Using missing ordinal patterns to detect nonlinearity in time series data

Christopher W. Kulp
The Department of Astronomy and Physics, Lycoming College, Williamsport, Pennsylvania 17701, USA

Luciano Zunino
Centro de Investigaciones Ópticas (CONICET La Plata–CIC), C.C. 3, 1897 Gonnet, Argentina
and Departamento de Ciencias Básicas, Facultad de Ingeniería, Universidad Nacional de La Plata (UNLP), 1900 La Plata, Argentina

Thomas Osborne and Brianna Zawadzki
The Department of Astronomy and Physics, Lycoming College, Williamsport, Pennsylvania 17701, USA
(Received 14 June 2017; published 28 August 2017)

The number of missing ordinal patterns (NMP) is the number of ordinal patterns that do not appear in a series after it has been symbolized using the Bandt and Pompe methodology. In this paper, the NMP is demonstrated as a test for nonlinearity using a surrogate framework in order to see if the NMP for a series is statistically different from the NMP of iterative amplitude adjusted Fourier transform (IAAFT) surrogates. It is found that the NMP works well as a test statistic for nonlinearity, even in the cases of very short time series. Both model and experimental time series are used to demonstrate the efficacy of the NMP as a test for nonlinearity.

DOI: 10.1103/PhysRevE.96.022218

I. INTRODUCTION

Characterizing the dynamics of a system from time series data is an important problem in time series analysis. Properly characterizing the dynamics of a system requires the ability to distinguish between both determinism and stochasticity as well as linear and nonlinear dynamics. Sometimes it is useful to pair a test for nonlinearity with a test for determinism. For example, the 0 - 1 test for chaos [1] can be used to distinguish chaotic dynamics from regular dynamics, as long as the time series is measured from a deterministic series. In [2], it was shown that pairing the 0 - 1 test with a test for determinism can prevent false positives for chaotic dynamics. In this paper, we demonstrate the use of a test statistic, called the *number of missing ordinal patterns*, which can be used to test for both determinism and nonlinearity in model data and nonlinearity in experimental data.

There are many tests in the literature for both determinism and nonlinearity. Some tests for determinism, such as those presented in Refs. [3,4], use symbol spectra to identify recurring patterns in symbolized time series data. While symbol spectra-based methods work well in discriminating determinism from stochasticity, it has been the authors' experience that they can be cumbersome to use when analyzing a large number of series. Other methods for detecting determinism, such as noise titration [5,6], involves a technique for detecting nonlinearity and adding noise to the series until the nonlinearity can no longer be detected. The noise titration technique is mainly focused on detecting low-dimensional chaos, however, the presence of low-dimensional chaos would then imply determinism. Noise titration has also been shown to struggle with series which are contaminated with colored noise [7] and more complicated deterministic systems [8]. One of the most common tests for nonlinearity involves the generation of surrogate time series from the series to be tested. We will discuss surrogate methods in more detail below. Additional methods for detecting nonlinearities include using a "representation space" [9], "permutation slopes" [10], and multiscale symbolic approaches [11]. The aforementioned list of tests for determinism and nonlinearity is not intended to be comprehensive, but rather, is intended to give the interested reader a brief survey of some of the methods in the literature developed over the last 20 years.

One of the simplest tests for determinism involves symbolization of the time series using the Bandt and Pompe (BP) methodology [12]. The BP methodology partitions the series into embedding delay vectors, similar to phase space reconstruction, and then maps each of the partitions into a permutation of the set $\{1,2,\ldots,D\}$, where D is called the embedding dimension, and is not necessarily the same as the embedding dimension used in phase space reconstruction. The result is a series with $D!$ possible symbols called ordinal patterns. In a stochastic series, all possible ordinal patterns will occur, if the series is long enough. However, certain ordinal patterns will not occur in deterministic series. Those patterns are called forbidden patterns and correspond to states inaccessible to the dynamics governing the system. Hence the presence of forbidden patterns is a sign of determinism [13,14]. Using forbidden patterns is an easy method of detecting determinism because all one needs to do is simply count the number of different patterns that occur in the series. The number of forbidden ordinal patterns (NFP) is then $D!$ minus the number of occurring patterns. A nonzero value for the NFP means that the system is deterministic. The NFP has been shown to be an effective measure for determinism, even in irregularly sampled time series [15–17].

One challenge to using the NFP as a means of detecting determinism is that it requires time series that are long enough to sample all possible ordinal patterns. Essentially, the length of the series N should satisfy $N \gg D!$, which can be an issue for series requiring large embedding dimensions, such as $D = 7$ or $D = 8$. While model data sets can be as long as one has the computational power and time to produce, experimental data is often short and may not be long enough to sample all possible patterns. Furthermore, correlations in the time series may also require longer data sets in order to sample

all possible ordinal patterns [18]. The reason for the increased length needed for correlated series is because in correlated stochastic series, some ordinal patterns are more probable than others due to the correlations. Whereas in random series, all ordinal patterns are equally probable. The difference in probabilities in correlated stochastic series means that longer series are needed in order for all ordinal patterns to be realized.

Ultimately, the question is that if the number of forbidden ordinal patterns $N_{\mathrm{FP}} \neq 0$ for short series, how confident can we be that the system is deterministic as opposed to the nonzero NFP being due to small sample effects? The ordinal patterns that are not present in the time series due to small sample effects are referred to as *missing ordinal patterns*. In other words, how can one be sure missing patterns are actually forbidden patterns and are those missing patterns a useful means of detecting nonlinearity?

In this paper, we will address the issue of using the number of missing ordinal patterns (NMP) to detect determinism and nonlinearity in time series data. This will be done by using the NMP in a standard surrogate framework [19] (and references therein). For each series to be studied, we produce a particular number of iterative amplitude adjusted Fourier transform (IAAFT) surrogates which will have the same power spectrum and amplitude probability distribution as the original series. Each series (original and all surrogates) will then be symbolized using the BP methodology and the NMP will be calculated. If the NMP of the original series is outside of the range of the NMP of the surrogates, then the null hypothesis (that the data is a rescaled Gaussian linear stochastic process) has been rejected up to a confidence level determined by the number of surrogates produced (see below). Through this analysis we will show that the NMP for nonlinear series will be statistically significantly different from those of the surrogates, even in very short time series data (where the condition $N \gg D!$ is not upheld). Furthermore, the NMP can be used as a test for determinism by finding whether or not the NMP of the series converges to that of the surrogates as the series length increases. To demonstrate the efficacy of the analysis, both model and experimental time series will be studied.

The work presented here was motivated by issues arising through the practice of using the BP methodology on real-world data. Part of the motivation was to better understand how to choose an embedding dimension D in order to avoid small sample effects but still have a large enough embedding dimension in order to capture the necessary dynamics. In past experience, the authors have found that the NMP can be zero for high dimensional deterministic systems if D is chosen to be too small. This is believed to be due to projecting the dynamics onto too small of a subspace. However, if an embedding dimension of say, 7, needs to be used to capture the appropriate dynamics but the time series is less than 1000 elements long, how can we be certain that the nonzero NMP is indicative of determinism and/or nonlinearity and not finite size effects?

The rest of this paper will be structured as follows. In Sec. II the BP methodology will be discussed as will be the procedure for finding the number of missing ordinal patterns. Section III contains a brief discussion of the IAAFT surrogates. Sections IV and V present the results of our analysis on model and experimental time series, respectively. Finally, we make concluding remarks in Sec. VI.

