Improved predictions of rare events using the Crooks fluctuation theorem

Julia Gundermann,[1,*] Stefan Siegert,[2] and Holger Kantz[1]
[1]*Max-Planck-Institut für Physik Komplexer Systeme, Nöthnitzer Strasse 38, 01187 Dresden, Germany*
[2]*College of Engineering, Mathematics and Physical Sciences, University of Exeter, Harrison Building, North Park Road, Exeter, EX4 4QF, United Kingdom*
(Received 17 December 2013; published 12 March 2014)

This article explores the applicability of concepts from nonequilibrium thermodynamics to rare event prediction. The Crooks fluctuation theorem is an equality constraint on the probability distribution of a thermodynamical observable. We consider as a prediction target the exceedance of a threshold of such an observable, where the magnitude of the threshold modulates the rareness of the event. A probability forecast is constructed for this event based on a small observational data set. A simple method is proposed that exploits the Crooks fluctuation theorem to estimate this probability. It is shown that this estimator has improved predictive skill compared to the relative frequency of exceedance in the data set. We quantify this improvement in two examples, and study its dependence on the threshold magnitude and sample size in different systems. Further improvements are achieved by combining the Crooks estimator with the frequency estimator.

DOI: 10.1103/PhysRevE.89.032112 PACS number(s): 05.70.Ln, 92.60.Wc, 05.40.−a

I. INTRODUCTION

A. Motivation

Our highly structured and interconnected civilization is quite vulnerable to a large diversity of rare but extreme events. Weather extremes are the most evident threat, but also earthquakes and other geophysical hazards have a much bigger impact on human life when taking place in a large city than when taking place in the sparsely populated countryside. In addition, communication networks [1], traffic networks, and the power grid [2] are such that local failure might have long range and hence large scale effect. For a cost efficient protection against natural hazards or other types of extreme events, a precise knowledge of their frequency of occurrence as a function of the magnitude of an event is needed. For example, U. S. and German flood protection plans only consider events up to magnitudes which occur at least once per 100 years on average, so-called 100-year-floods [3]. In other terminology, such a flood has a chance of 1% to occur within a given year and is called to have a *return level* or *return period* of 100 years. While it is a political and societal issue to decide which such return level is tolerable, it is a scientific issue to determine the return levels, i.e., the average recurrence time, as a function of the magnitude of a given event class.

It is evident that really extreme events are rare; they are in the tail of the distribution. Hence, the average recurrence time of events larger than a certain threshold τ is the inverse of the weight of the tail of the underlying magnitude distribution integrated over magnitudes larger than τ. Since data samples are finite and most often small (not many really extreme events have taken place within the observation period), it is a challenge to precisely estimate the weight of the tail of the underlying distribution. Furthermore, in the context of climate change there are hints that the frequency of extreme weather conditions might increase [4]. A statistical verification of this requires one to consider data on finite time windows (e.g., 10 years) in order to prove a time dependence. In this case, the data set for estimation of the tail weight is small by construction. In summary, the estimation of tail weights from finite data samples is a relevant and widespread task. This paper explores a method to improve such estimates from small data samples.

B. From observations to probabilistic predictions

Here, we rephrase the issue of rare events from the point of view of predictions. One of the goals of statistical data analysis is to make predictions of future observations of a physical system. If the future event is uncertain, this uncertainty should be reflected by the forecast. Uncertainty in the future can arise due to inherent randomness (such as thermal noise), limited observability of initial states and unknown parameters (such as in weather and climate forecasting), or simply due to a limited availability of observation data. The probabilistic calculus is useful in this case because it makes possible the quantification of uncertainty. Consider the simple case where a limited number of independent observations of the system of interest is available. A simple prediction problem would be the question: "Will the next observation exceed the value τ?" If the event is uncertain, such a prediction will take the form of a probability. If past observations are available, a simple unbiased estimator of the exceedance probability is k/N, where k is the number of observed exceedance events among the total of N available observations. Hence, observations from the tail of a probability distribution are converted into a probabilistic forecast.

C. Fluctuation theorems

There are cases where not only past observations of the system of interest are available, but also physical constraints. The Crooks fluctuation theorem [5] is a particular example, which relates two probability distributions $p_f(W)$ and $p_b(W)$ of a thermodynamical observable W. The distribution $p_f(W)$ is obtained by driving the system out of equilibrium by applying a certain protocol, and $p_b(W)$ is obtained under the reverse protocol. The Crooks theorem is usually stated as

$$\frac{p_f(W)}{p_b(-W)} = e^{\beta(W-\Delta F)}, \tag{1}$$

*juguma@pks.mpg.de

where W is some mechanical work to be performed on the system, ΔF is the free energy difference between the initial and final macrostates, and $\beta = 1/k_B T$ is the inverse temperature. There are examples where $p_f(W) = p_b(W) =: p(W)$. Since Crooks fluctuation theorem is a constraint on the probability distribution of the observable W, all probabilistic predictions about W (such as exceedance events) should be compatible with it. We will show that this relation offers a different method to estimate the tail weight of probability distributions. The Crooks relation and its usage will be discussed in Sec. II. While the original setting of Crooks, nonequilibrium thermodynamics, is quite far from our practical data analysis problem, we could recently show that the Crooks relation also holds for a process on a two-dimensional turbulent fluid flow [6], which offers potential for a broader applicability of this concept.

D. Evaluating probabilistic predictions

A provider or user of a probabilistic prediction f will be interested in the quality of the forecast [7]. Such a quality assessment should involve a comparison of the prediction f and the actually observed event O, where the occurrence (and nonoccurrence) of the event is coded by $O = 1$ (and $O = 0$). Proper scoring rules are a useful tool for this purpose because they encourage forecasters to communicate their probability assessments honestly, and because they reward desirable properties of probability forecasts, namely, calibration and sharpness [8]. Arguably the most prominent example of proper scoring rules for binary events is the (half) Brier score [9]. We will employ this score to demonstrate the superiority of the Crooks estimate over the direct estimate k/N for the probability of exceedance events. We will introduce the Brier score, and concepts derived from it, in Sec. III.

E. Present study

We consider the problem of predicting the exceedance of a threshold for a variable whose distribution obeys the Crooks relation and for which a small number N of independent observations is available. We propose a simple method to translate the available data and knowledge about Crooks relation into an exceedance probability. The Brier score is used to compare this prediction to predictions that only use the observed data, but also to another benchmark. We present analytical results for the performance of the Crooks estimator and illustrate them for Gaussian distributed random variables. We apply the very same concept to numerically generated data of hydrodynamic flow, where the observable relates to the change of kinetic energy in a transformed two-dimensional fluid field. The corresponding distribution satisfies the Crooks relation, but is not given in closed form. The performance of the Crooks estimator will be determined numerically. Both examples will be elaborated in Sec. IV. The article concludes with a discussion in Sec. V.

F. Terminology

We make frequent use of the words *forecast* and *estimator*. Since we want to perform forecasts on sequences of independent events, the predicted probability is unconditioned. Hence, the forecast problem is identical to the estimation problem of the tail weight of a probability distribution. We can therefore either discuss the performance in terms of skill scores of probability forecasts, or in terms of properties of statistical estimators such as bias, consistency, and variance. The terms forecast and prediction are used interchangeably.

II. GENERAL IDEA

Fluctuation relations are a subject of statistical physics and give deeper insight into systems in a nonequilibrium state [10–15]. They are equalities characterizing the distributions of thermodynamic variables, e.g., work or entropy production. The Crooks relation [5], which is of interest here, is a statement about the distribution of the work $p_f(W)$, which is performed on a system when it is pushed out of equilibrium by changing an external parameter L. This parameter could be, for example, the position of the piston pushed into a volume filled with gas. The probability can be related to the one of the corresponding backward process $p_b(W)$ (pulling the piston outwards) through parameters given by the initial (equilibrium) state of the system, i.e., inverse temperature β and free energy difference ΔF of initial and final states. This relation is called the *Crooks fluctuation theorem* and is given by Eq. (1). It was tested and verified for a number of systems experimentally, e.g., [16–20], which demonstrates its diverse applicability.

