\alpha(p) = \begin{cases} h_\alpha(c^- p) & \text{for } p \in \left[0, \frac{1}{2c^-}\right), \\ h_\alpha\left(c^+(1-p)\right) & \text{for } p \in \left[\frac{1}{2c^-}, 1\right], \end{cases} \tag{A.10}$$

such that

$$h^c_\alpha(p) = h_\alpha\left(\min\left\{c^- p, c^+(1-p)\right\}\right). \tag{A.11}$$

The function $h^c_\alpha(p)$ emerges naturally as a measure of information. The factors c^+ and c^- weight the importance (risk) of each event, whereas the order α controls how the different probabilities in the distribution contribute to the overall measure of uncertainty. For increasing val-

ues of α, the measure gives more weight to larger probabilities. For example, consider the process of tossing a coin where both heads and tails are assigned equal importance (c^+ and c^- for heads and tails, respectively, with $c^+ = c^- = 1$). The greatest amount of information (i.e., uncertainty) regarding the outcome corresponds to the probability $p = 1/2$ that maximizes the binary entropy function $h_\alpha(p)$. However, when the importance of the outcomes differs (e.g., heads is preferred over tails, $c^+ > c^-$), the greatest uncertainty is achieved at $p = \frac{c^+}{c^+ + c^-}$, maximizing cost-sensitive binary entropy $h^c_\alpha(p)$. In conclusion,

$$H^c_\alpha(E \mid \hat{\boldsymbol{Q}}^-) = \mathbb{E}\left[h_\alpha\left(I(\hat{\boldsymbol{Q}}^-)\right)\right] \leq h_\alpha\left(\mathbb{E}[I(\hat{\boldsymbol{Q}}^-)]\right) = h_\alpha\left(P^c_{e,\min}(\hat{\boldsymbol{Q}}^-)\right), \tag{A.12}$$

which implies

$$h^{-1}_\alpha\left(H^c_\alpha(E \mid \hat{\boldsymbol{Q}}^-)\right) \leq P^c_{e,\min}(\hat{\boldsymbol{Q}}^-) = \mathbb{E}\left[I(\hat{\boldsymbol{Q}}^-)\right]. \tag{A.13}$$

On the other hand, given the binary entropy function $h_\alpha(p)$ for $p \in [0, 0.5]$, $\alpha \in (0, 2]$, it is straight forward to show that

$$h_\alpha\left(I(\hat{\boldsymbol{Q}}^- = \hat{\boldsymbol{q}}^-)\right) \geq 2I(\hat{\boldsymbol{Q}}^- = \hat{\boldsymbol{q}}^-). \tag{A.14}$$

Applying the expectation operator to each side of the inequality,

$$H^c_\alpha(E \mid \hat{\boldsymbol{Q}}^-) = \mathbb{E}\left[h_\alpha\left(I(\hat{\boldsymbol{Q}}^-)\right)\right] \geq 2\mathbb{E}\left[I(\hat{\boldsymbol{Q}}^-)\right] = 2P^c_{e,\min}(\hat{\boldsymbol{Q}}^-). \tag{A.15}$$

In conclusion,

$$h^{-1}_\alpha\left(H^c_\alpha(E \mid \hat{\boldsymbol{Q}}^-)\right) \leq P^c_{e,\min}(\hat{\boldsymbol{Q}}^-) = \mathbb{E}\left[I(\hat{\boldsymbol{Q}}^-)\right] \leq \min\left\{\frac{1}{2}H^c_\alpha(E \mid \hat{\boldsymbol{Q}}^-), C\right\} \tag{A.16}$$

□

Appendix B. Proof of inequality of minimum probability of error for additional time lags

