Statistical evaluation of forecasts

Malenka Mader,[1,2,3,4,5,*] Wolfgang Mader,[2,3,4] Bruce J. Gluckman,[4,5,6,7] Jens Timmer,[2,3,8] and Björn Schelter[2,9,†]

[1]*Department of Neuropediatrics and Muscular Disease, University Medical Center of Freiburg, Freiburg, Germany*

[2]*Freiburg Center for Data Analysis and Modeling (FDM), University of Freiburg, Freiburg, Germany*

[3]*Institute for Physics, University of Freiburg, Freiburg, Germany*

[4]*Center for Neural Engineering, Pennsylvania State University, State College, Pennsylvania 16801, USA*

[5]*Department of Engineering Science and Mechanics, Pennsylvania State University, State College, Pennsylvania 16801, USA*

[6]*Department of Neurosurgery, Pennsylvania State University, Hershey, Pennsylvania 17033, USA*

[7]*Department of Biomedical Engineering, Pennsylvania State University, State College, Pennsylvania 16801, USA*

[8]*BIOSS, Center for Biological Signalling Studies, University of Freiburg, Freiburg, Germany*

[9]*Institute for Complex Systems and Mathematical Biology, University of Aberdeen, Aberdeen, UK*

(Received 21 January 2014; published 26 August 2014)

Reliable forecasts of extreme but rare events, such as earthquakes, financial crashes, and epileptic seizures, would render interventions and precautions possible. Therefore, forecasting methods have been developed which intend to raise an alarm if an extreme event is about to occur. In order to statistically validate the performance of a prediction system, it must be compared to the performance of a random predictor, which raises alarms independent of the events. Such a random predictor can be obtained by bootstrapping or analytically. We propose an analytic statistical framework which, in contrast to conventional methods, allows for validating independently the sensitivity and specificity of a forecasting method. Moreover, our method accounts for the periods during which an event has to remain absent or occur after a respective forecast.

DOI: 10.1103/PhysRevE.90.022133 PACS number(s): 02.50.−r, 95.75.Wx

I. INTRODUCTION

Extreme but rare events occur in ecology as in the absence/presence models of species [1], in weather in the form of storms [2,3], heavy precipitation [2,3], floods [4], and extreme temperatures [3], in geology as earthquakes, and in economics [5] as financial crises, as well as in medicine in the form of strokes and epileptic seizures [6,7]. The impacts of such extreme and unpredicted events are severe. A reliable forecasting system could alleviate those impacts, because precautions or interventions [6,7] could be taken. A forecasting system would be considered reliable when predicting both an upcoming event (sensitivity) and the absence of a future event (specificity). Depending on the application, the focus might vary on either the high sensitivity or the high specificity of the prediction.

In general, a forecasting system is based on several features, which quantify different properties of observations, such as frequency content and regularity (for reviews, see [6,8], and [9]). These features are supposed to quantify a change in the observed system before an extreme event occurs. In that case a probabilistic [2,10,11] or a binary [1,3,6,9,12] forecast can be derived from such features.

For probabilistic forecasts, the probability of the occurrence of an event is provided [2,10,11]. Such forecasts leave the user with the decision about the necessity to initiate precautions or preventions [5,13]. Probabilistic forecasts can be statistically assessed by the Brier score [14],

$$B = \frac{1}{N}\sum_{t=1}^{N}(\hat{p}(t) - I(t))^2, \tag{1}$$

which is the temporal mean of the squared deviations of the probabilities $\hat{p}(t)$ of an upcoming event at time t to the indicator function $I(t)$ of actually occurring events. The indicator function $I(t)$ is 0, except at times of events when $I(t) = 1$. Inserting a binary forecast

$$\hat{I}(t) = \begin{cases} 1 & \text{if an alarm is raised at time } t, \\ 0 & \text{otherwise,} \end{cases} \tag{2}$$

instead of the probabilistic forecast $\hat{p}(t)$ into Eq. (1), yields

$$B = \frac{1}{N}\left(\sum_{\tilde{t}'\in \mathrm{TP}} 0 + \sum_{\tilde{t}'\in \mathrm{FP}} 1 + \sum_{\tilde{t}''\in \mathrm{TN}} 0 + \sum_{\tilde{t}''\in \mathrm{FN}} 1\right) \tag{3}$$

$$= \frac{m_+ \times 0 + (T - m_+) \times 1 + m_- \times 0 + (N - T - m_-) \times 1}{N} \tag{4}$$

$$= 1 - \frac{m_+ + m_-}{N}, \tag{5}$$

where $m_\pm$ are the numbers of true positives and true negatives, T is the total number of predicted events, and N is the number of prediction-observation pairs available. For this, the sum in Eq. (1) is split up into four sums, as in Eq. (3). These sums contain the time points of true positive (TP), false positive (FP), true negative (TN), and false negative (FN) predictions $\hat{I}$.