There are examples where the forward and backward distributions collapse into the same distribution and the free energy difference is zero. This is the case, for instance, if the process is a thermodynamic cycle and symmetric under time reversal. In the example we will present in Sec. IV B, the equilibrium distribution of the final state is symmetric to that of the initial state. The distribution of work in the forward and backward process is then the same, $p(W)$, which leads to the following relation:

$$\frac{p(W)}{p(-W)} = e^{+\beta W}. \tag{2}$$

Here, we want to address the question of how the fact that such a relation holds for a random observable can be used to improve forecasts of this observable. To be precise: Given a set of training data (i.e., N independent observations of W) and the knowledge that they are drawn from a distribution which satisfies the Crooks relation with a known β, we want to estimate the probability of observing a value $W < \tau$. We can predict this by using two different forecasts:

(1) We count the number k of how often in our sample $W_i < \tau$. We call this approach, which is easy to implement and needs no further information, the *Basis forecast*. Its prediction is

$$f_B = \frac{k}{N}. \tag{3}$$

(2) We use the knowledge that $p(W)$ satisfies Eq. (2). One can show analytically

$$\begin{aligned} P(W < \tau) &= \int_{-\tau}^{\infty} dW\, p(W) e^{-\beta W} \\ &= \mathbb{E}[e^{-\beta W} \Theta(W + \tau)]. \end{aligned} \tag{4}$$

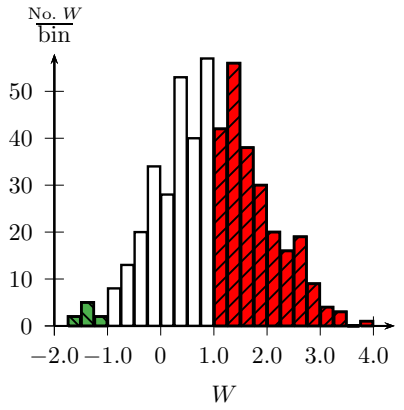

FIG. 1. (Color online) Histogram of data sampled from a distribution, which satisfies the Crooks relation with itself. One may ask for the probability of W to be below -1. This can be approximated by counting over the bins left of this threshold (in green). We call this the Basis forecast [Eq. (3)]. Alternatively, applying the knowledge about the Crooks relation, one can use the Crooks forecast [Eq. (5)], and approximate the probability on the basis of the red-marked part of the distribution. The values contributing to the Basis forecast (in green, striped top left to bottom right) and the Crooks forecast (in red, striped top right to bottom left), are 9 and 238 out of 500, respectively.

Here, $\Theta(x)$ is the Heaviside function, which is 1 for all $x \geqslant 0$ and 0 else, and $\mathbb{E}[X]$ denotes the expectation value of the random variable X. Replacing the analytical expectation value in Eq. (4) by the empirical average leads to the following estimator:

$$f_C = \frac{1}{N}\sum_{i=1}^{N} \Theta(W_i + \tau) e^{-\beta W_i}, \tag{5}$$

which we will name the *Crooks forecast* in the following.

We propose that the Crooks forecast offers an improvement compared to the Basis forecast, particularly in the region $\tau\beta < 0$. Intuitively, the Crooks forecast f_C is a better estimator in this region simply because it is calculated based on more data. Figure 1 visualizes the idea. In this article, we want to make the notion of "better" more explicit, and quantify how much better the Crooks forecast is under different settings. In a typical thermodynamic setup, where β refers to an inverse temperature and thus is positive, f_C should be better than f_B for negative thresholds. We will consider the case of positive β here for all our derivations, except in Sec. IV B where we used a negative β. We will see that for $\beta < 0$ one can predict exceedances of positive thresholds $\tau > 0$ in the same spirit. We further show that the Crooks forecast does not improve compared to the Basis forecast on the full range of thresholds. The reason is that the empirical mean of $e^{-\beta W}$ is heavily influenced by strongly weighted rare events from the diverging side of the exponential function, and thus is sensitive to whether such values are included in the sample or not.

III. FORECAST EVALUATION: BRIER SCORE, RELIABILITY, AND BRIER SKILL SCORE

Assessing the quality of a probability forecast or estimator requires a quantitative criterion that compares the forecast $f \in [0,1]$ to the corresponding binary observation $O \in \{0,1\}$. For this study, we will use the quadratic criterion introduced by Brier [9], given by

$$b = (f - O)^2. \tag{6}$$

This criterion is referred to as the half Brier score or simply *Brier score*. It is always non-negative and it is the smaller the better the forecast is. The perfect forecast, which predicts $f = 1$ whenever the event happens, and $f = 0$ otherwise, has a Brier Score of $b = 0$.

The Brier score has become a very popular scoring criterion for several reasons: The Brier score is *proper*, which means that a forecaster can not improve their score by issuing a forecast which is different from their subjective probability for the event O. This property of encouraging forecasters to be honest is the one originally advocated by Brier. The Brier score further rewards favorable attributes of probability forecasts (reliability and resolution [8,21]). The mean of the Brier score over a number of forecast-verification pairs is a mean squared error between forecasts and observations, which makes it interpretable in terms of estimator performance. Lastly, we are going to use the Brier score out of mere convenience because many of the calculations in this paper can be done analytically using the Brier score. There are further proper scoring rules, such as the Ignorance score $-\log_2|1 - O - f|$ [22], for which this is not the case. Being a logarithmic criterion, the latter penalizes the occurrence of an event missed by the forecaster more severely than the Brier score. Due to limited data, the Basis forecast $f_B = k/N$ can be zero, which causes the Ignorance score to diverge if the event does happen.

The average Brier score, which we will denote by B, can be decomposed additively into three components called *reliability, resolution*, and *uncertainty*,

$$B := \mathbb{E}[b(f,O)] = \mathrm{REL} - \mathrm{RES} + \mathrm{UNC}, \tag{7}$$

which was shown first for the empirical average by [21], and for the mathematical expectation by [23]. Define the *climatology* as $\Pi = \mathbb{E}[O]$ and the *calibration function* as $\pi(f) = P(O = 1|f)$. Then, the three terms are given by

$$\mathrm{REL} = \mathbb{E}[(f - \pi(f))^2], \tag{8a}$$

$$\mathrm{RES} = \mathbb{E}[(\Pi - \pi(f))^2], \text{ and} \tag{8b}$$

$$\mathrm{UNC} = \Pi(1 - \Pi). \tag{8c}$$

Reliability quantifies the agreement between the forecast f and the conditional frequency of occurrence $P(O = 1|f)$. Resolution, on the other hand, rewards meaningful variations of the forecast f from the limiting frequency of occurrence of the event. In systems where event probabilities change over time (as, for example, the probability of rain for weather forecasting), resolution is important, as it quantifies the knowledge of a forecaster about state-dependent changes. In our case, the "true" event probability $P(W < \tau)$ is not going to change, thus there can not be any meaningful variations of f around Π, and our forecasts can not have any resolution. Since $P(O = 1|f) = P(O = 1) = \Pi$, any variability in f is just noise. We will thus focus on the reliability of the forecast f as a measure of predictive skill.

We are interested in asymptotic properties of our probability estimators f_B and f_C. We would like to know which forecast

TABLE I. Overview over the different forecast's reliabilities and Brier scores.