A consequence of incorporating additional time lags into the vector observable variables is that

$$\begin{aligned} P^c_{e,\min,\mathrm{LB}}(\hat{\boldsymbol{Q}}^{-l}) &\leq P^c_{e,\min,\mathrm{LB}}(\hat{\boldsymbol{Q}}^{-p}), \text{ for } l > p, \\ P^c_{e,\min,\mathrm{UB}}(\hat{\boldsymbol{Q}}^{-l}) &\leq P^c_{e,\min,\mathrm{UB}}(\hat{\boldsymbol{Q}}^{-p}), \text{ for } l > p, \end{aligned} \tag{B.1}$$

where l and p denote the number of time lags in $\hat{\boldsymbol{Q}}^{-l}$ and $\hat{\boldsymbol{Q}}^{-p}$, respectively, i.e.,

$$\begin{aligned} \hat{\boldsymbol{Q}}^{-l} &= [\hat{\boldsymbol{Q}}(t), \hat{\boldsymbol{Q}}(t-\delta t_1), \ldots, \hat{\boldsymbol{Q}}(t-\delta t_l)] \pm \delta\hat{\boldsymbol{Q}}^{-l}, \\ \hat{\boldsymbol{Q}}^{-p} &= [\hat{\boldsymbol{Q}}(t), \hat{\boldsymbol{Q}}(t-\delta t_1), \ldots, \hat{\boldsymbol{Q}}(t-\delta t_p)] \pm \delta\hat{\boldsymbol{Q}}^{-p}. \end{aligned} \tag{B.2}$$

Proof. First, we prove the cost-sensitive conditional entropy inequality,

$$H^c_\alpha(E \mid \hat{\boldsymbol{Q}}^{-l}) \leq H^c_\alpha(E \mid \hat{\boldsymbol{Q}}^{-p}). \tag{B.3}$$

Noting that $\hat{\boldsymbol{Q}}^{-l} = [\hat{\boldsymbol{Q}}^{-p}, \hat{\boldsymbol{Q}}^{-r}]$, where $\hat{\boldsymbol{Q}}^{-r} = [\hat{\boldsymbol{Q}}(t-\delta t_{p+1}), \ldots, \hat{\boldsymbol{Q}}(t-\delta t_l)] \pm \delta\hat{\boldsymbol{Q}}^{-r}$, proving Eq (B.3) is equivalent to proving

$$H^c_\alpha(E \mid \hat{\boldsymbol{Q}}^{-p},\ \hat{\boldsymbol{Q}}^{-r}) \leq H^c_\alpha(E \mid \hat{\boldsymbol{Q}}^{-p}). \tag{B.4}$$

By the law of total probability applied to the conditional probability:

$$\begin{aligned} &P(E = 1 \mid \hat{\boldsymbol{Q}}^{-p} = \hat{\boldsymbol{q}}^{-p}) \\ &= \sum_{\hat{\boldsymbol{q}}^{-r}} P(E = 1 \mid \hat{\boldsymbol{Q}}^{-p} = \hat{\boldsymbol{q}}^{-p},\ \hat{\boldsymbol{Q}}^{-r} = \hat{\boldsymbol{q}}^{-r}) P(\hat{\boldsymbol{Q}}^{-r} = \hat{\boldsymbol{q}}^{-r} \mid \hat{\boldsymbol{Q}}^{-p} = \hat{\boldsymbol{q}}^{-p}). \end{aligned} \tag{B.5}$$

The cost-sensitive entropy function h^c_α is a concave function when $\alpha \in (0, 2]$, applying Jensen's inequality,

$$\begin{aligned} &h^c_\alpha\left(P(E = 1 \mid \hat{\boldsymbol{Q}}^{-p} = \hat{\boldsymbol{q}}^{-p})\right) \\ &\overset{(B.5)}{=} h^c_\alpha\left(\sum_{\hat{\boldsymbol{q}}^{-r}} P(E = 1 \mid \hat{\boldsymbol{Q}}^{-p} = \hat{\boldsymbol{q}}^{-p},\ \hat{\boldsymbol{Q}}^{-r} = \hat{\boldsymbol{q}}^{-r}) \right. \\ &\quad \left. \times\ P(\hat{\boldsymbol{Q}}^{-r} = \hat{\boldsymbol{q}}^{-r} \mid \hat{\boldsymbol{Q}}^{-p} = \hat{\boldsymbol{q}}^{-p})\right) \\ &\geq \sum_{\hat{\boldsymbol{q}}^{-r}} h^c_\alpha\left(P(E = 1 \mid \hat{\boldsymbol{Q}}^{-p} = \hat{\boldsymbol{q}}^{-p},\ \hat{\boldsymbol{Q}}^{-r} = \hat{\boldsymbol{q}}^{-r})\right) P(\hat{\boldsymbol{Q}}^{-r} = \hat{\boldsymbol{q}}^{-r} \mid \hat{\boldsymbol{Q}}^{-p} = \hat{\boldsymbol{q}}^{-p}). \end{aligned} \tag{B.6}$$