II. MISSING AND FORBIDDEN ORDINAL PATTERNS

The Bandt and Pompe (BP) method [12] for symbolizing time series has been widely used in a variety of applications as an alternative means to threshold-based symbolization techniques. There are many advantages of the BP methodology including its simplicity, speed, and noise robustness. Additionally, the method is invariant to nonlinear monotonous transformations. As opposed to using an amplitude-based threshold, the BP methodology is based on partitioning the series and then using the relative amplitude of the values in the partition to produce the symbol for the partition.

As a demonstration of the BP methodology, consider the series $\{x_1,x_2,\ldots,x_N\}$. The first step is to create embedding vectors similar to those used in phase space reconstruction. An embedding delay τ and dimension D must be chosen. The embedding delay can be chosen using methods such as the first zero crossing of the autocorrelation or the first minimum of the mutual information. However, the authors have had success in past work (see, for example, [4]) simply by choosing $\tau = 1$. In this paper, we will use $\tau = 1$ and we will see that we will get good results with that choice. Using $\tau = 1$ can be very convenient when analyzing a large number of data sets where computing a value of τ for each series can be prohibitive. The choice of D has been traditionally made keeping in mind that $D! \ll N$. However, as discussed in Sec. I, such a restriction on D can be problematic for short series. In this paper, we will perform our analysis on several different values of D to show that the NMP can be an effective measure even for short series analyzed with large values of D. It is worth noting that Bandt and Pompe [12] suggest that $\tau = 1$ and $3 \leqslant D \leqslant 7$ be chosen for the purposes of computational efficiency and practicality.

Once a value for τ and D are chosen, the next step is to create ordinal patterns from the embedding vectors. This is done by mapping each vector to a permutation of the set $\{1,2,\ldots,D\}$ by using the rank of each value in the sequence. For example, the sequence $\{10,5,11,13\}$ would be mapped to $\{2,1,3,4\}$ because $x_2 < x_1 < x_3 < x_4$. In the case where two or more elements are equal, the element with the lowest index would come first in the permutation. For example, $\{5,10,11,10\}$ would map to $\{1,2,4,3\}$.

The measure used in this paper as a test statistic is called the number of missing ordinal patterns (NMP). Missing patterns are ordinal patterns that do not occur in the ordinal pattern series. If the series is long enough to sample all possible patterns ($N \gg D!$), and there are still ordinal patterns that do not occur in the series, then those patterns are referred to as forbidden patterns. Deterministic series will have forbidden patterns [13,14] which represent states inaccessible to the dynamics. However, for short series, one cannot be certain whether a particular set of patterns are forbidden or only missing due to small sample effects. The number of missing ordinal patterns is found by counting the number of observed patterns (NOP), simply the number of different ordinal patterns appearing in the symbolized series, and computing, $N_{\mathrm{MP}} = D! - N_{\mathrm{OP}}$. Because we will be comparing the NMPs of a system using a variety of embedding dimensions, we will

measure the ratio of the NMP to the number of possible patterns, $R_{\mathrm{MP}} = N_{\mathrm{MP}}/D!$, as was done in Ref. [16].

III. SURROGATE TIME SERIES

As mentioned in Sec. I, the goals of this work are to test the reliability of the number of missing patterns (NMP) as a metric for nonlinearity in time series and to better understand how to use NMP to detect determinism in short time series. In this case, short time series means that the length of the series N does not satisfy $N \gg D!$. Although a nonzero NMP value for a short series is not a guarantee for determinism (the patterns could simply be missing due to small sample effects), it is still possible that the number of missing patterns can be used as a measure of nonlinearity.

To address the above issue, we can compare the NMP of a series (the one being analyzed) to the NMP of a random series, called a surrogate, with the same length. In this case, using the same values of τ and D for the original and random series. The simplest way to generate a random series would be to generate a list of pseudorandom numbers with the same length as the series to be analyzed. Then one could compute the NMP of the pseudorandom list and compare that to the original series. Of course for better results, one could generate an ensemble of pseudorandom series and find the range of NMPs for the ensemble. The result would be a type of confidence interval for how significantly different the NMP of the original series is to the ensemble, the idea being that if all of the pseudorandom series had an $N_{\mathrm{MP}} = 0$, then one could be confident that the choice of D was appropriate for the original series and a nonzero NMP would not be due to small sample effects. The problem with this technique, however, is that other than length, the pseudorandom series shares nothing in common with the original series. Hence such a comparison might not be appropriate.

A more sophisticated surrogate would involve a random shuffling of the series before symbolization. In this case, the amplitude probability distribution of the surrogate would match that of the original series. However, the temporal relationship between the values will be destroyed. This means that any linear correlations or other linear properties (such as the power spectrum) of the series will not be preserved in the surrogate. It has been the authors experience that strongly correlated stochastic processes, such as fractional Gaussian noise (fGn) with a large Hurst exponent, require longer series to sample all possible ordinal patterns than independent and identically distributed (iid) sequences. Hence, a nonzero NMP for the series may result if the series is stochastic with strong linear correlations.

Therefore, an even more sophisticated surrogate generating method is needed, one that preserves the power spectrum of the original series. One of the earliest methods for this type of surrogate was introduced by [20]. By preserving the original series' power spectrum, the surrogate will also possess the same linear correlations. In addition to preserving the power spectrum, it is possible to preserve the amplitude probability distribution. The result is the so-called iterative amplitude adjusted Fourier transform (IAAFT) scheme which produces surrogates with the same amplitude probability distribution and power spectrum with potential higher order correlations being randomized. The IAAFT was introduced in Ref. [21], and the interested reader is directed there and to Ref. [19] for more details on the algorithm. Modifications to the IAAFT method have been developed [22]. Furthermore, a surrogate method specifically directed towards series with a strong periodic component has also been developed [23,24]. Finally, it is worth remarking that high frequency components are spuriously added to the surrogates if there is a mismatch in the beginning and end of the series (in both value and derivative). Consequently, an end-to-end mismatch criterion needs to be satisfied to avoid false rejections of the null hypothesis [19]. In this paper, IAAFT surrogates, as presented in Ref. [21] will be used and computed via the TISEAN surrogates algorithm [25].

When working with surrogates, it is necessary to identify a null hypothesis. With the IAAFT surrogates the null hypothesis is that the original data is a rescaled Gaussian linear stochastic process. If the null hypothesis is rejected, then the system is considered to be nonlinear with a confidence level α. The confidence level is determined by the number of surrogates M, analyzed using the relationship $M = 2/\alpha - 1$. The decision to accept or reject the null hypothesis is made by choosing a value of τ and D and computing the R_{MP} of the IAAFT surrogates and the original series. If the R_{MP} of the series lies outside the range of R_{MPs} obtained for the surrogates, then the null hypothesis has been rejected with a significance level α, and the series is determined to be nonlinear. In the language of surrogate testing, the R_{MP} is the test statistic that is compared to the original series and its IAAFT surrogates through a two-sided rank order test of size α.

In the next two sections, the surrogate analysis is applied to model and experimental data in order to demonstrate the efficacy of the analysis on time series data.

IV. APPLICATION TO MODEL SYSTEMS

In this section, the R_{MP} surrogate analysis is performed on several model systems, an AR(1) process, a fractional Gaussian noise (fGn) process, the Lorenz equations, and a nonlinear correlated stochastic process. To demonstrate the efficacy of the analysis, it must be shown that the test does not reject the null hypothesis of a linear stochastic system AR(1) and for a linearly correlated stochastic system (fGn). In addition it must also be shown that the test rejects the null hypothesis for nonlinear systems (Lorenz equations and the nonlinear stochastic correlated process).