Equation (5) shows that for the Brier score the number of TPs m_+ and true negatives m_- are weighted equally, when applied to binary forecasts. This is particularly inexact when faced with rare events because a high number of true negatives then obscures a low number of true positives.

For binary forecasts, contingency tables are typically considered. They contain information about true and false

*Corresponding author: M.Mader@fdm.uni-freiburg.de

†Corresponding author: B.Schelter@abdn.ac.uk

positives, as well as true and false negatives. The statistics of the contingency table can be assessed by the Fisher-Irwin test [15]. For the evaluation of a prediction, a time frame within which the event must or must not occur needs to be defined. This time frame enhances the probability of true positive and true negative predictions, which is not accounted for in the statistics of the Fisher-Irwin or similar tests. As an alternative to statistically testing the entries in contingency tables directly, certain measures [12,16,17] and according statistical tests [18,19] are available that take into account the time frames in which events must or must not occur. However, they do not quantify the fraction of true positives, i.e., the sensitivity, and the fraction of true negatives, i.e., the specificity, independently [12,16,17]. The statistics [19] of the seizure prediction characteristics [18], for instance, is based on a Poissonian random predictor with a binomial success distribution when assuming a certain false prediction rate, i.e., specificity. This dependency of sensitivity and specificity is also shown for the Brier score in Eq. (5). To summarize, neither for binary nor for probabilistic forecasts has independent statistical assessment of sensitivity and specificity in the framework of predictions been proposed so far. We close this gap by presenting an adaptive Poissonian random predictor to obtain binomial success distributions for the sensitivity and specificity, independently.

The article is organized as follows. In Sec. II we review the statistical assessment of binary forecasting methods. In Sec. III we propose and test a statistical method to independently assess the number of true positives and true negatives of forecasting methods based on random predictors. For two different settings we show the adaptablity of our method in Secs. III A and III B. By simulations we test the efficiency of the proposed statistics for both settings. We complete this paper by discussing the method proposed (Sec. IV).

II. STATISTICAL PROPERTIES OF BINARY PREDICTION METHODS

A binary forecast, which we denote a prediction, in contrast to the probabilistic nature of forecasts, can be either positive or negative. A positive prediction yields an alarm raised to announce that an event is about to occur. If correctly predicting, a positive prediction leads to a true positive. True negatives are negative predictions followed by the absence of events. Standard ways of quantitatively assessing true and false positives and true and false negatives are described in Sec. II A.

In Sec. II B, the idea of statistical evaluation of the contingency tables is reviewed. The performance of the prediction system needs to be tested in terms of superiority to a random predictor. We propose an analytic solution for this step.

A. Quantification of prediction-observation matches

A useful framework for quantifying the matches of predictions and observations is based on the introduction of an occurrence period (OP) and an intervention period (IP) (Fig. 1). The period during which the anticipated event has to occur after an alarm is the OP [6]. This renders a classification of true positives [Fig. 1(a)] and false positives [Fig. 1(b)] possible. Typically, an event-free period is desirable after the alarm before the OP sets on. It allows for an intervention or precaution to be initiated after the alarm, before the occurrence of the event [Fig. 1; dotted (green) box]. This period is denoted the IP [6,8], prediction horizon [20], or lead time [4].

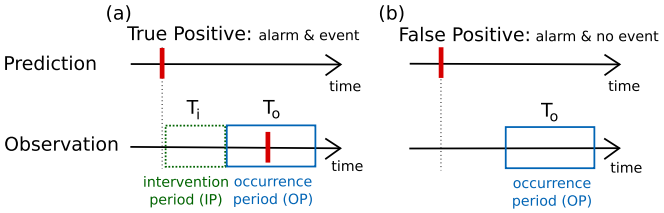

FIG. 1. (Color online) A prediction method that raises an alarm [thick (red) vertical line] can be validated only if the event has to occur in a specific time window [solid (blue) box], the OP of length T_o, after the alarm was triggered. The period between the alarm and the onset of the OP is the IP of length T_i, in which intervention systems can be initiated [dotted (green) box]. (a) A true positive alarm is then defined as an alarm which is followed by an event in the OP. (b) If an event remains absentin the OP, the alarm is a false positive.

A negative prediction (Fig. 2) is followed by the IP in the sense that a nonoccurrence of an event should be guaranteed during this phase to obtain a true negative prediction [Fig. 2(a)]. If an event occurred, on the other hand, the negative prediction would be wrong, leading to a false negative [Fig. 2(b)].