Applying the inequality to the right hand side of Eq. (B.3), we get

$$\begin{aligned} &H^c_\alpha(E \mid \hat{\boldsymbol{Q}}^{-p}) \\ &= \sum_{\hat{\boldsymbol{q}}^{-p}} h^c_\alpha\left(P(E = 1 \mid \hat{\boldsymbol{Q}}^{-p} = \hat{\boldsymbol{q}}^{-p})\right) P(\hat{\boldsymbol{Q}}^{-p} = \hat{\boldsymbol{q}}^{-p}) \\ &\overset{(B.6)}{\geq} \sum_{\hat{\boldsymbol{q}}^{-p}} \sum_{\hat{\boldsymbol{q}}^{-r}} h^c_\alpha\left(P(E = 1 \mid \hat{\boldsymbol{Q}}^{-p} = \hat{\boldsymbol{q}}^{-p},\ \hat{\boldsymbol{Q}}^{-r} = \hat{\boldsymbol{q}}^{-r})\right) \\ &\quad P(\hat{\boldsymbol{Q}}^{-r} = \hat{\boldsymbol{q}}^{-r} \mid \hat{\boldsymbol{Q}}^{-p} = \hat{\boldsymbol{q}}^{-p}) P(\hat{\boldsymbol{Q}}^{-p} = \hat{\boldsymbol{q}}^{-p}) \\ &= \sum_{\hat{\boldsymbol{q}}^{-p}} \sum_{\hat{\boldsymbol{q}}^{-r}} h^c_\alpha\left(P(E = 1 \mid \hat{\boldsymbol{Q}}^{-p} = \hat{\boldsymbol{q}}^{-p},\ \hat{\boldsymbol{Q}}^{-r} = \hat{\boldsymbol{q}}^{-r})\right) \\ &\quad \times\ P(\hat{\boldsymbol{Q}}^{-r} = \hat{\boldsymbol{q}}^{-r},\ \hat{\boldsymbol{Q}}^{-p} = \hat{\boldsymbol{q}}^{-p}) \\ &= H^c_\alpha(E \mid \hat{\boldsymbol{Q}}^{-p}, \hat{\boldsymbol{Q}}^{-r}) = H^c_\alpha(E \mid \hat{\boldsymbol{Q}}^{-l}). \end{aligned} \tag{B.7}$$

Finally, the conditional entropy inequality is applied to the information theoretic-bounds

$$\begin{aligned} P^c_{e,\min,\mathrm{LB}}(\hat{\boldsymbol{Q}}^{-l}) &= h^{-1}_\alpha\left(H^c_\alpha(E \mid \hat{\boldsymbol{Q}}^{-l})\right), \\ P^c_{e,\min,\mathrm{LB}}(\hat{\boldsymbol{Q}}^{-p}) &= h^{-1}_\alpha\left(H^c_\alpha(E \mid \hat{\boldsymbol{Q}}^{-p})\right), \\ P^c_{e,\min,\mathrm{UB}}(\hat{\boldsymbol{Q}}^{-l}) &= \min\left\{\frac{1}{2}H^c_\alpha(E \mid \hat{\boldsymbol{Q}}^{-l}), C\right\}, \\ P^c_{e,\min,\mathrm{UB}}(\hat{\boldsymbol{Q}}^{-p}) &= \min\left\{\frac{1}{2}H^c_\alpha(E \mid \hat{\boldsymbol{Q}}^{-p}), C\right\}, \end{aligned} \tag{B.8}$$

by taking into account that $P^c_{e,\min}(\hat{\boldsymbol{Q}}^{-p}) \leq 1/2$,

$$\begin{aligned} P^c_{e,\min,\mathrm{LB}}(\hat{\boldsymbol{Q}}^{-l}) &\leq P^c_{e,\min,\mathrm{LB}}(\hat{\boldsymbol{Q}}^{-p}), \text{ for } l > p, \\ P^c_{e,\min,\mathrm{UB}}(\hat{\boldsymbol{Q}}^{-l}) &\leq P^c_{e,\min,\mathrm{UB}}(\hat{\boldsymbol{Q}}^{-p}), \text{ for } l > p, \end{aligned} \tag{B.9}$$