Forecast	Reliability (REL)	Brier score (B)
Basis: $f = \frac{1}{N}\sum_{i=1}^{N}\Theta(\tau - W_i)$	$\frac{\Pi - \Pi^2}{N}$	$\frac{N+1}{N}\Pi(1-\Pi)$
Crooks: $f = \frac{1}{N}\sum_{i=1}^{N} e^{-\beta W_i}\Theta(W_i - \tau)$	$\frac{\mathcal{E} - \Pi^2}{N}$	$\frac{N+1}{N}\Pi(1-\Pi) - \frac{\Pi - \mathcal{E}}{N}$
Zero: $f = 0$	Π^2	Π

can be trusted more, i.e., which forecast is on average closer to the true event probability Π under the distribution $p(W)$. Given a specific value of the estimator, one way to assess this difference would be to calculate average Brier scores of this estimator over a large number of test cases. Another, much more convenient way is to set up an experiment with artificial data for which we know the true event probability which f_B and f_C then try to estimate. For a binary event O, whose true probability of occurrence is Π, the conditional expected Brier score, given an estimator f, is given by

$$\begin{aligned} B &= \mathbb{E}[(O-f)^2|f] \\ &= \Pi(1-f)^2 + (1-\Pi)f^2 \\ &= \underbrace{(f-\Pi)^2}_{\text{REL}} + \underbrace{\Pi(1-\Pi)}_{\text{UNC}}, \end{aligned} \tag{9}$$

which was decomposed into its reliability and uncertainty behind the last equals sign. The uncertainty of a forecast is equal to the Brier score of the climatology (which predicts the exact probability $f = \Pi$). The Brier score of the Basis forecast can be calculated analytically as the expectation value of Eq. (9) over many realizations f (see [24] for a general examination of this issue). These are discrete and restricted to k/N. The random variable k is binomially distributed with size N and probability Π. Using this, one can calculate the expectation value of f^2 by taking the expectation of Eq. (9) over f, and the expected Brier score for the Basis forecast becomes

$$B_B = \frac{N+1}{N}\Pi(1-\Pi) = \text{UNC} + \frac{1}{N}\Pi(1-\Pi). \tag{10}$$

The reliability term for the Crooks forecast, and thus its Brier score due to Eq. (9), can also be calculated analytically:

$$\begin{aligned} \text{REL}_C(\tau,N) &= \mathbb{E}\left[\left(\frac{1}{N}\sum_{i=1}^{N} e^{-\beta W_i}\Theta(W_i+\tau) - \Pi(\tau)\right)^2\right] \\ &= \frac{1}{N^2}\sum_{i=1}^{N}\mathbb{E}\left[e^{-2\beta W_i}\Theta(W_i+\tau)\right] + \frac{1}{N^2}\sum_{i\neq j=1}^{N}\mathbb{E}\left[e^{-\beta W_i}\Theta(W_i+\tau)e^{-\beta W_j}\Theta(W_j+\tau)\right] - \Pi^2 \\ &= \frac{N}{N^2}\mathbb{E}[e^{\beta W}\Theta(\tau - W)] + \frac{N(N-1)}{N^2}\Pi^2 - \Pi^2 \\ &= \frac{\mathcal{E} - \Pi^2}{N}. \end{aligned} \tag{11}$$

Here, we used the Crooks relation (2) to get from second to third line, and abbreviated the expectation value of $e^{\beta W}$ over $W < \tau$ with $\mathcal{E}$,

$$\mathcal{E} = \mathbb{E}[e^{\beta W}\Theta(\tau - W)]. \tag{12}$$

Note that both the integrand and domain of integration differ from those used to construct the Crooks forecast [see Eq. (4)].

As a third and very simple forecasting method, we introduce the *Zero forecast*. It unconditionally predicts zero probability, independent of the data set. The Zero forecast serves as a good benchmark for events that happen with low probability. Its Brier score is $B_Z = \Pi$, with reliability $\text{REL} = \Pi^2$.

See Table I for an overview of the information we collected so far, from which we can already draw the following conclusions:

(i) Both the Crooks and Basis reliabilities (and Brier score) are inversely proportional to the training data set size N. Their ratio will be independent of the sample size.

(ii) The expectation value $\mathcal{E}$ will be larger than or equal to Π^2 because the reliability, derived in Eq. (11), is a quadratic quantity.

(iii) The reliability of the Crooks forecast will be smaller than for the Basis forecast if $\mathcal{E} < \Pi$. This is the case for negative thresholds, as can be seen in Fig. 2: the integrand for calculating $\mathcal{E}$ is smaller than the integrand for Π by a factor $e^{\beta\tau}$. Thus, in Fig. 2 in the region of negative thresholds, the red (striped right to left) area is always smaller than the green (striped left to right) one. This means that, for negative thresholds, the reliability of the Crooks forecast is always lower than the reliability of the Basis forecast, and therefore

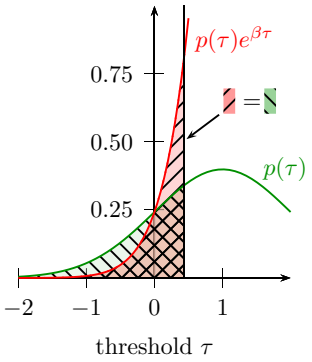

FIG. 2. (Color online) Visualization of the curves used to calculate the expectation value $\mathcal{E}(\tau) = \int_{-\infty}^{\tau} d\tau' p(\tau') e^{\beta\tau'}$ (in red, striped top right to bottom left) and the cumulative probability $\Pi = \int_{-\infty}^{\tau} d\tau' p(\tau)'$ (in green, striped top left to bottom right). For negative thresholds $\mathcal{E}$ will be lower than Π. For positive thresholds, this difference gets compensated by the now expanding effect of the exponential weight. At a certain τ (depending on the exact distribution) $\mathcal{E} = \Pi$. At this point, the Crooks forecast becomes worse than the Basis forecast.

its Brier score is always lower, i.e., better. With this finding, we confirm our main hypothesis.

(iv) There will be a value $\tau > 0$, depending on the distribution, but independent of N, where the Basis forecast becomes better than the Crooks forecast. The reason is that the effect explained in the item above happens for positive thresholds in the opposite direction and at a certain τ both parts compensate each other (see Fig. 2). The behavior of the forecasts on the range of positive thresholds is addressed further at the end of Sec. IV A.

(v) The comparison of Crooks and Basis forecasts was easy due to the fact that they both depend in the same way on the sample size N. However, it is difficult to make general statements for the comparison of the Zero forecast versus Basis, respectively, Crooks. The following can be said: All forecasts have zero reliability for $\Pi = 0$ and a reliability greater than zero otherwise. For all finite τ (i.e., $\Pi > 0$), there is a minimal sample size $N_C(\tau)$, for which the Crooks reliability is smaller than the Zero forecast reliability. Similarly, there is a minimal sample size N_B, for which the Basis reliability is below the Zero reliability. The same holds for the scores. The exact $N_C(\tau)$ and $N_B(\tau)$ depend on the distribution; a general statement is not possible.

In the following, we use the *Brier skill score* as a normalized measure of forecast skill that is less dependent on the event rate Π [25]. It is defined as

$$S = 1 - \frac{B}{B_{\text{ref}}}, \tag{13}$$

where B_{ref} is the Brier score of a suitable reference prediction. A forecast has a positive Brier skill score if it is better than the reference forecast, and a Brier skill score of one indicates a perfect forecast with Brier score $B = 0$. The choice of the reference forecast will not change the ordering of forecast schemes with respect to their skill. However, whether or not a forecast scheme has positive (or negative) skill can change.