The respective numbers of true and false positives and true and false negatives are summarized in contingency tables, which then can be statistically validated, as described in the next section.

B. Evaluation of prediction-observation matches

To statistically validate contingency tables derived from the assessment of positive and negative predictions (Sec. II A), chi-square tests or the Fisher-Irwin test can be used to test the independence of predictions and events [15,21]. However, they do not account for the statistical effect of the OP. After a positive prediction an event can occur within the duration of the OP. For a long OP, the probability of a

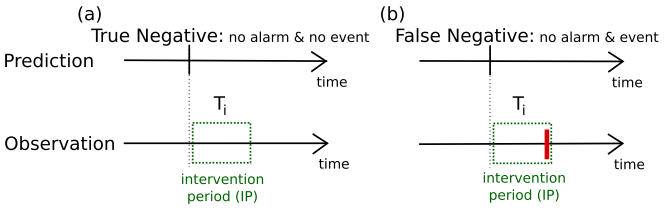

FIG. 2. (Color online) Negative predictions refer to the IP [dotted (green) box], since the time points after the IP could belong to an occurrence period of a positive prediction after the negative prediction. (a) A true negative prediction refers to a negative prediction after which no event occurs within the IP. (b) A false negative prediction refers to one after which an event occurs at the end of the IP.

true positive prediction is increased compared to that for a short one. In order to circumvent this statistical challenge, measures [1,6,9,12] based on the contingency tables have been introduced. They quantify the performance of the predictor by weighting the sensitivity and specificity differently. One example of such a measure is the fraction correct, which is the sum of the sensitivity and specificity [9]. For some of those measures, the distributions of the test statistics are known; for others they are approximated by Gaussian distributions. These distributions of test statistics are obtained for the case of linkage and the case of independence of prediction-observation pairs. In the case of linkage, the predictions of a prediction method are correlated with the observations. On the other hand, predictions and observations are uncorrelated in the case of independence. Correct predictions might still occur; for instance, a positive prediction is followed by an event. However, such correct predictions are due to chance. The aim of any statistical assessment of prediction methods is to exclude that the performance is just as good as it is for such an independent prediction method, for which predictions and observations coincide by chance. That is, the null hypothesis of independence of predictions and observations needs to be rejected. To this end, a threshold which separates the two distributions corresponding to linkage and independence for a given error probability can be obtained analytically [9]. Since error probabilities of positives and negatives are linked by the choice of the threshold, an independent statistical assessment of sensitivity and specificity is not possible. In contrast to the methods used so far, we propose an analytic method of validating predictions with respect to the temporal extent during which predicted events can occur. Moreover, the proposed method renders independent statistical control of true positives and true negatives possible.

III. STATISTICAL ASSESSMENT OF PREDICTIONS WITH INDEPENDENT CONTROL OF SENSITIVITY AND SPECIFICITY

In this section, we propose a statistics for evaluation of the performance of a prediction method based on the number of true positives and true negatives with respect to actually observed events. The statistics proposed is based on the idea of evaluating the performance of a prediction method by comparing it to the performance of a random predictor. Here, we assume a homogeneous Poissonian random predictor that raises alarms at a constant rate but independent of the actually occurring events. Adaptions to other types of random predictors are necessary if other null hypotheses are to be tested. In case, e.g., it is known that the events occur at a time-dependent rate such that the numbers of events vary throughout the day, month, or year, the random predictor needs to fulfill the same time dependence. Then the null hypothesis of performing quite as well as the prediction method investigated given that time dependence is tested. Such adaptions, however, do not change the idea of the subsequent statistics. For binary random predictions, the numbers of correct predictions, i.e., accordances of predictions and observations, are binomially distributed. Therefore, a threshold of the maximum number of correct predictions obtainable by chance can be determined from a binomial distribution. This leads to a binomial statistics, which can be adapted to different settings of testing the accordance of predictions and observations.

To assess the performance of the statistics proposed, prediction-observation pairs are simulated. These prediction-observation pairs mimic the series of prediction-observation pairs derived from any prediction method for which the performance should be evaluated. Since in the simulations the linkage of predictions and observations is known, the effectiveness of the statistics is investigated.

In the following, two prediction scenarios are investigated, illustrating the adaptivity of the binomial distribution ansatz. In the first scenario, a prediction method with an OP of one sample point and without an IP is simulated (Sec. III A). In the second scenario, IPs and OPs are simulated as extended periods in time (Sec. III B). For each scenario, the two steps of proposing the statistics and assessing its effectiveness are conducted.