□

Appendix C. Continuous extension of information-theoretic bound

We show that the bound for minimum cost-sensitive probability of error also holds for continuous observables just by replacing the cost-sensitive Rényi entropy H^c_α by the cost-sensitive Rényi entropy $\mathcal{H}^c_\alpha$ conditioned on a continuous variable. The minimum cost-sensitive probability of error achievable by any model with continuous $\hat{\boldsymbol{Q}}^-$ is

$$\begin{aligned} &h^{-1}_\alpha\left(\mathcal{H}^c_\alpha(E \mid \hat{\boldsymbol{Q}}^-)\right) \leq P^c_{e,\min}(\hat{\boldsymbol{Q}}^-) \\ &= \mathbb{E}[I(\hat{\boldsymbol{Q}}^-)] \leq \min\left\{\frac{1}{2}\left(\mathcal{H}^c_\alpha(E \mid \hat{\boldsymbol{Q}}^-)\right), C\right\}. \end{aligned} \tag{C.1}$$

The cost-sensitive conditional Rényi entropy is defined as

$$\mathcal{H}^c_\alpha(E \mid \hat{\boldsymbol{Q}}^-) = \int_{\hat{\boldsymbol{Q}}^-} h^c_\alpha\left(P(E = 1 \mid \hat{\boldsymbol{Q}}^- = \hat{\boldsymbol{q}}^-)\right) \rho_{\hat{\boldsymbol{Q}}^-}(\hat{\boldsymbol{q}}^-) \mathrm{d}\hat{\boldsymbol{q}}^-, \tag{C.2}$$

where $\rho_{\hat{\boldsymbol{Q}}^-}(\hat{\boldsymbol{q}}^-)$ is the probability density function of $\hat{\boldsymbol{Q}}^-$ and the integral is taken over the support of $\hat{\boldsymbol{Q}}^-$. The minimum probability of error at each value is defined as

$$\begin{aligned} &\mathcal{I}(\hat{\boldsymbol{Q}}^- = \hat{\boldsymbol{q}}^-) \\ &= \min\left\{c^- P(E = 1 \mid \hat{\boldsymbol{Q}}^- = \hat{\boldsymbol{q}}^-), c^+\left(1 - P(E = 1 \mid \hat{\boldsymbol{Q}}^- = \hat{\boldsymbol{q}}^-)\right)\right\}. \end{aligned} \tag{C.3}$$

Proof. Similar to Eq. (A.3), the overall cost-sensitive probability of error is a weighted integral of the probability of error for each value of

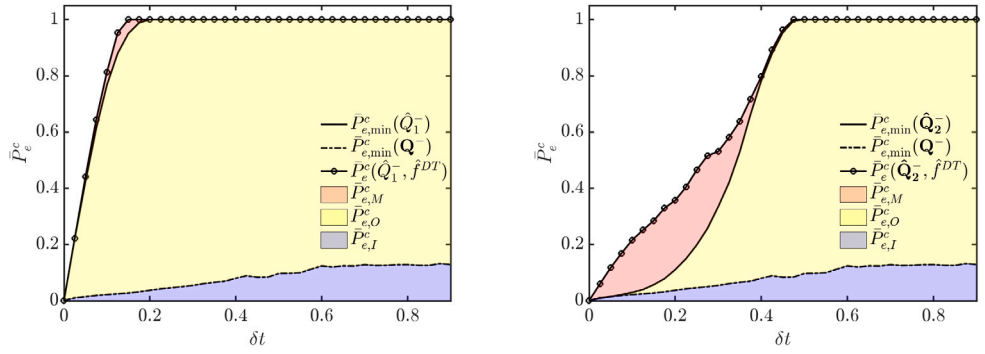

Fig. D.7. Normalized probability of error for extreme events prediction in the Rössler system for the threshold $\eta = \bar{\theta}_3 + 6\sigma_{\theta_3}$ using observable (a) $\hat{Q}_1^-$ and (b) $\hat{\mathbf{Q}}_2^-$. $\bar{P}^c_{e,\min}(\hat{Q}_1^-)$, $\bar{P}^c_{e,\min}(\hat{\mathbf{Q}}_2^-)$, and $\bar{P}^c_{e,\min}(\mathbf{Q}^-)$ are the minimum probability of error using the observable $\hat{Q}_1^-$, $\hat{\mathbf{Q}}_2^-$, and $\mathbf{Q}^-$, respectively.