We will choose the forecast scheme predicting the exact probability (or the best approximation from a huge data set, which we suppose to be exact), which we introduced as climatology, as the reference. Forecast schemes without resolution then will always have negative skill. This is the case for all the forecast schemes introduced here. In general, comparing with (well-approximated or exact) climatology negatively biases the skill of any forecast scheme because the former is calculated on the basis of a lot more knowledge (or data) than the latter. This effect is very strong for small sample sizes [26–28]. Correction proposals are available [24,29]. In [24], the authors suggest to use the expectation value of the Brier score of the Basis forecast as the reference. This is considered as a version of climatological score corrected for the effect of the finite sample. Nonetheless, we choose the climatology forecast predicting the exact probability as the reference, to be deliberately independent of the sample size, which makes comparisons between different sample sizes easier, and visualizes better how well the exact probability Π can be forecasted.

From the skill of a forecast compared to climatology we can also easily read its skill compared to the Basis forecast. It is given by $S_{\text{ref}=B} = \frac{N+1}{N} S_{\text{ref}=cl} - \frac{1}{N}$, and thus is only a rescaling of the former. Interpreting forecasts $P(W < \tau)$ as probability estimators for Π we can use methods from statistical estimation theory to evaluate their quality [30]. The quality of estimators can be characterized by their *consistency, bias*, and *variance*. Both Crooks and the Basis forecast are consistent and unbiased, which follows from their definitions. Because they are unbiased, their variance is equal to their Brier score reliability (see Table I). The Zero forecast, however, has zero variance and its bias is equal to Π, independent of the sample size N. Thus, it is not consistent.

To help interpreting the Brier skill score, consider for a moment the Basis estimator for a data set that consists of only one sample point. This estimator is equal to one with probability Π and zero with probability $1 - \Pi$. It is unbiased and its variance is given by $\Pi(1 - \Pi)$. From the definition of the Brier skill score [Eq. (13)], with Π as the reference forecast, and the Brier score decomposition, the Brier skill score of a forecast with zero resolution can be written as

$$S = -\frac{\text{REL}}{\Pi(1 - \Pi)}. \tag{14}$$

We can interpret the denominator as a normalization and read the Brier skill score as the variance of an estimator (with respect to the correct probability) normalized by the variance of the one-point Basis estimator introduced above.

The ordering of forecasts in terms of their Brier skill score is the same as in terms of their Brier score. One just has to keep in mind that (for forecast schemes without resolution and the climatology as the reference, as given here) the optimal skill score is zero, and, the larger the skill score (i.e., the closer to zero), the better is the forecast.

IV. TWO ILLUSTRATING EXAMPLES

In the following, we want to support our findings with illustrative examples. We show two examples, first a Gaussian distribution, and second, data from a numerical experiment,

simulating the transformation of a two-dimensional fluid field. Gaussian distributions with mean μ and variance σ^2 always satisfy the Crooks relation with $\beta = 2\mu/\sigma^2$. Here, we can use the advantage that we have the expectation value $\mathcal{E}$ and the cumulative density Π, needed for the Brier (skill) scores, available in closed form. We complement the analytical results with numerical experiments, if necessary or if it helps illustration. Further, we address the effect of different β to the forecasts.

In the second, physically motivated example, the distribution is highly asymmetric, and not known in closed form. This adds an error in estimating the cumulative density Π because all expectation values must be approximated by empirical mean values. The parameter β is negative in this example, leading to improved predictions in the range of positive thresholds.

A. Gaussian distribution

1. Setup

We consider Gaussian distributions with the same standard deviation $\sigma = 1$, and varying $\beta = 2\mu$ ($1 \leqslant \beta \leqslant 7$), because increasing β moves the body of the distribution towards positive values and changes the expectation value $\mathcal{E}$ and the cumulative density Π for a given threshold τ. Gaussian distributions with different standard deviation can be transformed to $\sigma = 1$ by changing β correspondingly. Further, we investigate the effect of the training data set size N as well as the effect of the calculation of empirically estimated expectation values. The basis of our numerically obtained data are mean values over $n = 10^4$ repetitions of the following, what we call hereafter an "experiment":

(1) Sample N training values from a Gaussian distribution, parametrized by β.

(2) Calculate forecasts f_B and f_C of the probability that $W < \tau$, as given in Table I, first column. (The Zero forecast f_Z is always zero, anyway.)

(3) Calculate Brier scores B_C and B_B, B_Z by inserting the respective f into Eq. (9), using the analytical value of the probability Π.

Then, we calculate empirical mean values of the Brier scores from the $n = 10^4$ experiments, and Brier skill scores using Eq. (13) with the climatology score $\Pi(1 - \Pi)$ as the reference. These empirical averages will complement our analytical results.

Figure 3 shows a histogram of forecasts for a training data set of size $N = 10$. While the distribution of the Crooks forecast is smooth, the Basis forecast is restricted to $N + 1$ discrete values, and the Zero forecast obviously predicts zero in all 10 000 experiments.

2. Crooks vs Basis vs Zero forecast

As mentioned before, for the Gaussian distribution analytical results are available (at least in terms of the error function). The numerical results confirm, by and large, the findings of Sec. III. Figure 4 shows Brier skill scores of the Crooks, Basis, and Zero forecasts (and weighted composites of forecasts that will be explained later) for different training data set sizes N. They are plotted logarithmically over threshold τ, respectively,

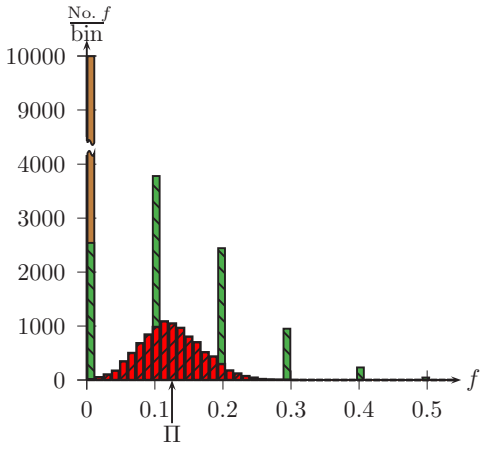

FIG. 3. (Color online) Histogram of $n = 10^4$ Zero (brown resp. long, plain, and interrupted column at $f = 0$), Basis (green, striped top left to bottom right), and Crooks (red, striped bottom left to top right) forecasts, each estimated using data sets of size $N = 10$. The exact probability is $\Pi = 0.1252$.

linearly over probability (to exceed this negative threshold) Π, which emphasizes different regions.

The green unfilled circles and (straight gray) lines at $-1/N$ show the Brier skill score for the Basis forecast, calculated empirically, i.e., over $n = 10^4$ experiments, and analytically, as given by Table I. The green lines give the reference for the skill, which a forecast has to exceed to be better than using only the relative frequency in the data set. This reference increases towards zero for increasing training data set size, as the approximation of the climatology can be made more exact having more data points available.

The approximation of the analytical limit of the Basis Brier skill score by the empirical estimate deviates strongly in the range of very low probability, i.e., low threshold τ. Some of the points lie on the line of the Zero Brier skill score [see left column of plots, green (unfilled) circles on top of the brown (short dashed) line]. The reason is that in none of the n experiments was the Basis forecast nonzero, which illustrates the rareness of the event we are predicting.

Forecasting zero for very low thresholds seems to be a good strategy, as can be seen in the brown (short dashed) curves, which show the Brier skill score for the Zero forecast. There is a range of low probabilities where the Zero forecast improves compared to the Basis forecast. This range is broader for smaller training data sets. This is reasonable, as the lowest nonzero Basis forecast that can be made is $1/N$, which is already too high for a certain range of probability Π.