A. Evaluating prediction-observation pairs without an IP and with a short OP

A prediction method performs significantly better than chance if it exceeds the performance of a random predictor given a certain error probability. The random predictor raises alarms randomly throughout time, independent of properties used in the investigated prediction method to predict the observed events. Given that the alarm rate is the same as the alarm rate of the prediction method investigated, the probability of a positive prediction before an event is

$$p_+ = \frac{n_a}{N}, \tag{6}$$

where n_a is the number of alarms of the prediction method and N is the number of data points of the prediction-observation time series. Both properties can be derived from the time series of predictions and observations.

Since the random predictor can be either right or wrong, the number of true positive predictions is binomially distributed. The probability for m_+ true positive predictions out of M_+ positive predictions is thus

$$\lambda_+ = \mathcal{B}(m_+, M_+; p_+) = \binom{m_+}{M_+} p_+^{m_+}(1-p_+)^{M_+-m_+}. \tag{7}$$

From this binomial distribution the quantiles are derived from the inverse of the cumulative binomial distribution Λ^{-1}. This leads to the critical value

$$m_{c+} = \Lambda^{-1}(1-\alpha_+, M_+; p_+), \tag{8}$$

which is the maximum number of true positives, obtainable by the random predictor with an error probability α_+. This error probability is the significance level.

The critical value m_{c-} of true negatives is obtained analogously, with the number of negative predictions $M_- = N - M_+$ and a negative-prediction probability,

$$p_- = (1-p_a), \tag{9}$$

instead of M_+ and p_+ in Eq. (8), and p_a the probability for an alarm. If desired, a different significance level α_- instead of α_+ can be used for control of the number of true negatives.

In order to assess the statistical properties of the investigated prediction method the following four-step algorithm has to be applied.

(1) Determine the significance level $\alpha_\pm$ for true positives and true negatives, respectively. This is the error probability of deciding that the prediction method is better at positive and negative predictions than chance, although it actually is not.

(2) Estimate the probability of a positive and negative prediction $p_\pm$ based on the number of positive and negative predictions n_a and $N - n_a$ in the prediction time series.

(3) Derive the critical values of true positives m_{c+} and true negatives m_{c-} from the binomial distributions, Eq. (8), with significance levels $\alpha_\pm$, numbers of positive and negative prediction $M_\pm$, and positive and negative prediction probabilities $p_\pm$, Eqs. (6) and (9).

(4) Compare the critical values $m_{c\pm}$ obtained in step 3 to the number of true positives m_+ and negatives m_- obtained from the prediction-observation time series as described in Sec. II A. If the number of true positives and true negatives exceeds the respective critical values, the prediction system performs significantly better than chance.

In the following subsections we present a simulation study. First, the simulation of data is presented. Second, the simulation study is described and results are presented.

1. Simulation of data

In order to test the statistics proposed in Sec. III A, we simulated prediction-observation time series as in [10]. To this end, first a time series of probabilistic forecasts f was drawn from a β distribution with density

$$s(f) = \frac{f^{v-1}(1-f)^{w-1}}{F(v,w)}, \tag{10}$$

where $F(v,w)$ is the β function and v and w parameters of the distribution. From the parameters the expected value of probabilistic forecasts can be derived, $\mu_f = \frac{v}{v+w}$. The β distribution is constrained to the range $[0,1]$ and, thus, can be thresholded to 0's and 1's. To obtain binary predictions a ($a = 1$, alarm, vs $a = 0$, no alarm), the forecasts f were thresholded such that the probability of an alarm was μ_f. To this end, the threshold was selected to be the inverse of the cumulative β distribution with parameters v and w evaluated at $1 - \mu_f$. This ensured the expected value of the forecast to be the same as for the prediction, such that $\mu_f = \mu_a$.

In the second step, the probability of an event, i.e., $x = 1$, after a prediction a was modeled by

$$\mu_{x|a} = ca + d, \tag{11}$$

a linear linkage c of predictions a to the probability of an event after a prediction $\mu_{x|a}$, with an offset $d = \mu_x(1-c)$ (Fig. 3), as in [10]. The offset d was chosen such that the expected value of an event was equal to the expected value of predictions, $\mu_x = \mu_a$. Thus, depending on whether an alarm was raised [$a = 1$; Fig. 3, solid (red) line] or not [$a = 0$; Fig. 3, dashed (blue) line], the probability of an event $\mu_{x|a}$ was a linear function of the linkage strength c. The observation x was obtained based on the prediction a by drawing a random number x' of a uniform distribution in the interval $[0,1]$ and comparing it to $\mu_{x|a}$, such that

$$x = \begin{cases} 1 & \text{if} \quad x' < \mu_{x|a} \text{ (event occurs)}, \\ 0 & \text{otherwise (no event occurs)}. \end{cases} \tag{12}$$

This way, prediction-observation pairs with known linkage could be generated. At high linkage the probability of an event after a positive prediction [solid (red) line] was high, while the probability of an event after a negative prediction [dashed (blue) line] was low. At low linkage strengths ($c \approx 0$) the probability of events was similar no matter whether a positive or negative prediction was made. While zero linkage thus refers to independence of predictions and observations, such that the null hypothesis is met, increasing linkage refers to an increasing probability of accordances of predictions and observations.