A. AR(1) process

The first model system studied is a first-order autoregressive process AR(1), $x_i = ax_{i-1} + \epsilon_i$, where ϵ are pseudorandom values from a standard normal distribution and the parameter $a \in \{0.05, 0.1, \ldots, 0.95\}$. This system is studied in order to see if the R_{MP} analysis correctly fails to reject the null hypothesis for a linear stochastic system.

One hundred realizations of the AR(1) process were produced for each value of a. Each realization consisted of $N = 1024$ data points after discarding 10 000 previous elements in order to discard transients. Then, for each realization, $M = 199$ surrogates were generated, producing a confidence level, $\alpha = 0.01$. The fraction q of rejections of the

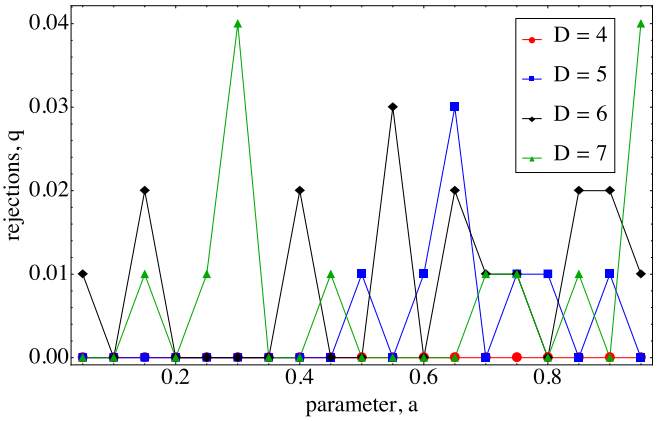

FIG. 1. The fraction of rejections q as a function of parameter a for an AR(1) process for several embedding dimensions D and $\tau = 1$. The graph is made using $M = 199$ surrogates producing a significance level of $\alpha = 0.01$.

null hypothesis were found by counting the number of times the R_{MP} of the series was outside the range of those of the surrogates. The results are shown in Fig. 1.

Notice that in Fig. 1, the value of q is very low, less than 4% for all embedding dimensions. Hence, the R_{MP} analysis correctly fails to reject the null hypothesis at worst, 96% of the time in the case of this linear stochastic system. Notice, however, that for $D = 4$ the test fails to reject the null hypothesis 100% of the time. The difference in performance for the different values of D is due to the short length of the series, $N = 1024$. Next, a linearly correlated stochastic process is studied.

B. Fractional Gaussian noise

A fractional Gaussian noise (fGn) process is a stochastic process for which the null hypothesis should not be rejected. Fractional Gaussian noise is derived from fractional Brownian motion (fBm), a generalization of Brownian motion, which is a linearly correlated stochastic process. The degree of correlation is measured by the Hurst exponent, $H \in (0,1)$. If $H = 0.5$, the fBm processes is actually Brownian motion, whereas if $H < 0.5$, then the process is negatively correlated, and if $H > 0.5$, then the process is positively correlated. Fractional Gaussian noise is the difference in successive values of a fBm process. In other words, if $f(H) = \{x_1, x_2, \ldots x_{N+1}\}$ is a fBm process with Hurst exponent H, then the fGn process is $g(H) = \{x_2 - x_1, x_3 - x_2, \ldots, x_{N+1} - x_N\}$. Fractional Gaussian noise processes are used in this work instead of fractional Brownian motion because fBm processes are nonstationary, unlike fGn processes which are stationary. Surrogate realizations for nonstationary processes are difficult to obtain, hence the focus on fGn processes. In the work presented here, fGn processes were produced using MATHEMATICA's FractionalGaussianNoiseProcess command.

In this work, $H = 0.9$ was used to generate the fGn process. When $H = 0.9$, there is a very strong positive linear correlation in the time series. It has been the authors' experience that fGn processes with $H = 0.9$ require the most amount of data in order to sample all possible ordinal patterns. Therefore, the fGn analysis begins with $H = 0.9$ and should serve as a difficult test case on which to apply our analysis.

To analyze the fGn data, series of length N were produced from $N = 50$ to $N = 1000$ in steps of 50. Then $M = 199$ surrogates were produced for each series of length N, providing a significance level of $\alpha = 0.01$. We then used the BP methodology to symbolize each series and each surrogate using $\tau = 1$ and $D = 4$, 5, 6, and 7. The results are displayed in Fig. 2. Each graph consists of a box and whisker plot and a scatter plot, which provides a clear visual demonstration of whether or not the R_{MP} of the series falls within the range of R_{MPs} of the surrogates. The box and whisker plot shows the range of values of the R_{MP} for the IAAFT surrogates and the scatter plot (blue dot) represents the R_{MP} of the original series.

Notice that in Fig. 2, for $D = 4$ (a) to $D = 6$ (c), the R_{MP} for the series is consistent with that of the surrogate at all lengths. In particular, for $D = 4$, the R_{MP} quickly reaches 0. Hence, the test is correctly failing to reject the null hypothesis at level of $\alpha = 0.01$ for $4 \leqslant D \leqslant 6$. The same is generally true for $D = 7$ [Fig. 2(d)], except for $N = 350$, 850, and 950 and in each case, the R_{MP} of the series is less than that of the range of the surrogates. For example, for $N = 950$, the R_{MP} for the fGn process is 0.8341 and the surrogates have R_{MP} values ranging from 0.8377 to 0.8474. The percent difference between the series R_{MP} and the surrogate minimum is 0.43%, which is negligible. However, the result suggests that the series are too short to be analyzed reliably with $D = 7$.

In addition to the analysis for $H = 0.9$, multiple realizations of fGn processes with a variety of Hurst exponents were analyzed. For $H = \{0.1, 0.2, \ldots, 0.9\}$, 100 realizations were produced, each with $N = 1000$ elements. For each realization, $M = 199$ surrogates were generated for a confidence level of $\alpha = 0.01$. The NMP analysis was then performed and the fraction of null hypothesis rejections q for each value of H were found. The results of this study are shown in Fig. 3. We see that for all values of H, the number of rejections remains quite small, demonstrating that the NMP analysis again appropriately fails to reject the null hypothesis of a linear stochastic system.

Next, data generated from the Lorenz equations are analyzed in order to understand under what conditions the null hypothesis is correctly rejected. Furthermore, the Lorenz system can demonstrate the use of NMP to identify not only nonlinear but also deterministic dynamics.

C. The Lorenz equations

The Lorenz equations,

$$\dot{x} = \sigma(y - x), \quad \dot{y} = x(\rho - z), \quad \dot{z} = xy - \gamma z, \qquad (1)$$

are well known to be an example of a nonlinear deterministic system. Here, initial conditions $(x_0, y_0, z_0) = (1,5,10)$ and parameter values $(\sigma, \rho, \gamma) = (10, 28, 8/3)$, for the chaotic case, and $(\sigma, \rho, \gamma) = (100, 200, 8/3)$ for the periodic case were used. Equation (1) was numerically solved using MATHEMATICA's NDSolve command and the solution was sampled using sampling periods of $\Delta t = 0.15$ for the chaotic case and $\Delta t = 0.01$ for the periodic case. In each case, we allowed transients to decay before sampling the solution and we