The red (dark gray solid) curves in Fig. 4 show the Brier skill score of the Crooks forecast. We did not plot the empirical mean values because they lie on top of their analytic expectation values for the whole range plotted here. As expected, the Crooks forecast improves compared to the Basis forecast for the whole range of thresholds τ shown here. It becomes better when decreasing the threshold. Neglecting the range of the ordinate, the relative position of the Brier skill scores for the Crooks and the Basis forecasts is the same in all of the plots. The reason is that they are inversely proportional to the sample size, and thus their ratio is independent of N. In

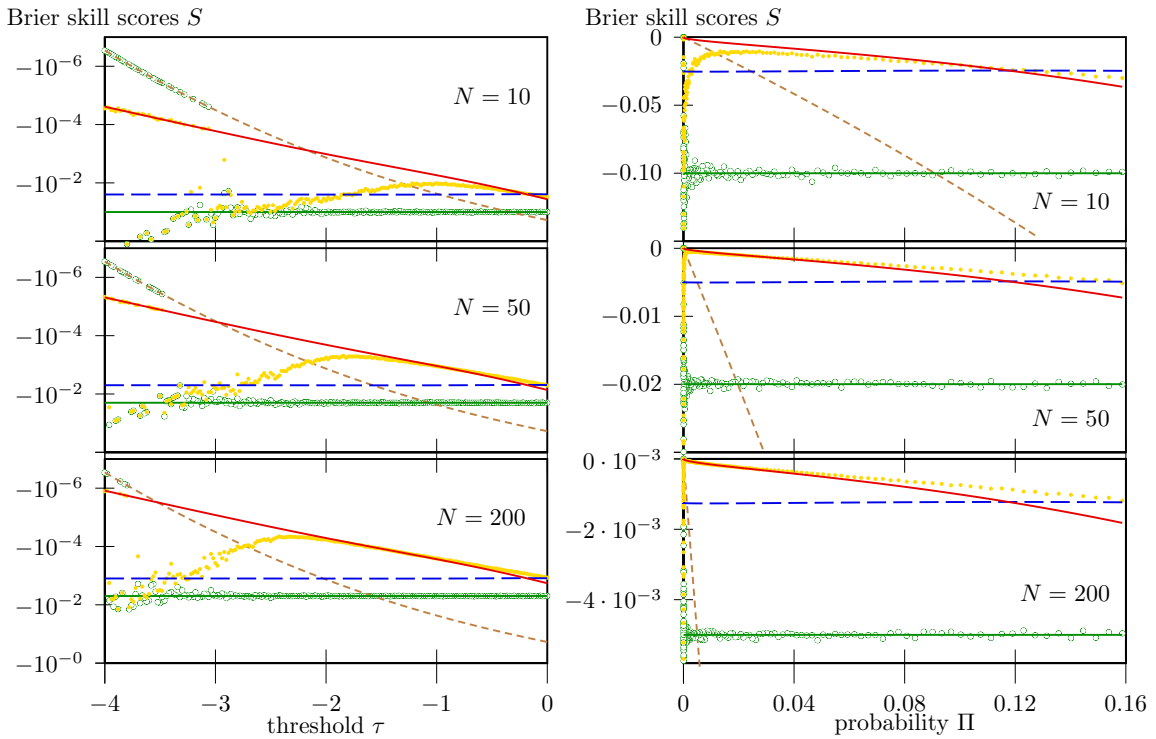

FIG. 4. (Color online) Mean Brier skill scores for training data sets of size N (10,50,200), plotted logarithmically over threshold τ, respectively, linearly over probability Π to be below the threshold τ. Samples are drawn from a Gaussian distribution with mean and standard deviation equal to 1, i.e., $\beta = 2$. The color coding of the curves and points is as follows: red (dark gray solid) line: Crooks forecast; green (gray solid) line and unfilled circles: Basis forecast, analytical and numerical values; blue (long dashed) line: composite 0.5; orange (light gray filled) dots: composite *cm*; brown (short-dashed) line: Zero forecast.

the range of low thresholds (see left column), the Zero forecast is even better than the Crooks forecast. This effect decreases with increasing N.

Figure 4 describes the situation for one exemplary β. Several facts still hold for the other values of β: There are still jumps of the Basis skill onto the skill for the Zero forecast for low thresholds because none of the experiments ever provides a nonzero Basis forecast. The improvement of the Crooks skill compared to the Basis skill increases with decreasing threshold. The intersection of the curves, where the Zero forecast becomes worse than Crooks, decreases towards smaller thresholds with increasing β. For a given τ, the Crooks forecast is the better the higher the parameter β. For low β the distribution is closer to being symmetric around zero, i.e., the number of points that are available for the Crooks forecast are not much larger than for the Basis forecast. Unexpectedly, for large β and small samples (e.g., $\beta = 4, N = 10$) the Zero forecast is better than Crooks not only at very small thresholds, but starts to improve upon the Crooks forecast again at thresholds close to zero. [The red (dark gray solid) curve falls steeper than the brown (short dashed) curve there.] This effect is not addressed further here, but might be due to the unsuitability of the Brier skill score as a criterion of predictive skill. (A logarithmic criterion, which is $-\infty$, when the forecast probability is zero, would not suffer from this effect.)

One might ask in this special case as to how good a prediction can be using the knowledge that we make forecasts about a Gaussian variable. Therefore, we calculate mean and standard deviation empirically for each training data set and then obtain the prediction as an integral over the Gaussian distribution with these parameters. The curves are not shown. In general, the skill of this method is between the Basis' and the Crooks' skill. For low β and low thresholds, this method is better than the Crooks forecast (this happens in a region where the Zero forecast is better, too). The reason is the same as above: there is no great advantage by the amount of data that is used to calculate the Crooks forecast.

3. *Composites of Basis and Crooks forecasts*

Each of the two estimators (Basis and Crooks) neglects that part of the data, which is used by the other estimator. Therefore, it seems likely that predictions can be further improved by combining them. In the following, two composites of f_C and f_B are proposed, both of which are of the following form:

$$f_\alpha = \alpha f_C + (1-\alpha) f_B, \qquad \alpha \in [0,1] \tag{15}$$

(i) $f_{0.5}$: arithmetic mean. Both predictions are equally weighted $\alpha = \frac{1}{2}$.

(ii) f_{cm}: forecast weighted according to the number of contributing sample members. The Crooks and the Basis forecasts are weighted proportionally to the number of sample members a_C, respectively a_B, that are above $-\tau$, respectively, below τ, i.e.,

$$\alpha = \frac{a_C}{a_C + a_B}. \tag{16}$$

In general, if α does not depend on the exact forecast (like in the first composite), the reliability of the f_α forecast can be calculated. It is

$$\mathrm{REL}_\alpha = \frac{\alpha^2\mathcal{E} + (1-\alpha)^2\Pi - \Pi^2}{N}. \tag{17}$$

With this, the Brier score and skill score are given as well.

In Fig. 4, the composites $f_{0.5}$ and f_{cm} are indicated by blue (long-dashed) curves and orange (filled) dots, respectively. If the deviations in the Crooks and the Basis forecasts are to different sides of the perfect probability Π, then the arithmetic mean $f_{0.5}$ can cancel these deviations. This happens in the range of high thresholds, where the two forecasts are anticorrelated because the data which is in one part of the histogram is missing in the other. In the range of low thresholds, the skill of the arithmetic mean forecast lies between the Crooks and the Basis forecasts. There is no range where this forecast is worse than the Basis forecast.

The Brier skill score of the composite f_{cm} can only be approximated empirically. It is better than the Crooks forecast for thresholds close to zero. The range where it is better increases with increasing sample size N. For low thresholds, it jumps to the Crooks forecast, whenever all of the Basis forecasts are zero. We conclude that combining the knowledge of the two forecasts can lead to an even better prediction, but it is important how much both forecasts are weighted.