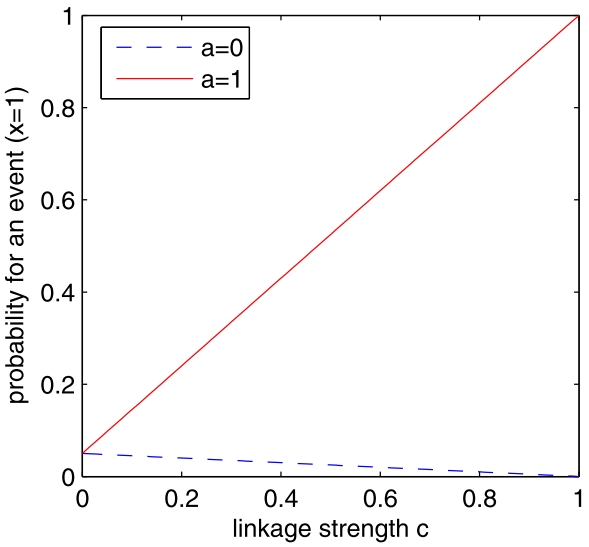

FIG. 3. (Color online) The prediction method simulated here is based on a linear function of the linkage c (x axis) of predictions a [solid (red) or dashed (blue) line] to the probability of an event $x = 1$ (y axis). While the probability of an event increases linearly with the linkage strength for positive predictions [$a = 1$; solid (red) line], it decreases for negative ones [$a = 0$; dashed (blue) line]; see Eq. 11.

2. Simulation study

To assess the performance of the proposed statistics, we simulated prediction-observation time series consisting of $N = 10\,000$ prediction-observation pairs as described in Sec. III A 1. Investigating, particularly, the case of few events, in which measures like the Brier score fail, we chose the probability of an event to be $\mu_x = \mu_a = 0.05$ with $v = 1$ and $w = 19$ (Sec. IIIA1). Linkage strengths were increased from no linkage, $c = 0$, to maximum linkage, $c = 1$. That way, it was known whether the null hypothesis or the alternative actually applied to the data. This yielded power analysis possible. Power refers to the probability of the statistics to reject the null hypothesis when actually the alternative is true. Applied to the prediction-observation pairs assessed here, the power is expected to increase with increasing linkage strength. On the other hand, coverage was investigated. Coverage refers to the probability to reject the null hypothesis when it actually is

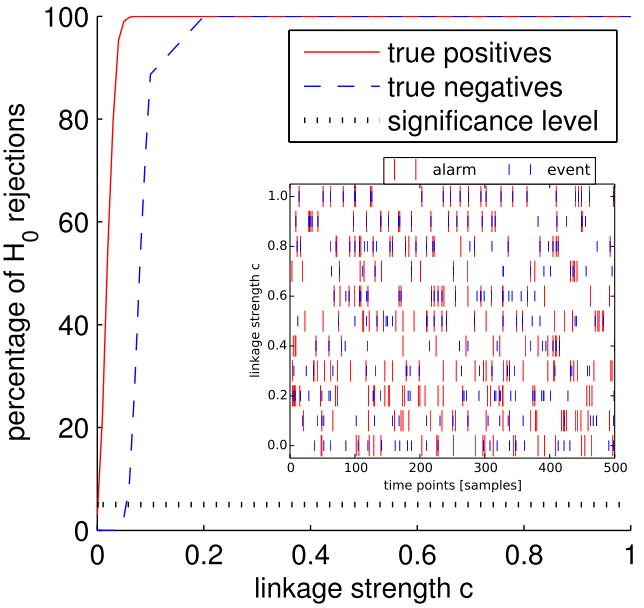

FIG. 4. (Color online) Percentage at which the simulated prediction-observation time series for linkage strengths c were statistically tested to be significantly better than chance. The solid (red) line shows the coverage ($c = 0$) and power ($c > 0$) of true positives. The dashed (blue) line shows the coverage and power of the true negatives. For this, 1000 repetitions of prediction-observation time series consisting of 10 000 data points were simulated for each linkage strength. Inset: The first 500 of these 10 000 data points. Positive predictions, i.e., alarms, are shown by long (red) vertical lines; positive observations, i.e., events, by short (blue) vertical lines.

true. To ensure that coverage is kept correct, the probability of erroneously rejecting the null hypothesis should not exceed the significance level. Coverage and power of the statistics lead to a reliable and powerful statistics.