$\hat{\mathbf{Q}}^-$,

$$P^c_e(\hat{\mathbf{Q}}^-, \hat{f}) = \int_{\hat{\mathbf{Q}}^-} \Big(c^- P(\hat{E}=0, E=1 \mid \hat{\mathbf{Q}}^- = \hat{\mathbf{q}}^-) + c^+ P(\hat{E}=1, E=0 \mid \hat{\mathbf{Q}}^- = \hat{\mathbf{q}}^-) \Big) \rho_{\hat{\mathbf{Q}}^-}(\hat{\mathbf{q}}^-) \mathrm{d}\hat{\mathbf{q}}^-, \tag{C.4}$$

Applying Eqs. (C.4) and (A.5) we have,

$$P^c_{e,\min}(\hat{\mathbf{Q}}^-) = \int_{\hat{\mathbf{Q}}^-} \mathcal{I}(\hat{\mathbf{Q}}^- = \hat{\mathbf{q}}^-) \rho_{\hat{\mathbf{Q}}^-}(\hat{\mathbf{q}}^-) \mathrm{d}\hat{\mathbf{q}}^- = \mathbb{E}[\mathcal{I}(\hat{\mathbf{Q}}^-)]. \tag{C.5}$$

Analogous to Eq. (A.12),

$$\mathbb{E}\left[h_\alpha\left(\mathcal{I}(\hat{\mathbf{Q}}^-)\right)\right] \le h_\alpha\left(\mathbb{E}[\mathcal{I}(\hat{\mathbf{Q}}^-)]\right) = h_\alpha\left(P^c_{e,\min}(\hat{\mathbf{Q}}^-)\right), \tag{C.6}$$

where

$$\begin{aligned}\mathbb{E}\left[h_\alpha\left(I(\hat{\mathbf{Q}}^-)\right)\right] &= \int_{\hat{\mathbf{Q}}^-} h_\alpha\left(I(\hat{\mathbf{q}}^-)\right) \rho_{\hat{\mathbf{Q}}^-}(\hat{\mathbf{q}}^-) \mathrm{d}\hat{\mathbf{q}}^- \\ &= \int_{\hat{\mathbf{Q}}^-} h^c_\alpha\left(P(E=1 \mid \hat{\mathbf{Q}}^- = \hat{\mathbf{q}}^-)\right) \rho_{\hat{\mathbf{Q}}^-}(\hat{\mathbf{q}}^-) \mathrm{d}\hat{\mathbf{q}}^- \\ &= \mathcal{H}^c_\alpha(E \mid \hat{\mathbf{Q}}^-).\end{aligned} \tag{C.7}$$

On the other hand, similar to Eq. (A.15),

$$\mathcal{H}^c_\alpha(E \mid \hat{\mathbf{Q}}^-) = \mathbb{E}\left[h_\alpha\left(\mathcal{I}(\hat{\mathbf{Q}}^-)\right)\right] \ge 2\mathbb{E}\left[\mathcal{I}(\hat{\mathbf{Q}}^-)\right] = 2P^c_{e,\min}(\hat{\mathbf{Q}}^-). \tag{C.8}$$

In conclusion,

$$\begin{aligned}&h_\alpha^{-1}\left(\mathcal{H}^c_\alpha(E \mid \hat{\mathbf{Q}}^-)\right) \le P^c_{e,\min}(\hat{\mathbf{Q}}^-) \\ &= \mathbb{E}\left[\mathcal{I}(\hat{\mathbf{Q}}^-)\right] \le \min\left\{\frac{1}{2}\left(\mathcal{H}^c_\alpha(E \mid \hat{\mathbf{Q}}^-)\right), C\right\}.\end{aligned} \tag{C.9}$$

□

Appendix D. Results for higher thresholds and confusion matrix for $\hat{f}^{DT}$

To address the sensitivity of the results to the intensity of the extreme events in the Rössler system and the Kolmogorov flow, we repeated the analysis with higher threshold values in both cases. The minimum probability of error is shown in Figs. D.7 and D.8 as a function of δt for the cases discussed above. The main observation is that increasing the threshold for extreme events makes the prediction more challenging, due to the higher scarcity of events, which renders them more unpredictable. Nonetheless, the trends discussed in the main text regarding the behavior of δt remain unchanged.