The question appears as to what is the best possible composite of the Basis and Crooks forecasts. Therefore, we optimize the α in the composite such that on average this forecast performs best. Keep in mind that for single realizations of the experiment this optimized forecast can be worse than either of the Basis or Crooks forecast. Further, notice that the following calculation uses the knowledge of the exact distribution, which is in general not available. Nonetheless, it provides some helpful insights. We find the optimal composite by minimizing the reliability [Eq. (17)] with respect to α. This corresponds to minimizing the Brier score or maximizing the Brier skill score. The optimal α is given by

$$\alpha_{\mathrm{opt}} = \frac{\Pi}{\Pi + \mathcal{E}}, \tag{18}$$

where $\mathcal{E}$ is the expectation value of $e^{\beta W}$ over $W < \tau$. It is independent of the sample size. From the analysis above (see items in Sec. III), we know that $\Pi > \mathcal{E} \geqslant \Pi^2$ (for negative thresholds). It follows that $\frac{1}{2} < \alpha_{\mathrm{opt}} \leqslant \frac{1}{1+\Pi}$. This means that, for an optimal composite, the Crooks forecast should always be weighted stronger than the Basis forecast. On the other hand, for a nonzero probability Π the Basis forecast should always be included because $\alpha < 1$. The corresponding Brier score is

$$S_{\alpha_{\mathrm{opt}}} = -\frac{1}{N}\frac{\mathcal{E} - \Pi^2 - \Pi\mathcal{E}}{(\Pi + \mathcal{E})(1+\Pi)}. \tag{19}$$

It is larger (i.e., closer to zero) than the Crooks skill score (see Table I) by having a smaller numerator and a larger denominator than the latter.

Figure 5 shows the optimal α depending on the threshold, for a Gaussian distribution with $\beta = 1$. For low β, the range, where the Basis forecast is used to a considerable amount, is quite broad. The higher the β, the faster the optimal α approaches 1. The inset shows the corresponding Brier skill scores. The optimal α forecast (violet curve) can improve by up to 78% compared to the pure Crooks forecast (or 91% compared to the Basis forecast) in the regime of thresholds close to zero.

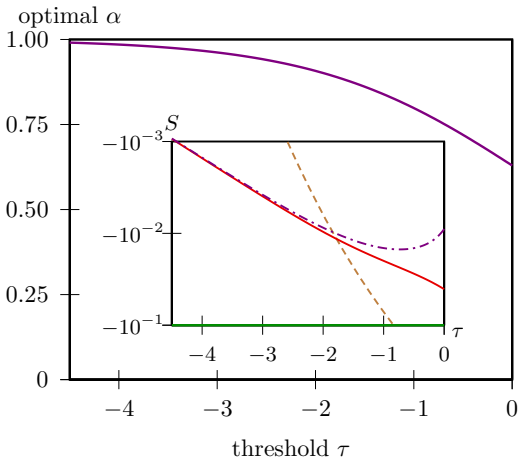

FIG. 5. (Color online) α parameter for the optimal composite of Crooks and Basis forecast [see Eq. (15)] in order to maximize the Brier skill score for a Gaussian distribution with $\beta = 1$. The inset shows the Brier skill score for the optimal α forecast (violet, dashed-dotted line) for training data set size $N = 10$. The color coding is like before: Crooks: red (dark gray solid line); Basis: green (gray solid line at $S = -10^{-1}$); Zero: brown (dashed line).

If the reliabilities of the Crooks and Basis forecasts are not available as analytical expressions, one can estimate the optimal α from empirical data. This is done by minimizing the reliability of the forecast f_α [Eq. (15)], given by $\mathbb{E}[(f_\alpha - \Pi)^2]$ with respect to α. One finds

$$\alpha = \frac{\mathbb{E}\left[(f_B - f_C)(f_B - \Pi)\right]}{\mathbb{E}[(f_B - f_C)^2]}. \tag{20}$$

The expectation values can be approximated by empirical mean values. In the range of low thresholds, where none of the Basis forecasts were nonzero, the optimal α also makes jumps from a value near to 1 down to below $\frac{1}{2}$. This is because the Zero forecast usually is better in this range. Within this unsteady behavior the empirical calculation should no longer be trusted. The derivation of Eq. (20) seems rather artificial here, as we could have already used Eq. (18) to calculate the optimal composite. However, we will use Eq. (20) in Sec. IV B, where the distribution is not given in analytical form and we can only estimate the cumulative probability Π.

4. Diverging side of the distribution

In the Introduction, we claimed that we can improve forecasts if we use the knowledge that a distribution satisfies the Crooks relation. We were interested in estimating the probability that a value W is below a certain threshold τ. It can be shown that the Crooks forecast we constructed improves compared to the Basis forecast, if we measure their quality with the Brier score. The above holds true if we have negative thresholds τ (and a positive parameter β). There, the individual contributions to the Crooks forecast, which are equal to $e^{-\beta W}$,

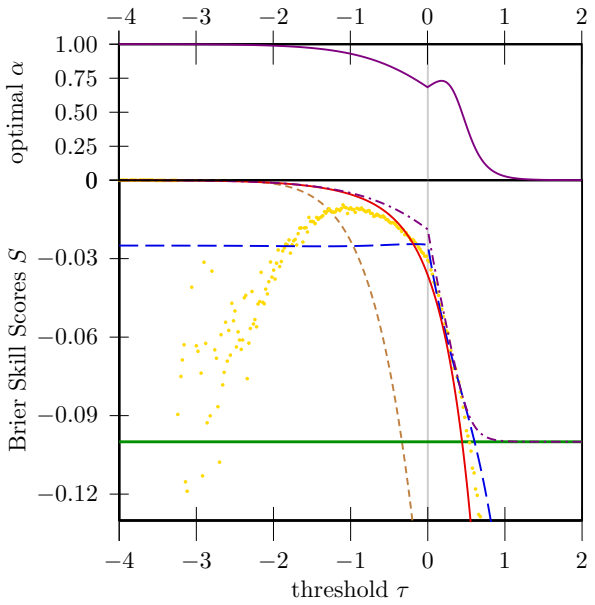

FIG. 6. (Color online) Optimal α (top) and Brier skill scores (bottom) plotted over threshold τ. Data basis are sets of size $N = 10$ drawn from a Gaussian distribution with $\sigma = 1$ and $\mu = \beta/2$ with $\beta = 2$. The lower plot shows familiar curves for the Crooks (red resp. dark gray solid line), Basis (green resp. gray solid line at $S = -0.1$), Zero (brown, short dashed line), composite 0.5 (blue, long dashed line), and composite *cm* (orange dots) forecast. The latter can only be given by empirical approximation and is thus not plotted as a continuous line.

are small. For $\tau > 0$, these contributions diverge exponentially. In this paragraph, we extend the range of thresholds to positive values (given that $\beta > 0$). Doing so, the two forecasts share a set of values, namely, in the range $-|\tau| < W < +|\tau|$.

Figure 2 shows that there is a certain range of positive thresholds, where the Crooks forecast is still better, i.e., $\mathcal{E} < \Pi$, but at some τ the reducing effect of the negative range on the expectation value $\mathcal{E}$ is compensated. From this τ on the Basis forecast is better. See Fig. 6 (lower plot), where we show Brier skill scores of the different forecasts for thresholds extended into the positive regime. Very quickly, the Crooks forecast becomes worse than the Basis forecast. This affects the composites as well. The Brier skill scores of the diverging curves reach down to -10 in the regime shown here.

The analytical calculation of some quantities, which include the covariance term $\mathbb{E}[f_C f_B]$, has to be extended because the latter contains a term which is proportional to $\Theta(\tau - W)\Theta(W + \tau)$. This term vanishes for negative τ, but does not for positive. This affects the calculation of the reliability of mean values with a fixed α such as the arithmetic mean, and the optimal α. Equations (17) and (18) have to be modified in this regime.