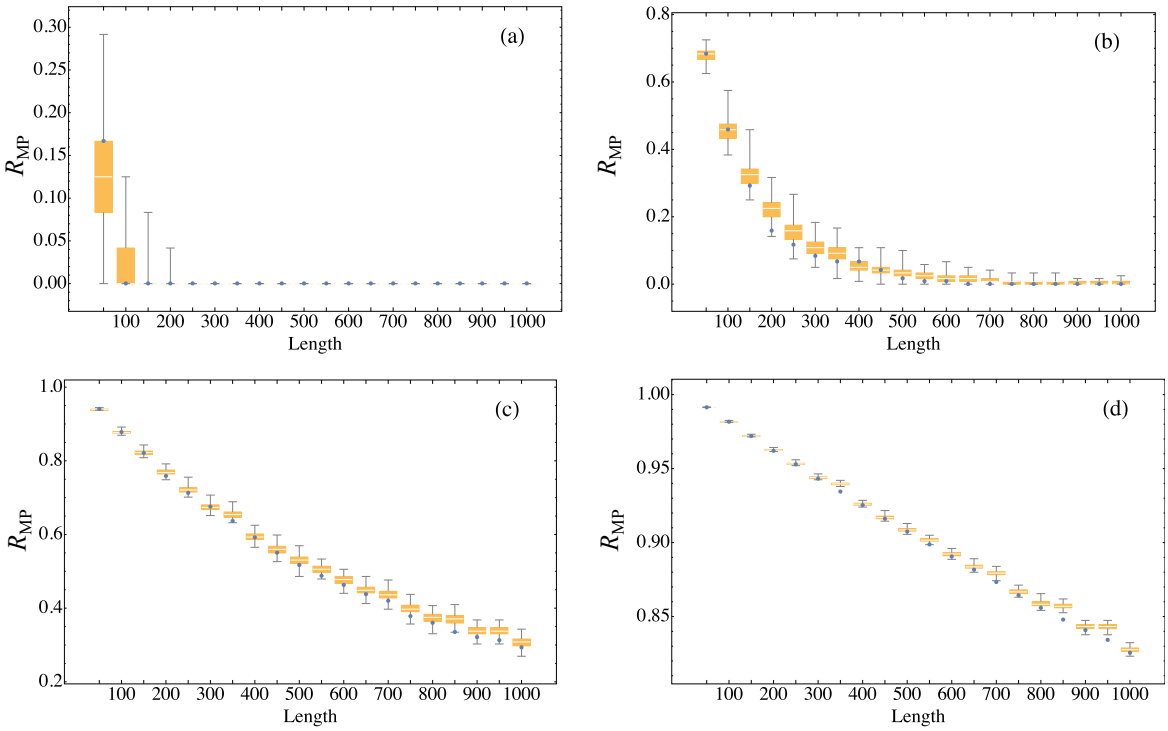

FIG. 2. Results of the surrogate analysis of an fGn process with $H = 0.9$ using $M = 199$ surrogates producing a significance level of $\alpha = 0.01$. The box and whisker plot shows the range of values of the R_{MP} for the IAAFT surrogates and the scatter plot (blue dot) represents the R_{MP} of the original series. Results for embedding dimensions $D = 4$ (a), 5 (b), 6 (c), and 7 (d), and $\tau = 1$ have been included.

produced time series of various lengths starting with $N = 50$ and ending with $N = 1000$ in steps of 50 elements.

For each series, $M = 199$ IAAFT surrogates were generated for a significance level of $\alpha = 0.01$. The BP methodology was then used to symbolize the series with $\tau = 1$ and $D = 4$, 5, 6, and 7. The results are displayed in Fig. 4 for the periodic case and in Fig. 5 for the chaotic case; both are similar to Fig. 2 in format.

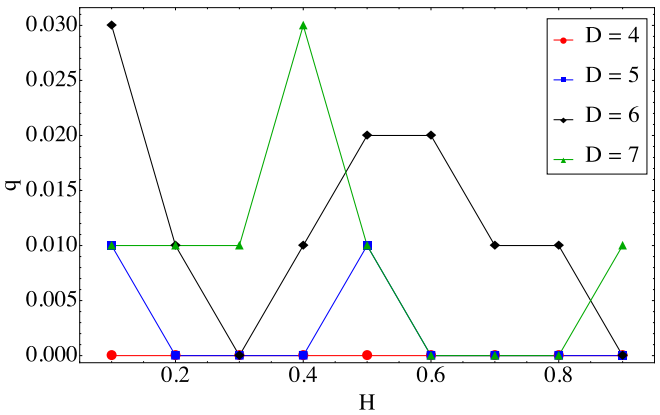

FIG. 3. The fraction of rejections of the null hypothesis q for fGn systems as a function of the Hurst exponent. One hundred realizations with $N = 1000$ for each value of H were produced, and for each realization $M = 199$ surrogates were generated leading to a significance level of $\alpha = 0.01$. Results for embedding dimensions $D = 4$, 5, 6, and 7, and $\tau = 1$ have been included.

In Fig. 4, where the periodic case was studied, the R_{MP} of the series is greater than that of the surrogates for all lengths of data and all embedding dimensions. Hence, the R_{MP} is able to reliably reject the null hypothesis for the periodic Lorenz equations, even in the cases where $N \ll D!$.

In Fig. 5, where the chaotic case was studied, the R_{MP} of the original series is outside the range of values for the surrogates for all data lengths except for $N = 50$ in the cases of $D = 6$ and $D = 7$ [Figs. 5(c) and 5(d), respectively]. This results tells us that the R_{MP} of the original series is statistically significantly different for almost all values of N and D, suggesting that although the surrogates have missing ordinal patterns (missing not forbidden; missing patterns simply have not been sampled, yet), the number of missing ordinal patterns in the surrogates are significantly fewer than those of the original series. Hence, the NMP is capable of rejecting the null hypothesis even in very short data sets for both chaotic and periodic Lorenz series.

In addition, the Lorenz equations were studied for very long series, $N = 100\,000$ using the same values of Δt and M as above and the results are shown in Fig. 6. The point of repeating the NMP analysis on long series is to demonstrate the presence of *persistent missing patterns*. The NMP remains nonzero for long series in both cases. The presence of persistent NMPs suggest that some of the missing patterns are, in fact, forbidden, and therefore, providing evidence for determinism.

Although not shown here for the sake of brevity, we also studied the chaotic Lorenz equation in the case of a sampling time of $\Delta t = 1.0$, where the series was significantly undersampled. It was found that the R_{MP} analysis did not reject the null hypothesis for $D = 4$ and 5 and $N \leqslant 1000$. However, the null hypothesis was rejected for $D = 6$ and 7