The confusion matrix for the decision tree model to predict extreme events in the Rössler system using observable $\hat{Q}_1^-$ is shown in Table D.1 for $\delta t = 0.075$.

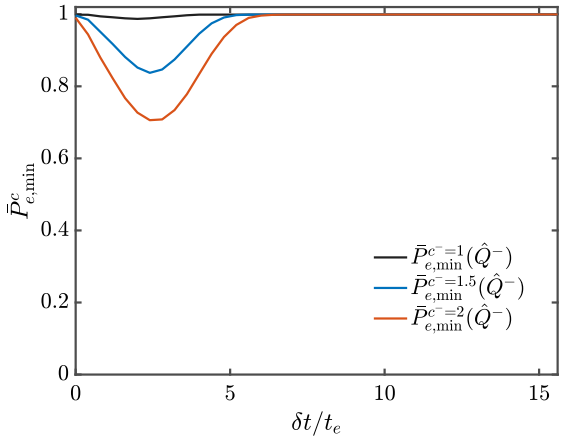

Fig. D.8. Normalized, cost-sensitive minimum probability of error for extreme event prediction in the Kolmogorov flow for $\eta = \bar{D} + 3\sigma_D$.

Table D.1
Confusion matrix of decision tree model. The normalized probability of error of the model can be calculated as $\bar{P}^c_e(\hat{Q}_1^-, \hat{f}^{DT}; \delta t = 0.075) = \frac{0.0060+0.0030}{0.0030+0.0194} = 0.3554..$

Decision tree	$\hat{E}=0$	$\hat{E}=1$
$E=0$	0.9716	0.0060
$E=1$	0.0030	0.0194

References

[1] S. Albeverio, V. Jentsch, H. Kantz, D. Dragoman, M. Dragoman, A.C. Elitzur, M.P. Silverman, J. Tuszynski, H.D. Zeh (Eds.), Extreme events in nature and society, The Frontiers Collection, Springer, Berlin, Heidelberg, 2006.
[2] B. Etkin, Turbulent wind and its effect on flight, J. Aircr. 18 (1981) 327.
[3] S. Rahmstorf, D. Coumou, Increase of extreme events in a warming world, Proc. Natl. Acad. Sci. USA 108 (2011) 17905.
[4] B. Schäfer, D. Witthaut, M. Timme, V. Latora, Dynamically induced cascading failures in power grids, Nature Commun. 9 (2018) 1975.
[5] A. Ray, S. Rakshit, G.K. Basak, S.K. Dana, D. Ghosh, Understanding the origin of extreme events in El Niño southern oscillation, Phys. Rev. E 101 (6) (2020) 062210.
[6] A. Ray, T. Bröhl, A. Mishra, S. Ghosh, D. Ghosh, T. Kapitaniak, S.K. Dana, C. Hens, Extreme events in a complex network: Interplay between degree distribution and repulsive interaction, Chaos 32 (12) (2022).
[7] A. Kumar, S. Kulkarni, M.S. Santhanam, Extreme events in stochastic transport on networks, Chaos 30 (2020) 043111.
[8] M. Farazmand, T.P. Sapsis, Extreme events: Mechanisms and prediction, Appl. Mech. Rev. 71 (2019).
[9] T.P. Sapsis, Statistics of extreme events in fluid flows and waves, Annu. Rev. Fluid Mech. 53 (2021) 85.
[10] S.E. Otto, C.W. Rowley, Koopman operators for estimation and control of dynamical systems, Annu. Rev. Control Robot. Auton. Syst. 4 (2021) 59.