For such power and coverage analysis, we simulated $K = 1000$ repetitions of prediction-observation time series, each $N = 10\,000$ data points long, for a set of linkage strengths $c \in [0,1]$. The first 500 of 10 000 data points of exemplary prediction-observation time series are shown in the inset in Fig. 4 for various linkage strengths (y axis). Positive predictions, i.e., alarms, are shown as long (red) vertical lines, while positive observations, i.e., events, are shown by short (blue) lines. While at zero linkage alarms and events occur independently, the number of overlapping alarms and events increases with increasing linkage strength. According to our four-step algorithm we statistically assessed the performance of the simulated prediction-observation pairs. The probability of the rejection of the null hypothesis of independence of prediction-observation pairs was then derived from the percentage of repetitions for which the null hypothesis of this independence was rejected.

In Fig. 4 the results of the power analysis for both true positives [solid (red) line] and true negatives [dashed (blue) line] are shown for linkage strengths $c \in [0,1]$. For absent coupling $c = 0$, the number of true positives and true negatives remains below the significance level $\alpha_\pm = 0.05$ (dotted black line). Therefore, the statistics is considered reliable. Since for true negatives, the percentage is below the significance level, the proposed statistics is conservative. With increasing linkage strength, the power of the statistical test proposed increases quickly, for both true positives and true negatives. This yields a powerful statistics.

As the simulations show, the statistics of the number of true predictions based on a binomal distribution is reliable and powerful for assessing the performance of a prediction method.

B. Evaluating prediction-observation pairs with an extended IP and OP

If the predictions are evaluated using the framework of temporally extended IPs and OPs, their durations T_o and T_i need to be incorporated into the statistics. The critical values for true positives (+) and true negatives (−) are derived from the inverse of a binomial cumulative distribution as before, Eq. (8), with different prediction-observation probabilities p_+ and p_-. As shown in Fig. 1, for a true positive prediction, no event must occur in the IP, but at least one event has to occur in the OP. This yields a probability

$$p_+ = (1 - p_a)^{T_i}(1 - (1 - p_a)^{T_o}) \tag{13}$$

of a true positive prediction. After a true positive prediction, the next prediction can be made after the end of the event.

Similarly to the true positive prediction, the probability of a true negative prediction can be derived. For a true negative prediction (Fig. 2), the occurrence of no event must be ensured within the IP. Therefore, given a negative observation, the probability of T_i negative predictions is $p_- = (1 - p_a)^{T_i}$. However, at the next time point the probability of a true negative is increased since it is already known from the previous time point that $T_i - 1$ of the next predictions are negative. Thus the probability of a true negative prediction at the next time point is

$$p_- = (1 - p_a). \tag{14}$$

In the case of few events, this is an approximation of a global probability of true negative predictions.

The probabilities of true positives p_+ and true negatives p_- describe the probability of the random predictor to predict the observations correctly. From the inverse binomial cumulative distribution, $\Lambda^{-1}(1 - \alpha_\pm, M_\pm; p_\pm)$, the critical values of true positives and true negatives can be derived as in Sec. III A. The four-step algorithm presented in Sec. III A, thus can be applied when substituting $M_\pm$ and $p_\pm$ by the values obtained when incorporated extended durations of the IP and OP as presented here.

1. Simulation of data

Similarly to Sec. IIIA1, we simulated prediction-observation time series based on β-distributed forecasts. In contrast to the simulations above, the observations were constructed from the thresholded forecasts such that the probability of an event was low during the IP and high during the OP after a positive prediction. Additionally, we introduced a refractory period (RP), which occurs after each event. This simulates the period after the event during which no predictions can be made because, e.g., the system investigated does not

provide reliable predictions during or in a transition time after the event since the measurements are still affected.