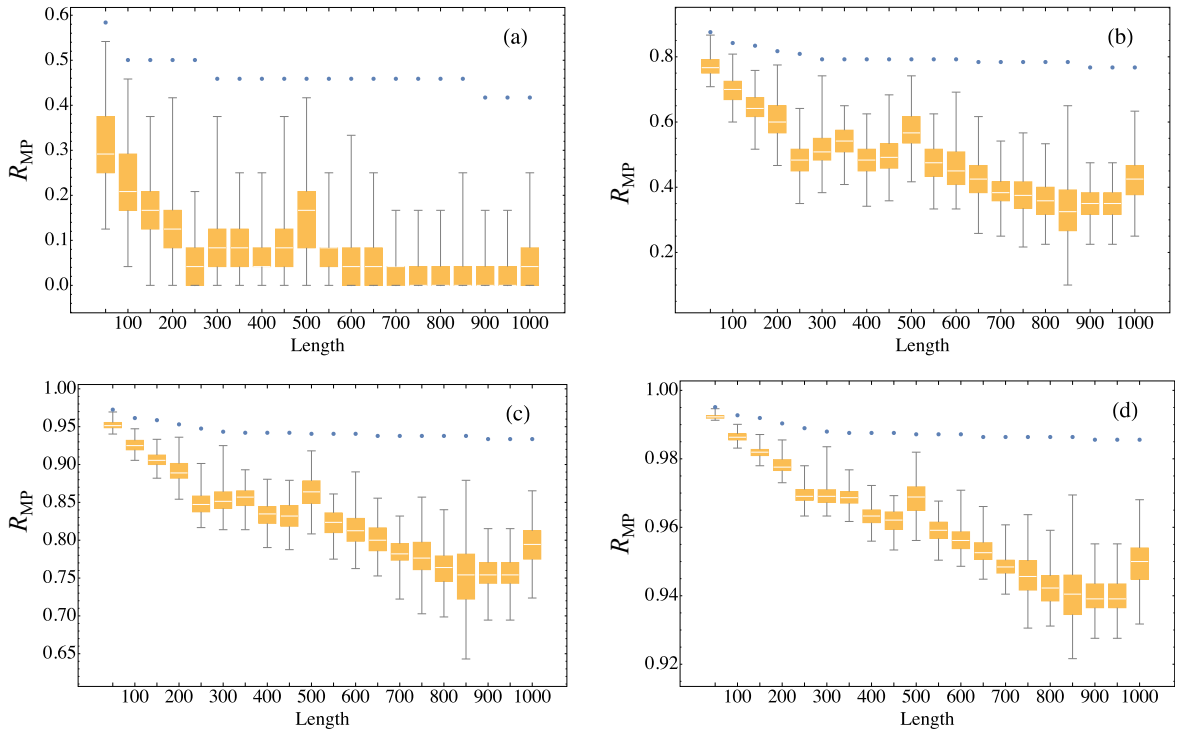

FIG. 4. Results of the surrogate analysis of a periodic series generated from (1) using $M = 199$ surrogates producing a significance level of $\alpha = 0.01$. The box and whisker plot shows the range of values of the R_{MP} for the IAAFT surrogates and the scatter plot (blue dot) represents the R_{MP} of the original series. Results for embedding dimensions $D = 4$ (a), 5 (b), 6 (c), and 7 (d), and $\tau = 1$ have been included.

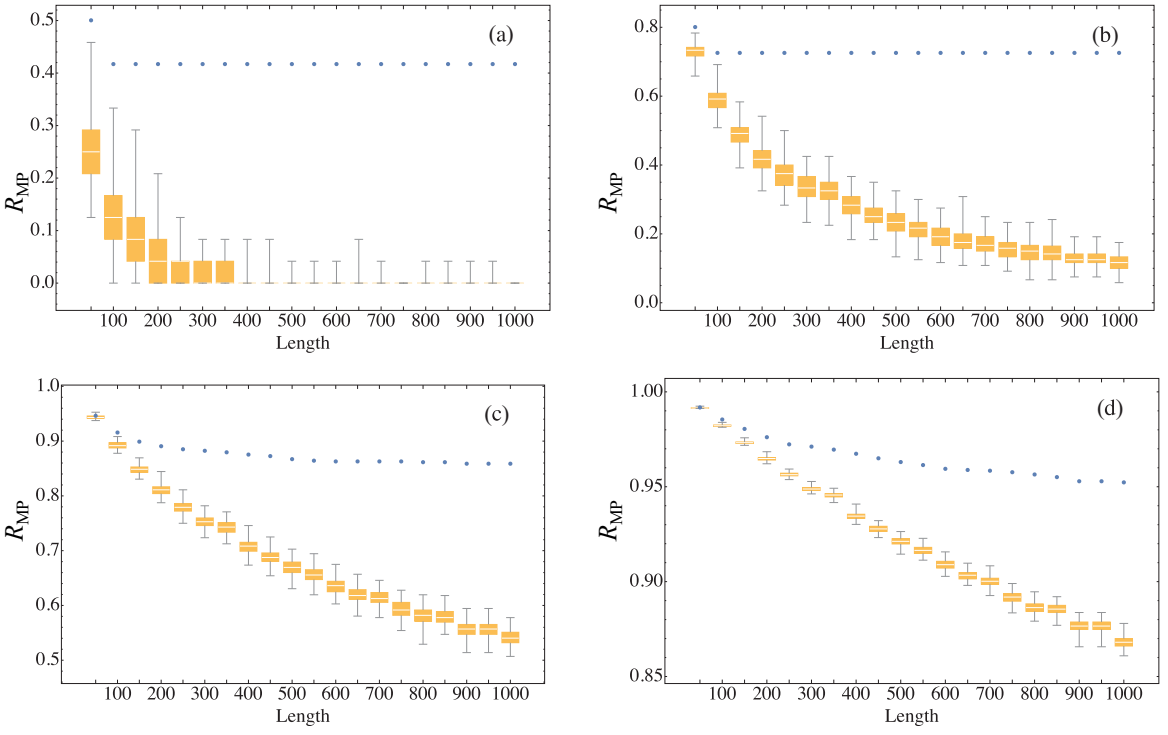

FIG. 5. Results of the surrogate analysis of a chaotic series generated from (1) using $M = 199$ surrogates producing a significance level of $\alpha = 0.01$. The box and whisker plot shows the range of values of the R_{MP} for the IAAFT surrogates and the scatter plot (blue dot) represents the R_{MP} of the original series. Results for embedding dimensions $D = 4$ (a), 5 (b), 6 (c), and 7 (d), and $\tau = 1$ have been included.

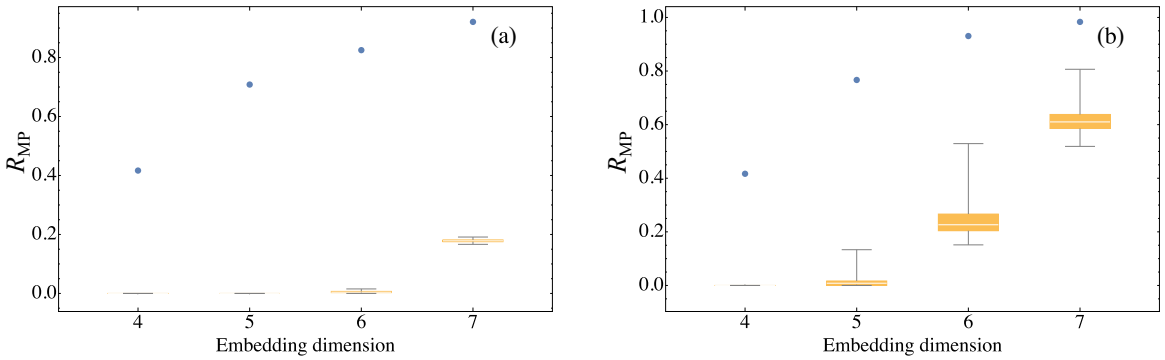

FIG. 6. Results of the surrogate analysis of (1) for (a) chaotic parameters and (b) periodic parameters using $N = 100\,000$ with $M = 199$ surrogates producing a significance level of $\alpha = 0.01$. The box and whisker plot shows the range of values of the R_{MP} for the IAAFT surrogates and the scatter plot (blue dot) represents the R_{MP} of the original series. Results for embedding dimensions $D = 4$, 5, 6, and 7, and $\tau = 1$ have been included.

when $N > 500$. This is suggesting that nonlinearity can be detected for short undersampled series if the embedding dimension is high enough. However, the R_{MP} was equal to zero for all embedding dimensions when $N = 100\,000$, suggesting that undersampling was equivalent to producing a stochastic series, as was expected due to the aliasing effect related to undersampling. A more thorough study of how the sampling time, especially oversampling, affects the R_{MP} analysis is needed for continuous systems.

D. Stochastic nonlinear correlated system

The final model system studied is a stochastic nonlinear correlated system. An example of such a system is

$$x_{k+1} = a\nu_k + b\nu_{k-1}(1 - \nu_k), \tag{2}$$

where ν is a uniformly independent identically distributed (iid) random variable between 0 and 1, $a = 3$, and $b = 4$.