The forecasts f_B and f_C are strongly anticorrelated for $\tau \lesssim 0$ because a data point which was missing in one forecast (and making its predicted value smaller) is included in the interval used by the other forecast. At $\tau > 0$ this changes because of the commonly used regime the correlation gets positive. The correlation and anticorrelation are stronger for a low parameter β because then the values $e^{-\beta W}/N$ and $1/N$

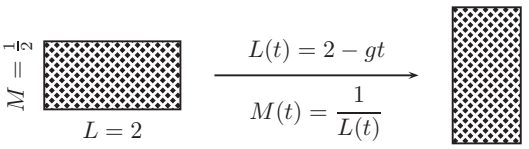

FIG. 7. Visualization of the transformation of the domain from a lying rectangle to a standing one. This is done by linearly decreasing the domain width L from 2 to $\frac{1}{2}$ with a speed g. In the setup here $g = 100$. (See Sec. IV B for details.)

included in the prediction f_C and f_B, respectively, become closer to each other. The optimal α becomes

$$\alpha_{\text{opt}} = \frac{\Pi - \xi\Theta(\tau)}{\Pi + \mathcal{E} - 2\xi\Theta(\tau)}, \quad \text{with} \quad \xi = \int_{-|\tau|}^{|\tau|} p(W)dW \tag{21}$$

and $\Theta(\tau)$ being the Heaviside function. It is plotted in Fig. 6 (upper plot). The parameter α has a dip at $\tau = 0$. One can show that decreasing the denominator and numerator by ξ and 2ξ, respectively, pulls α away from 0.5. That is why it is increasing again. At some point, this effect is outweighed by the strong increase of the expectation value $\mathcal{E}$ in the denominator. The curve starts decreasing, passes $\alpha = 0.5$, where $\Pi = \mathcal{E}$, and goes to zero for $\tau \to \infty$.

B. Experimental data

In this second example, we use data obtained from a numerical experiment. We do not have the exact distribution of the random variable available here, but we know that it satisfies Crooks relation (2), and we know the parameter β. The random variable used in this example is the work, i.e., the change in kinetic energy that has to be imposed when deforming a rectangular domain in which two-dimensional fluid flow evolves. The fluid flow is approximated obeying two-dimensional truncated Euler equations. The process of transformation is visualized in Fig. 7. During the transformation work has to be imposed on the field (or can be extracted, this is the general case here). This work is dependent on the exact field composition. Here, we take the distribution of work and the parameter β as given. For details illustrating the system, numerical details and the deduction of Crooks' relation for this system and process (see [6]).

The final and initial states of the transformation shown in Fig. 7 are symmetric under a rotation by 90°. That is why the backward process [see Eq. (1)] is identical to the forward process and the original Crooks relation collapses to a equation for only one distribution $p(W)$ [see Eq. (2)]. This distribution over various repetitions of the experiment is shown in Fig. 8. It is highly asymmetric and satisfies Eq. (2) with a certain β derived from the initial conditions of the fields, here $\beta = -19$.

Having a negative sign for β we can improve predictions of work values being larger than positive thresholds τ, i.e., $P(W > +\tau)$ because then the Crooks forecast does not use the diverging side of the distribution. As we do not have the exact distribution of work given in analytic terms, we approximate the cumulative probability by using two million data points. Despite this abundance of sample points, the accuracy of the Brier skill score is affected, especially for

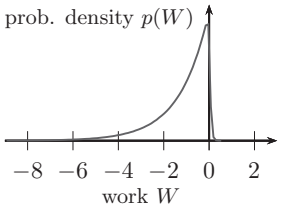

FIG. 8. Distribution of the work imposed in the transformation described in Sec. IV B.

very high thresholds, because there Π is approximated poorly. That is why we sampled the region of small τ denser than for large τ. Figure 9 shows the Brier skill scores for the Crooks and the Basis forecasts for different training data set sizes N. The courses of the curves are in general the same as in the previous example. For all sample sizes, the Crooks forecast improves compared to the Basis forecast. This improvement is better the further one is in the tail of the distribution. However, for small training data set sizes (e.g., $N = 10$), the Zero prediction is still better.

We did not include the skills for the composites here. The general behavior is like for the Gaussian distribution in Sec. IV A. For the calculation of the optimal α forecast we used Eq. (20) with the approximated cumulative density. This forecast gives some improvement for low thresholds τ, where the Basis forecast is weighted with up to 40%. For high thresholds, the pure Crooks forecast gives the best prediction, indicated by α increasing up to one. Again, this effect is independent of the sample size. This time, the forecast composite weighted by the number of contributing members is very close to the pure Crooks forecast. The reason is the high asymmetry of the distribution (see Fig. 8); the probability to have a nonzero Basis forecast is considerably low. The arithmetic mean, however, improves a lot compared to the Crooks forecast in the range of low thresholds.

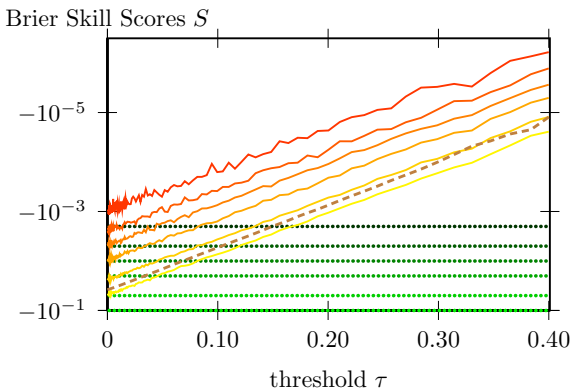

FIG. 9. (Color online) Brier skill scores for the Crooks forecast (in red, solid line) and the Basis forecast (in green, dotted line) plotted over threshold values τ. The colors vary from light colored to dark for increasing training data set size $N = 10,20,50,100,200,500$. The training data are taken from the distribution introduced Sec. IV B. As the absolute number of values W for the training data sets was 5×10^5, the number of experiments n varied from 5×10^4 down to 10^3. The Zero forecast (brown dashed line) is included as a reference.

V. CONCLUSIONS AND DISCUSSION

By the analytical derivations in this paper, and the supporting examples, we found that we can improve predictions of rare events using the Crooks fluctuation theorem. We confirmed our main hypothesis: If we use the knowledge that a random variable is drawn from a distribution which satisfies the Crooks fluctuation theorem with itself, we can give an estimate of the probability to exceed a threshold which is better than simple histogram counting. We measured the notion of "better" in terms of the Brier score reliability and Brier skill score and were able to show that the Crooks forecast always improves compared to the Basis forecast, if $\tau\beta < 0$. This improvement is universal, independent of the exact distribution of the prediction target. We explored the strengths and limitations of this new estimator for two examples over a wide range of parameter values (such as threshold and sample size), and proposed further improvement by combining the estimators.

Throughout the article, we assumed the parameter β to be given. In the nonequilibrium thermodynamic framework, where the Crooks relation originates, this parameter is a quantity given by the equilibrium description of the system. It is known *a priori*, for example, as the inverse temperature of a heat bath. To the authors, no report about any experiment or process is known, which provides a density fulfilling Eq. (2), but is not intrinsically thermodynamic. Maybe this is due to the fact that no one has ever looked for this outside the thermodynamic framework. Estimating the parameter β only from the data could then be an issue. One simple way is to estimate the distribution $p(W)$ by a histogram, and plotting the ratio of $p(W)$ and $p(-W)$ logarithmically. The fitted slope approximates the parameter β [see Eq. (2)]. For small data set sizes (like the ones considered here), this method underestimates the absolute value of β. We refer the interested reader to [6], where this effect of underestimating the slope of the curve is addressed. Anyway, the underestimated β will lead to biased estimates of the exceedance probability. As a topic for future investigation, more sophisticated techniques have to be devised, which turns out to be a difficult task, if one wants to estimate β from data set sizes as small as considered here (e.g., $N = 10$).