The parameters of the β distribution were changed to ensure that the probability of an event during a period as long as the IP-OP complex was lower than 10%. This yielded parameters $v = 1$, $w = 199$, and $\mu_f = 0.005$. Since now the OP was extended, the number of events was decreased under the null hypothesis of independent predictions and observations. To ensure $\mu_a = \mu_x$ as before, events were added randomly at instances of no events, whenever the number of alarms exceeded the number of events. This kept the null hypothesis correct, i.e., if no linkage was present. For increasing linkage, the number of events converges to the number of alarms due to linkage. In order to leave the exact number of added events flexible, we drew uniformly distributed random numbers between 0 and 1 for all time points at which no event occurred. The random numbers were thresholded at $\frac{\Delta}{\#\text{ no events}}$ such that the expected number of additional random events is the difference Δ in the numbers of alarms and events. As in Fig. 4, the first 500 of 10 000 data points of exemplary prediction-observation time series are shown for various linkage strengths in the inset in Fig. 5. As before, the correlation of alarms [long (red) vertical lines] and events [short (blue) vertical lines] increases with increasing linkage. As expected from the simulation design, the alarms tend to precede the events, in contrast to the simulation corresponding to the inset in Fig. 4. Furthermore, since the alarm rate is reduced in this simulation, fewer alarms and events occur.

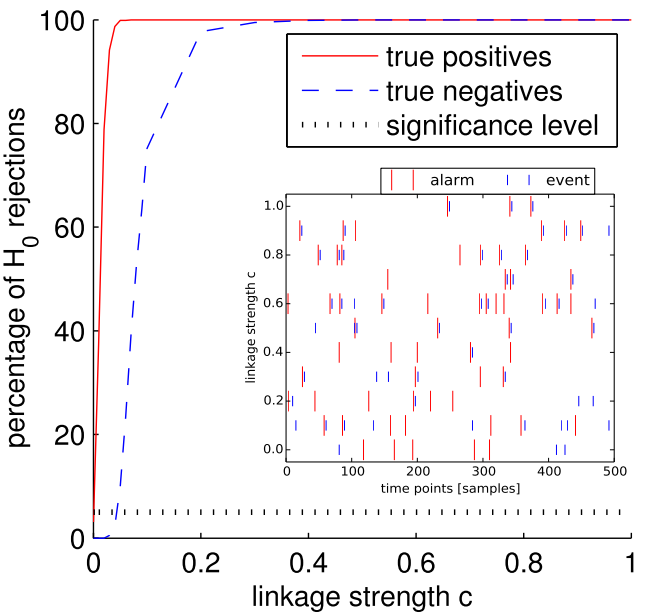

FIG. 5. (Color online) Percentage at which the simulated prediction-observation time series for linkage strengths c were statistically tested to be significantly better than chance. The solid (red) line shows the coverage ($c = 0$) and power ($c > 0$) of true positives. The dashed (blue) line shows the coverage and power of the true negatives. For this, 1000 repetitions of the prediction-observation time series consisting of 10 000 data points were simulated for each linkage strength with IP = 2, OP = 4, and RP = 2. Inset: The first 500 of these 10 000 data points. Positive predictions, i.e., alarms, are displayed by long (red) vertical lines; positive observations, i.e., events, by short (blue) vertical lines.

2. *Simulation study*

As described in Sec. IIIA2 we performed a power analysis based on 1000 repetitions of prediction-observation time series including IPs and OPs as described in Sec. IIIB1. We assumed the IP to be shorter than the OP, choosing $T_i = 2$ and $T_o = 4$ data points. In our simulations, we chose the RP to be as long as the IP, such that $T_r = 2$. The resulting power curve for linkages $c \in [0,1]$ is shown in Fig. 5.

Both true positives [solid (red) line] and true negatives [dashed (blue) line] kept the coverage at $c = 0$ correct. Again, the statistics is conservative for true negatives, since the percentage of rejections is lower than 5%. For increasing linkage strength, the power increases quickly for both true positives and true negatives. Thus, the statistics could be adapted to the case of extended IPs and OPs, remaining powerful while keeping the coverage correct.

IV. DISCUSSION

We have introduced a statistical method that evaluates the performance of binary forecasting systems by a random predictor. Unlike previous statistical methods, it allows independent evaluation of true positive and true negative predictions, even when extended phases such as the OP and IP are used. We proposed an algorithm for the statistical assessment of predictions obtained from a prediction method. This algorithm includes the proposed statistics of a random predictor. The random predictor is built such that the properties of prediction equal those of the investigated prediction method except for the linkage of prediction and observations. For this random predictor the statistics of obtaining true positive and true negative predictions by chance is given by a binomial distribution. From this distribution the quantiles of the random predictor can be derived and compared to the number of true positives and true negatives given by the prediction method. Here, we have described how critical values of true positives and true negatives can be obtained from the respective inverse of the cumulative binomial distribution. If the number of true predictions obtained from the prediction method exceeds the number of true predictions of the random predictor, the prediction method is considered significantly better than chance. Instead of computing critical values, p values can be obtained from the binomial cumulative distribution. Instead of computing the inverse of the cumulative binomial distribution evaluated at 1 minus the significance level α, the cumulative binomial distribution has to be evaluated at the number of true predictions made by the prediction method investigated.