For a discussion on the stochasticity of this system, the interested reader is directed to [7]. The goal of applying the R_{MP} surrogate analysis is to see whether or not the analysis can detect nonlinearity (by rejecting the null hypothesis) in a nonlinear correlated stochastic system. The results of the analysis are shown in Figs. 7 and 8. For each embedding dimension and for each length N, $M = 199$ surrogates were used for a confidence of $\alpha = 0.01$.

In Fig. 7, we can see that for $D = 4$ [Fig. 7(a)], the R_{MP} of the series is consistent with the R_{MP} of the surrogates for almost all data lengths, suggesting that $D = 4$ is too small of an embedding dimension to use with this system. However, for larger values of D, the R_{MP} surrogate analysis is able to detect the nonlinear behavior of the system as the series length increases.

Figure 8 shows the results of the R_{MP} analysis when longer series are studied. The R_{MP} of the series becomes consistent with those of the surrogates. Therefore, the missing patterns from Fig. 7 are, in fact, not forbidden. The nonzero R_{MP}, which is significantly different from the surrogates in Fig. 7, is demonstrating nonlinearity, however, the convergence of the R_{MP} to that of the surrogates for longer series provides evidence that the system is stochastic.

In order to reliably distinguish between missing and forbidden patterns, very long time series were needed for the model systems. Unfortunately, in real world applications, series of such length are rare. Hence, the use of R_{MP} to detect determinism in data sets which do not satisfy $N \gg D!$ is difficult (especially when $D = 7$ or $D = 8$ is used) and may not be possible. While this is not a surprising result, the interesting finding is the amount of data needed for the nonlinear stochastic series, especially for $D = 7$, where 1 000 000 points were needed to demonstrate stochasticity even though $7! = 5040$.

Next, the insights learned from applying the R_{MP} analysis to the model systems will be applied to study experimental time series. Because of the long series needed to identify determinism or stochasticity in the model systems, the next section focuses only on detecting nonlinearity in experimental data.

V. APPLICATION TO EXPERIMENTAL DATA

In this section, the R_{MP} surrogate analysis is applied to three experimental data sets, measurements of the North Atlantic Oscillation (NAO), smoothed sunspot numbers, and EEG data.

A. North Atlantic oscillation index

The first example, the North Atlantic Oscillation (NAO) index which is measured as the difference in normalized pressures at the Azores High and Icelandic Low [26]. The NAO has a significant influence on the winter weather in Western and Central Europe. The monthly mean NAO index was downloaded from the Climate Prediction Center Web site [27]. The downloaded data spans from January 1950 until January 2017. However, in order to satisfy the end-to-end mismatch criterion, data from December 1953 until June 2016 was analyzed, for a total of 751 data points. Once the end-to-end mismatch criterion was satisfied, $M = 999$ surrogates were generated resulting in $\alpha = 0.002$. The series and its surrogates were symbolized using $\tau = 1$ and $D = 4$, 5, 6, and 7. The results of the analysis are shown in Fig. 9. Notice that in Fig. 9, the R_{MP} of the NAO time series is consistent with those of the surrogates for all embedding dimensions. Hence, the null hypothesis cannot be rejected for all embedding dimensions tested, supporting the conclusion that the data are consistent with a linear stochastic system and in agreement with the literature [26,28–32].

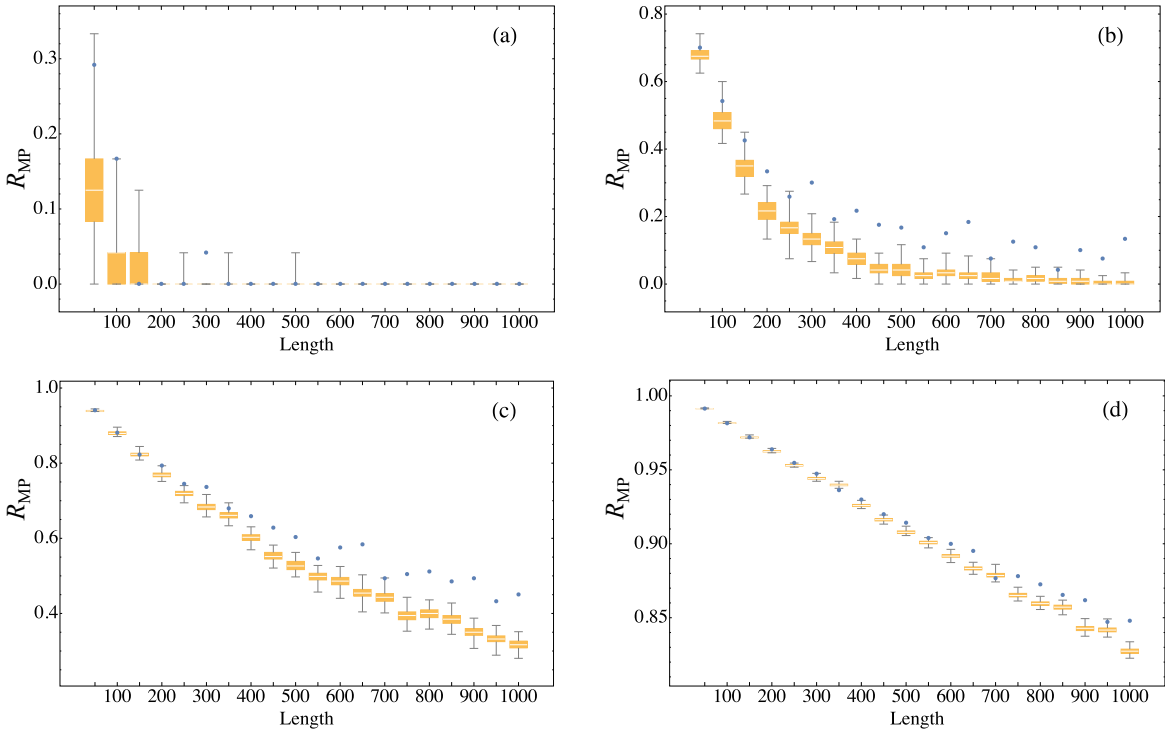

FIG. 7. Results of the surrogate analysis of (2). For each length N, $M = 199$ surrogates were generated for $\alpha = 0.01$. The box and whisker plot shows the range of values of the R_{MP} for the IAAFT surrogates and the scatter plot (blue dot) represents the R_{MP} of the original series. Results for embedding dimensions $D = 4$ (a), 5 (b), 6 (c), and 7 (d), and $\tau = 1$ have been included.