Like the issue of knowing or estimating the parameter β, the question appears as to what extent this method could be applied to the prediction on densities, which are Crooks-like, in the sense that they fulfill the Crooks relation approximately. This question can not be answered in a short sentence. It contains the issue to define how Crooks-like is Crooks-like enough, and depends on the thresholds and sample sizes considered.

The logarithmic ratio of a density $\ln \frac{p(W)}{p(-W)}$ fulfilling Crooks' relation (2) fits a line with slope β. In general, the logarithmic ratio of any distribution is to first order around $W = 0$ a linear function of W. From this linear slope one could read the parameter β. But, for any non-Crooksian distribution, the higher order terms of this probability ratio are nonzero. Thus, for values W, which are not in the vicinity of the origin, the slope of the curve deviates from a linear function. It is easy to show that this deviation leads in any case to a biased Crooks forecast. Exactly this can be seen for $p(W) \sim \exp\{-\frac{(W-\mu)^4}{2\sigma^4}\}$, a distribution which at first sight looks similar to a Gaussian.

As it falls steeper than a Gaussian distribution in the tails, the logarithmic probability ratio is pulled away from a linear behavior towards lower values for positive W (and higher for negative). The Crooks forecast, naively calculated for this distribution (with the parameter β obtained from the slope of the logarithmic ratio in the vicinity of $W = 0$), is a biased estimator, predicting a probability which is too high. Even though its variance might be smaller than the variance of the Basis estimator at thresholds close to zero, this is not the case for thresholds in the tail of the distribution. Thus, such a forecast is in general inferior to the Basis estimator.

We quantified the predictive skill of the forecast by the Brier score. The choice of a scoring rule is crucial for the ranking of forecast schemes. Using the Brier score made the Zero forecast appear like a favorable choice, as its skill can be larger than the skills of the Crooks or the Basis forecast. However, predicting zero probability simply ignores the possibility of an event to happen. If this event is extreme, possibly associated with life-threatening dangers for people or devastating damages or losses to properties, then its predicted impossibility may cause strong consequences, e.g., high monetary compensation costs, if the event does happen. A scoring criterion, which penalizes false zero forecasts more heavily than the Brier score, would certainly be favorable in such a scenario. The problem of suitable evaluation criteria for predictions of rare events is addressed in, e.g., [31].

A well established mathematical tool for extrapolation from extremes in a finite data set to an infinite data set is (generalized) extreme value theory (GEV) [32]. Universality of asymptotic distributions yields a way to determine, by fitting three parameters to an empirical histogram, the return level of events of a magnitude which are larger than all observed events. There is no need to make any quantitative comparison between this method and the one of this article since sample sizes needed to apply GEV are by orders of magnitude larger than those used here, so that GEV would not make any meaningful statement in setups like those considered in this article. However, the improvement of methods from GEV in situations where the Crooks fluctuation theorem is valid (and sufficient data are available) is an interesting subject that will be considered in forthcoming studies.

[1] A. Boin, P. Lagadec, E. Michel-Kerjan, and W. Overdijk, J. Contingencies Crisis Manag. **11**, 99 (2003).

[2] R. Albert, I. Albert, and G. L. Nakarado, Phys. Rev. E **69**, 025103 (2004).

[3] R. R. Holmes, Jr. and K. Dinicola, U.S. Geological Survey General Information Product **106**, 1 (2010).

[4] G. A. Meehl, T. Karl, D. R. Easterling, S. Changnon, R. Pielke, D. Changnon, J. Evans, P. Y. Groisman, T. R. Knutson, K. E. Kunkel, L. O. Mearns, C. Parmesan, R. Pulwarty, T. Root, R. T. Sylves, P. Whetton, and F. Zwiers, Bull. Am. Meteorol. Soc. **81**, 413 (2000).

[5] G. E. Crooks, Phys. Rev. E **60**, 2721 (1999).

[6] J. Gundermann, H. Kantz, and J. Bröcker, Phys. Rev. Lett. **110**, 234502 (2013).

[7] *Forecast Verification: A Practitioner's Guide in Atmospheric Science*, 2nd ed., edited by I. T. Jolliffe and D. B. Stephenson (Wiley, Chichester, 2011).

[8] T. Gneiting, F. Balabdaoui, and A. E. Raftery, J. R. Stat. Soc. B **69**, 243 (2007).

[9] G. W. Brier, Mon. Weather Rev. **78**, 1 (1950).

[10] D. J. Evans, E. G. D. Cohen, and G. P. Morriss, Phys. Rev. Lett. **71**, 2401 (1993).

[11] D. J. Evans and D. J. Searles, Phys. Rev. E **50**, 1645 (1994).

[12] G. Gallavotti and E. G. D. Cohen, Phys. Rev. Lett. **74**, 2694 (1995).

[13] C. Jarzynski, Phys. Rev. Lett. **78**, 2690 (1997).

[14] R. Kawai, J. M. R. Parrondo, and C. Van den Broeck, Phys. Rev. Lett. **98**, 080602 (2007).

[15] M. Esposito and C. Van den Broeck, Phys. Rev. Lett. **104**, 090601 (2010).

[16] F. Douarche, S. Ciliberto, and A. Petrosyan, J. Stat. Mech.: Theory Exp. (2005) P09011.

[17] O.-P. Saira, Y. Yoon, T. Tanttu, M. Möttönen, D. V. Averin, and J. P. Pekola, Phys. Rev. Lett. **109**, 180601 (2012).

[18] K. Hayashi, H. Ueno, R. Iino, and H. Noji, Phys. Rev. Lett. **104**, 218103 (2010).

[19] D. Collin, F. Ritort, C. Jarzynski, S. B. Smith, I. Tinoco Jr., and C. Bustamante, Nature (London) **437**, 231 (2005).

[20] C. Bustamante, Q. Rev. Biophys. **38**, 291 (2005).

[21] A. H. Murphy, J. Appl. Meteorol. **12**, 595 (1973).

[22] M. S. Roulston and L. A. Smith, Mon Weather Rev. **130**, 1653 (2002).

[23] J. Bröcker, Q. J. R. Meteorol. Soc. **135**, 1512 (2009).

[24] A. P. Weigel, M. A. Liniger, and C. Appenzeller, Mon. Wea. Rev. **135**, 118 (2007).

[25] D. S. Wilks, *Statistical Methods in the Atmospheric Sciences*, 3rd ed., International Geophysics Series, Vol. 100 (Academic, Oxford, 2011).

[26] A. Kumar, A. G. Barnston, and M. P. Hoerling, J. Clim. **14**, 1671 (2001).

[27] D. S. Richardson, Q. J. R. Meteorol. Soc. **127**, 2473 (2001).

[28] S. J. Mason, Mon. Weather Rev. **132**, 1891 (2004).

[29] W. A. Müller, F. J. Doblas-Reyes, M. A. Liniger, and C. Appenzeller, J. Clim. **18**, 1513 (2005).

[30] A. M. Mood, F. A. Graybill, and D. C. Boes, *Introduction to the Theory of Statistics*, 3rd ed. (McGraw-Hill., New York, 1974).

[31] B. Casati, L. J. Wilson, D. B. Stephenson, P. Nurmi, A. Ghelli, M. Pocernich, U. Damrath, E. E. Ebert, B. G. Brown, and S. Mason, Meteorol. Appl. **15**, 3 (2008).

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