In two sections, we have shown that our statistics is adaptive to frameworks based on occurrence and intervention periods. Furthermore, we have shown that the statistics proposed is powerful by testing the performance in controlled settings. Power is high to identify prediction methods which are linked to the observations such that predictions are better than chance. On the other hand, coverage is kept correct in cases in which the prediction method does not predict observations better than chance.

Our framework proposed here also implies the assessment of strengths of events, e.g., strengths of thunderstorms or earthquakes. To this end, the event strengths can be categorized into several groups. For each group the performance of true positives and true negatives can then be assessed with the statistics proposed. Alternatively, the binomial statistics can be adapted to multinomial instead of binomial distributions.

Furthermore, the binomial statistics proposed can be applied to the statistical assessment of event detection. In that case, the IP and OP can be replaced by the respective periods of interest. Adaptions of the statistics are then necessary. They are implementable analogously to those presented here.

As the simulations show, the proposed method is a reliable, powerful, adaptive, and widely applicable statistics for assessing the performance of prediction or detection methods, even in the case of rare events.

ACKNOWLEDGMENTS

This research was supported in part by the US National Institutes of Health through Grant No. R01NS065096, as well as E!5185 PANG.

[1] A. H. Fielding and J. F. Bell, Environ. Conserv. **24**, 38 (1997).
[2] A. H. Murphy, Weather Forecast. **6**, 302 (1991).
[3] C. Frei and C. Schär, J. Climate **14**, 1568 (2001).
[4] P. Lardet and C. Obled, J. Hydrol. **162**, 391 (1994).
[5] I. T. Jolliffe and D. B. Stephenson (eds.), *Forecast Verification: A Practitioner's Guide in Atmospheric Science*, 2nd ed. (Wiley, West Sussex, UK, 2012).
[6] F. Mormann, R. G. Andrzejak, C. E. Elger, and K. Lehnertz, Brain **130**, 314 (2007).
[7] B. Schelter, J. Timmer, and A. Schulze-Bonhage (eds.), *Seizure Prediction in Epilepsy: From Basic Mechanisms to Clinical Applications* (Wiley, Weinheim, Germany, 2008), p. 334.
[8] H. Feldwisch-Drentrup, B. Schelter, M. Jachan, J. Nawrath, J. Timmer, and A. Schulze-Bonhage, Epilepsia **51**, 1598 (2010).
[9] C. Marzban, Weather Forecast. **13**, 753 (1998).
[10] A. A. Bradley, S. S. Schwartz, and T. Hashino, Weather Forecast. **23**, 992 (2008).
[11] M. Jachan, H. Feldwisch-Drentrup, F. Posdziech, A. Brandt, D.-M. Altenmüller, A. Schulze-Bonhage, J. Timmer, and B. Schelter, in *4th European Conference of the International Federation for Medical and Biological Engineering*, IFMBE Proceedings, Vol. 22, edited by J. V. Sloten, P. Verdonck, M. Nyssen, and J. Haueisen (Springer, Berlin, 2009), pp. 1701–1705.
[12] C. A. Doswell III, R. Davies-Jones, and D. L. Keller, Weather Forecast. **5**, 576 (1990).
[13] 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).
[14] G. W. Brier, Monthly Weather Rev. **78**, 1 (1950).
[15] R. A. Fisher, J. Roy. Stat. Soc. **85**, 87 (1922).
[16] C. Marzban, J. Appl. Meteorol. **37**, 72 (1998).
[17] D. S. Wilks, in *International Geophysics Series: Statistical Methods in the Atmospheric Sciences*, *Vol. 91*, 2nd ed., edited by R. Dmowska, D. Hartmann, and H. T. Rossby (Academic Press, Elsevier, Burlington, VT, 2005), p. 648.
[18] T. Maiwald, M. Winterhalder, R. Aschenbrenner-Scheibe, H. U. Voss, A. Schulze-Bonhage, and J. Timmer, Physica D **194**, 357 (2004).
[19] B. Schelter, M. Winterhalder, T. Maiwald, A. Brandt, A. Schad, A. Schulze-Bonhage, and J. Timmer, Chaos **16**, 013108 (2006).
[20] M. Winterhalder, T. Maiwald, H. U. Voss, R. Aschenbrenner-Scheibe, J. Timmer, and A. Schulze-Bonhage, Epilepsy Behav. **4**, 318 (2003).
[21] I. Campbell, Stat. Med. **26**, 3661 (2007).