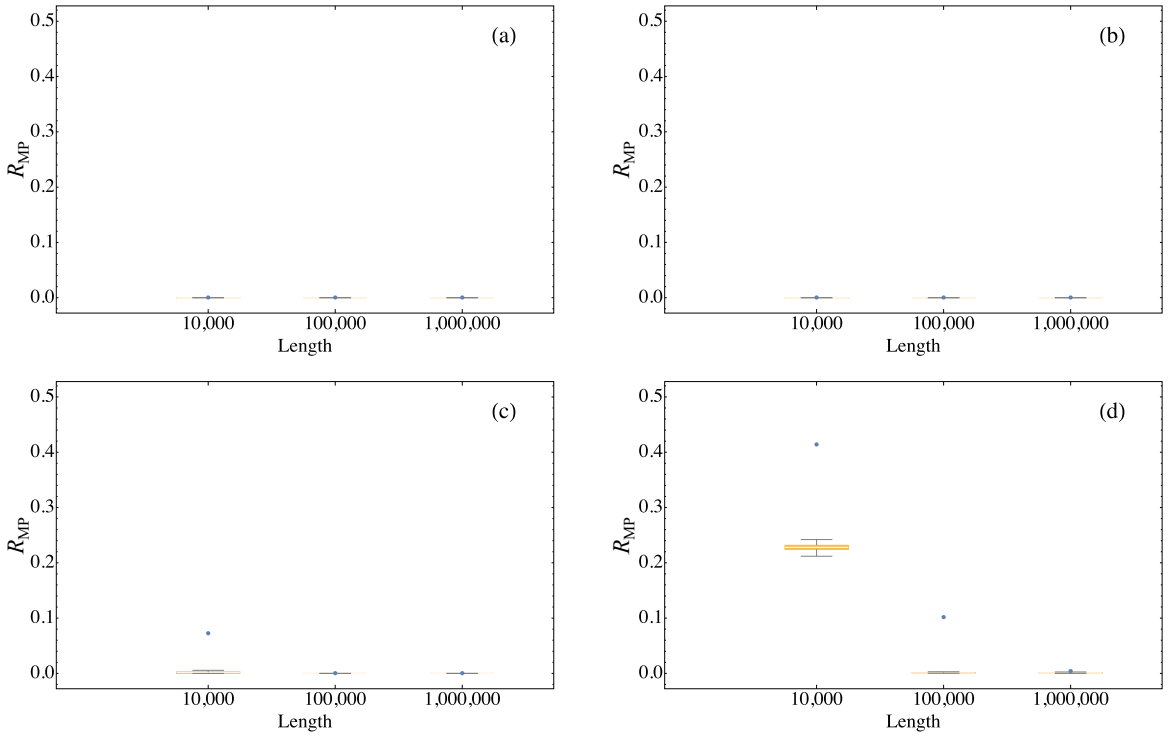

FIG. 8. Results of the surrogate analysis of (2). For each length N, $M = 199$ surrogates were generated for $\alpha = 0.01$. The box and whisker plot shows the range of values of the R_{MP} for the IAAFT surrogates and the scatter plot (blue dot) represents the R_{MP} of the original series. Results for embedding dimensions $D = 4$ (a), 5 (b), 6 (c), and 7 (d), and $\tau = 1$ have been included.

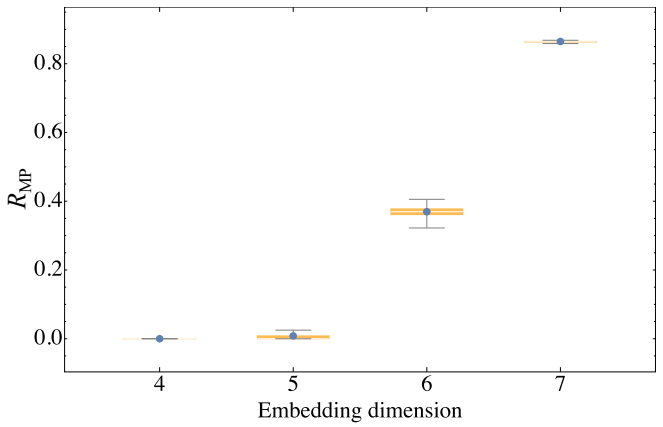

FIG. 9. Results of the surrogate analysis of NAO data using $M = 999$ surrogates resulting in $\alpha = 0.002$. The box and whisker plot shows the range of values of the R_{MP} for the IAAFT surrogates and the scatter plot (blue dot) represents the R_{MP} of the original series. Results for embedding dimensions $D = 4$, 5, 6, and 7, and $\tau = 1$ have been included.

B. Monthly smoothed sunspot numbers

The next data set we analyzed is the 13-month smoothed monthly sunspot numbers downloaded from the World Data Center SILSO, Royal Observatory of Belgium, Brussels [33]. The data used to satisfy the end-to-end mismatch criterion began in January 1749 and ended January 2017 for a total of 3205 data points. A total of $M = 999$ surrogates were created from the series and each series was symbolized using the same value of τ and D used for the NAO data.

The results from Fig. 10 suggest that nonlinearities are present in the data with a 99.8% confidence level for $D > 4$. This result is in agreement with [34] which detected nonlinearity with a 98% confidence level.

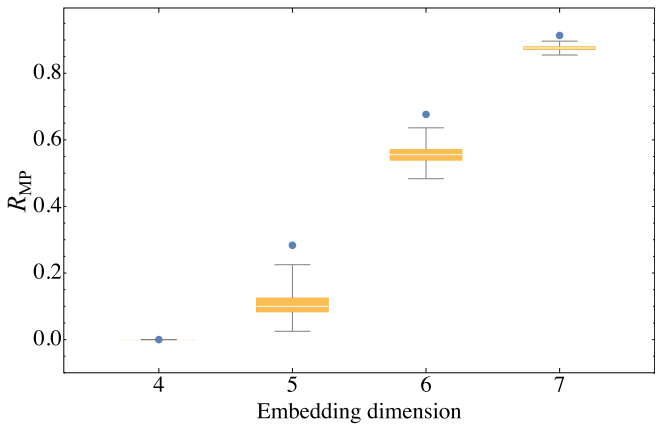

FIG. 10. Results of the surrogate analysis of 13-month smoothed monthly sunspot numbers using $M = 999$ surrogates resulting in $\alpha = 0.002$. The box and whisker plot shows the range of values of the R_{MP} for the IAAFT surrogates and the scatter plot (blue dot) represents the R_{MP} of the original series. Results for embedding dimensions $D = 4$, 5, 6, and 7, and $\tau = 1$ have been included.

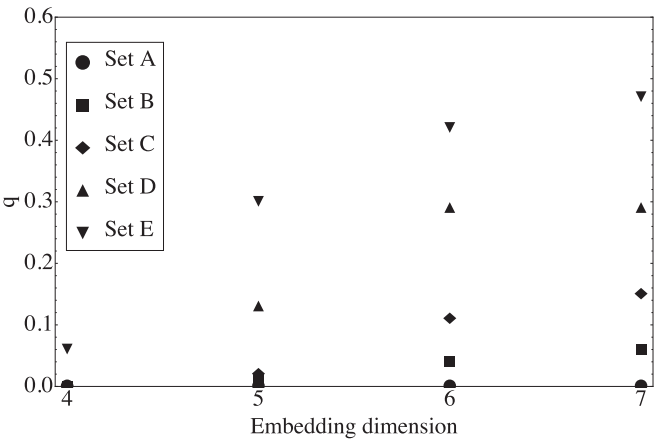

FIG. 11. Fraction of rejections of the null hypothesis q as a function of the embedding dimension for the EEG data. Results for embedding dimensions $D = 4$, 5, 6, and 7, and $\tau = 1$ have been included.

C. Electroencephalogram data

The final experimental series studied were electroencephalogram (EEG) data measured from healthy people and people with epilepsy. We used EEG from a public database [35]. The data consists of two sets of surface EEG time series measured from five healthy volunteers who were awake with eyes open (set A), and awake with eyes closed (set B). Furthermore, three additional data sets were measured from five epilepsy patients. Sets C and D were measured from the patients during a seizure-free interval from outside and inside the seizure generating areas, respectively. The final set, set E, contains intracranial EEGs measured during an epileptic seizure. Each data set contains 100 single-channel EEG segments sampled at 173.61 Hz for 23.6 s, resulting in $N = 4097$ data points. Additional information about the EEG data can be found in Ref. [36].

The EEG data was analyzed by producing $M = 999$ surrogates for each of the 100 recordings contained with in the five data sets. The fraction of null hypothesis rejections q occurring in each data set was found and the results are shown in Fig. 11.