[11] A. Asch, E.J. Brady, H. Gallardo, J. Hood, B. Chu, M. Farazmand, Model-assisted deep learning of rare extreme events from partial observations, Chaos 32 (2022) 043112.
[12] S. Mukherjee, E. Osuna, F. Girosi, Nonlinear prediction of chaotic time series using support vector machines, in: Proc. IEEE Signal Processing Society Workshop, IEEE, Amelia Island, FL, USA, 1997, pp. 511–520.
[13] M. Ghil, M.R. Allen, M.D. Dettinger, K. Ide, D. Kondrashov, M.E. Mann, A.W. Robertson, A. Saunders, Y. Tian, F. Varadi, P. Yiou, Advanced spectral methods for climatic time series, Rev. Geophys. 40 (3) (2002).
[14] E. Racah, C. Beckham, T. Maharaj, S.E. Kahou, M. Prabhat, C. Pal, ExtremeWeather: A large-scale climate dataset for semi-supervised detection, localization, and understanding of extreme weather events, in: Adv. Neural Inf. Process. Syst, Vol. 30, Curran Associates, Inc., 2017.
[15] P.R. Vlachas, W. Byeon, Z.Y. Wan, T.P. Sapsis, P. Koumoutsakos, Data-driven forecasting of high-dimensional chaotic systems with long short-term memory networks, Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci. 474 (2018) 20170844.
[16] V. Pammi, M. Clerc, S. Coulibaly, S. Barbay, Extreme events prediction from nonlocal partial information in a spatiotemporally chaotic microcavity laser, Phys. Rev. Lett. 130 (2023) 223801.
[17] A. Lozano-Durán, G. Arranz, Information-theoretic formulation of dynamical systems: Causality, modeling, and control, Phys. Rev. Res. 4 (2022) 023195.
[18] L.A. Smith, Disentangling uncertainty and error: On the predictability of nonlinear systems, in: A.I. Mees (Ed.), Nonlinear Dynamics and Statistics, Birkhäuser, Boston, MA, 2001, pp. 31–64.
[19] E.N. Lorenz, Deterministic nonperiodic flow, J. Atmos. Sci. 20 (1963) 130.
[20] L. Paninski, Estimation of entropy and mutual information, Neural Comput. 15 (6) (2003) 1191–1253.
[21] T. Hastie, R. Tibshirani, J.H. Friedman, J.H. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Vol. 2, springer, 2009.
[22] I. Guyon, A. Elisseeff, An introduction to variable and feature selection, J. Mach. Learn. Res. 3 (Mar) (2003) 1157–1182.
[23] T.M. Cover, J.A. Thomas, Elements of Information Theory, second ed., Wiley, 2006.
[24] R.M. Fano, D. Hawkins, Transmission of information: A statistical theory of communications, Am. J. Phys. 29 (1961) 793.
[25] M. Hellman, J. Raviv, Probability of error, equivocation, and the chernoff bound, IEEE Trans. Inf. Theory. 16 (4) (1970) 368–372.
[26] A. Rényi, On measures of entropy and information, in: Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics, Vol. 4, University of California Press, 1961, pp. 547–562.
[27] M.-J. Zhao, N. Edakunni, A. Pocock, G. Brown, Beyond Fano's inequality: Bounds on the optimal F-score, BER, and cost-sensitive risk and their implications, J. Mach. Learn. Res. 14 (2013) 1033.
[28] C.E. Shannon, A mathematical theory of communication, Bell Syst. Tech. J. 27 (1948) 379.
[29] F. Takens, Detecting strange attractors in turbulence, in: D. Rand, L.-S. Young (Eds.), Dynamical Systems and Turbulence, Warwick 1980, Springer Berlin Heidelberg, Berlin, Heidelberg, ISBN: 978-3-540-38945-3, 1981, pp. 366–381.
[30] M. Farazmand, T.P. Sapsis, A variational approach to probing extreme events in turbulent dynamical systems, Sci. Adv. 3 (2017) e1701533.
[31] V. Arnold, L. Meshalkin, Seminar led by AN Kolmogorov on selected problems of analysis (1958–1959), Usp. Mat. Nauk 15 (1960) 20.
[32] B.W. Zeff, D.D. Lanterman, R. McAllister, R. Roy, E.J. Kostelich, D.P. Lathrop, Measuring intense rotation and dissipation in turbulent flows, Nature 421 (2003) 146.
[33] R.O. Duda, P.E. Hart, et al., Pattern Classification and Scene Analysis, Vol. 3, Wiley New York, 1973.
[34] M. Ben-Bassat, J. Raviv, Rényi's entropy and the probability of error, IEEE Trans. Inform. Theory 24 (3) (1978) 324–331.