Notice that in Fig. 11, data set A has no null hypothesis rejections while data set E has the most. Sets B–D have increasing numbers of null hypothesis rejections, suggesting an increase in the degree of nonlinearity for those sets. These results are similar to those obtained in Donges *et al.* [37].

VI. CONCLUSIONS

In this paper, the number of missing ordinal patterns (NMP) was demonstrated as an effective measure of nonlinearity in model and experimental time series. The efficacy of the NMP as a test for determinism was also studied in model time series where very long data sets can be obtained. The ability to obtain very long data sets for experimental series limits the applicability of the NMP as a test for determinism in experimental data, where the condition, $N \gg D!$, is often not satisfied.

One of the most important results in this paper is that the NMP can be an effective indicator of nonlinearity for very

short series when used within a surrogate framework. In many cases, the NMP is effective as a test for nonlinearity even if $N \ll D!$. Normally, when working with the Bandt and Pompe (BP) methodology, it is good practice to choose D such that $N \gg D!$. However, such a choice is not necessary when using the NMP as a test for nonlinearity.

Real world time series are often nonstationary. While the NMP test statistic will work for nonstationary series, the IAAFT surrogates will not. The surrogates produced by the IAAFT method are stationary. Hence, the NMP of a nonstationary series maybe different from those of its IAAFT surrogates, resulting in a potential false rejection of the null hypothesis. Improved surrogate methods have been proposed to overcome this drawback [38,39].

There are several outstanding questions that need further investigation. First, can the NMP be used, either within a surrogate framework or not, as an effective measure for short experimental time series? Are there methods with which one can study the NMP for an ensemble of randomly chosen subsets of the data in order to test how the NMP varies as a function of length, and if so, can that method provide a reliable means of detecting determinism? The authors' preliminary work on the aforementioned question suggest that such an analysis is difficult to interpret. Furthermore, testing the robustness to noise of the NMPs ability to detect nonlinearity is a problem that still needs to be addressed. The BP methodology's robustness to noise is well documented [12–14,40], and the NMP metric inherits that robustness. However, an open question is whether or not that robustness extends to the IAAFT surrogate framework. While it is believed that extension should carry over to the surrogate framework analysis; it is worthwhile testing that belief, because a full understanding of the limits of noise robustness is an important prerequisite to applying any time series test on real-world data. In addition to noise robustness, the NMP analysis' robustness to sampling irregularities and choice of τ is another avenue of future research. Included in such a study would be the sensitivity of the NMP analysis to sampling rates for continuous systems. Furthermore, it is often the case that real-world data are irregularly sampled. As was mentioned in Sec. I, there have been studies on the robustness of the number of forbidden patterns (NFP) to sampling irregularity. A similar study in a surrogate framework for irregularly sampled time series would also be of interest.

[1] G. A. Gottwald and I. Melbourne, Proc.: Math. Phys. Eng. Sci. **460**, 603 (2004).
[2] C. W. Kulp and S. Smith, Phys. Rev. E **83**, 026201 (2011).
[3] Z. Yang and G. Zhao, in *Proceedings of the 20th Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (IEEE, Piscataway, NJ, 1998), Vol. 20, p. 2670.
[4] C. W. Kulp and L. Zunino, Chaos **24**, 033116 (2014).
[5] M. Barahona and C.-S. Poon, Nature (London) **381**, 215 (1996).
[6] C.-S. Poon and M. Barahona, Proc. Natl. Acad. Sci. USA **98**, 7107 (2001).
[7] U. S. Freitas, C. Letellier, and L. A. Aguirre, Phys. Rev. E **79**, 035201(R) (2009).
[8] J. Gao, J. Hu, X. Mao, and W. wen Tung, Chaos Solitons Fractals **45**, 213 (2012).
[9] O. A. Rosso, H. A. Larrondo, M. T. Martín, A. Plastino, and M. A. Fuentes, Phys. Rev. Lett. **99**, 154102 (2007).
[10] J. S. A. E. Fouda and W. Koepf, Commun. Nonlinear. Sic. Numer. Simul. **27**, 216 (2015).
[11] L. Zunino, M. C. Soriano, and O. A. Rosso, Phys. Rev. E **86**, 046210 (2012).
[12] C. Bandt and B. Pompe, Phys. Rev. Lett. **88**, 174102 (2002).
[13] J. M. Amigó, S. Zambrano, and M. A. F. Sanjuán, Europhys. Lett. **79**, 50001 (2007).
[14] J. M. Amigó, S. Zambrano, and M. A. F. Sanjuán, Int. J. Bifurcation Chaos **20**, 2915 (2010).
[15] C. W. Kulp, J. M. Chobot, B. J. Niskala, and C. J. Needhammer, Chaos **26**, 023107 (2016).
[16] M. McCullough, K. Sakellariou, T. Stemler, and M. Small, Chaos **26**, 123103 (2016).
[17] K. Sakellariou, M. McCullough, T. Stemler, and M. Small, Chaos **26**, 123104 (2016).
[18] L. Carpi, P. Saco, and O. Rosso, Physica A **389**, 2020 (2010).
[19] T. Schreiber and A. Schmitz, Physica D **142**, 346 (2000).
[20] J. Theiler, S. Eubank, A. Longtin, B. Galdrikian, and J. D. Farmer, Physica D **58**, 77 (1992).
[21] T. Schreiber and A. Schmitz, Phys. Rev. Lett. **77**, 635 (1996).
[22] V. Venema, F. Ament, and C. Simmer, Nonlinear Processes. Geophys. **13**, 321 (2006).
[23] P. P. Kanjilal, J. Bhattacharya, and G. Saha, Phys. Rev. E **59**, 4013 (1999).
[24] J. Bhattacharya, IEEE Trans. on Syst., Man, and Cybern. - Part B: Cybern. **31**, 637 (2001).
[25] R. Hegger, H. Kantz, and T. Schreiber, Chaos **9**, 413 (1999).
[26] I. Fernández, C. N. Hernández, and J. M. Pacheco, Physica A **323**, 705 (2003).
[27] http://www.cpc.ncep.noaa.gov/products/precip/CWlink/pna/nao.shtm.
[28] D. B. Stephenson, V. Pavan, and R. Bojariu, Int. J. Climatol. **20**, 1 (2000).
[29] C. Collette and M. Ausloos, Int. J. Mod. Phys. C **15**, 1353 (2004).
[30] P. G. Lind, A. Mora, M. Haase, and J. A. C. Gallas, Int. J. Bifurcation Chaos **17**, 3461 (2007).
[31] M. D. Martínez, X. Lana, A. Burgeño, and C. Serra, Nonlinear Process. Geophys. **17**, 93 (2010).
[32] I. Fernández, J. M. Pacheco, and M. P. Quintana, Physica A **389**, 5801 (2010).
[33] http://www.sidc.be/silso/home.
[34] M. De Domenico and V. Latora, Europhys. Lett. **91**, 30005 (2010).
[35] http://epileptologie-bonn.de/cms/front_content.php?idcat=193.
[36] R. G. Andrzejak, K. Lehnertz, F. Mormann, C. Rieke, P. David, and C. E. Elger, Phys. Rev. E **64**, 061907 (2001).
[37] J. F. Donges, R. V. Donner, and J. Kurths, Europhys. Lett. **102**, 10004 (2013).
[38] J. H. Lucio, R. Valdés, and L. R. Rodriguez, Phys. Rev. E **85**, 056202 (2012).
[39] R. A. Rios, M. Small, and R. F. de Mello, Int. J. Bifurcation Chaos **25**, 1550013 (2015).
[40] J. M. Amigó, S. Zambrano, and M. A. F. Sanjuán, Europhys. Lett. **83**, 60005 (2008).