Application of computational mechanics to the analysis of natural data: An example in geomagnetism

Richard W. Clarke, Mervyn P. Freeman,* and Nicholas W. Watkins
British Antarctic Survey, Madingley Road, Cambridge CB3 OET, United Kingdom
(Received 9 October 2001; published 8 January 2003)

We discuss how the ideal formalism of computational mechanics can be adapted to apply to a noninfinite series of corrupted and correlated data, that is typical of most observed natural time series. Specifically, a simple filter that removes the corruption that creates rare unphysical causal states is demonstrated, and the concept of effective soficity is introduced. We believe that computational mechanics cannot be applied to a noisy and finite data series without invoking an argument based upon effective soficity. A related distinction between noise and unresolved structure is also defined: Noise can only be eliminated by increasing the length of the time series, whereas the resolution of previously unresolved structure only requires the finite memory of the analysis to be increased. The benefits of these concepts are demonstrated in a simulated times series by (a) the effective elimination of white noise corruption from a periodic signal using the expletive filter and (b) the appearance of an effectively sofic region in the statistical complexity of a biased Poisson switch time series that is insensitive to changes in the word length (memory) used in the analysis. The new algorithm is then applied to an analysis of a real geomagnetic time series measured at Halley, Antarctica. Two principal components in the structure are detected that are interpreted as the diurnal variation due to the rotation of the Earth-based station under an electrical current pattern that is fixed with respect to the Sun-Earth axis and the random occurrence of a signature likely to be that of the magnetic substorm. In conclusion, some useful terminology for the discussion of model construction in general is introduced.

DOI: 10.1103/PhysRevE.67.016203 PACS number(s): 89.75.−k

I. INTRODUCTION

Computational mechanics (CM) [1] has a formalism [2] that has been proved to construct the minimal model capable of statistically reproducing all the resolvable causal structure of any infinite sequence of discrete measurements (be they scalar, vector, tensor, or descriptive) [3]. The size of a model so defined, measured by a quantity termed statistical complexity C_μ [2] is a reliable and falsifiable indication of the amount of structure the data contain [4,5]. The model so defined, when using the formalism of Ref. [2], is known as an ''epsilon machine'' (see also Sec. II).

The particular strengths of this approach are that it enables the complexities and structures of different sets of data to be quantifiably compared and that it directly discovers detailed causal structure within those data. Most importantly, examining data in this way accurately identifies the scales at which structure exists within the series. This information can then be used to optimize the efficiency of physically plausible models [6].

As with all other analytical tools, CM has some limitations in the face of certain real-world problems that affect the information content of the signal under study. These problems may include the following:

(i) gaps in the data,
(ii) noise,
(iii) restricted sequence length,
(iv) correlations at a very wide range of scales.

The problem of correlations at a wide range of scales is particularly interesting and relevant to geophysical and other natural time series because of their typically power law (colored noise) Fourier spectra [7]. Theoretically, the minimum resolvable scale will be constrained by the data sampling interval and the maximum resolvable scale by the length of the data series. In practice, the range of resolvable scales will also be set by the available computational resources. Thus, the range of resolvable scales may be less than those necessary to evaluate correlations on all relevant scales. Consequently, it is important to understand how structural analysis is affected by unresolved structure due to correlation. A model is sought which is sofic: A system for which all left infinite sequences are followed by a finite number of semi-infinite right sequences (see p. 80 of Ref. [1]). A left infinite sequence is an arbitrarily long sequence of symbols and is followed by its arbitrarily long right-hand counterpart, referred in the literature [1] as the (right) semi-infinite sequence. These definitions are convenient for the analysis of mathematical constructions, but not for the practical analysis of real data.

In this paper, we address these issues in detail. In Sec. II, we discuss how the ideal formalism of CM can be adapted to apply to a noninfinite series of corrupted and correlated data. In particular, three concepts are defined and discussed: (i) A tolerance parameter [2] to account for the statistical uncertainty introduced by a noninfinite series that destroys the exact equivalence of different causal states sharing the same outcome. (ii) A new expletive filter that removes signal corruption by assuming that corruption creates rare causal states or words that are not in the dictionary of the true signal. (iii) The concept of effective soficity in which a data series has a finite set of equivalent causal states that is stable to small changes in the effective memory of those states.

*Electronic address: mpf@bas.ac.uk

The latter concept distinguishes between intrinsically unresolvable structure, ''noise,'' and as yet unresolved ''hidden'' structure whose discovery is only prevented by the effective memory being used and by the length of the data series.

In Sec. III, we apply the CM algorithm with these additional concepts to the analysis of structure in four simulated time series with wide applicability: (i) Uncorrelated, white noise. The epsilon machine for this structure is known [8]. (ii) Periodic signal with white noise corruption. This is a very widely used paradigm in applications as diverse as astronomy, biology, mechanical engineering, telecommunications, etc. (iii) A biased Poisson switch (i.e., a sequence of pulses whose pulse durations and interpulse intervals are determined by stationary Poisson processes). This is a (Markovian) case of the alternating renewal process (ARP) [9]. More general ARPs are models for the $1/f^D$ (red noise) spectra so prevalent in nature [9,10]. (iv) A sequence of bursts similar to (iii) but with fixed pulse duration. The structure of the time series is analyzed by searching for regions of effective soficity in maps of statistical complexity over the parameter space of the CM model.

The simulated time series also represent four types of signal thought to be present in time series measurements of the geomagnetic field. In Sec. IV, we use the CM algorithm to examine a real geomagnetic time series measured at Halley, Antarctica, in which deflections of the earth's magnetic field are due mainly to electrical currents in the ionosphere. The CM analysis yields a structural model that comprises a diurnal component corresponding to the oscillation of the measuring apparatus with the rotation of the Earth and a Poisson-switched, fixed-duration, pulse component that is likely associated with the magnetospheric substorm [11].

In Sec. V, we discuss some general principles that have been learned in applying CM to the analysis of structure in real data, and draw conclusions in Sec. VI.

II. METHOD

Here we give an introduction to the practical use of CM in the analysis of real data. We concentrate only on describing in detail the formalism for the parsing structure that we have used in the analyses. For a fuller description of the potential intricacies of the method see Ref. [3]. Defining some terminology, we highlight the difficulties associated with analyzing experimental data in this way, and explain solutions to these problems.

To start with, one has a set of measurements—either a spatial or temporal series where the separation between each point is known. The total time or length for which data exist is their span S. After coarse graining at a fixed scale s, the series has $N=S/s$ equally spaced measurements. Next, we digitize the signal amplitude. For reasons that will be apparent later, the number of possible digits should be low unless the series length is extremely large. The digitized sequence is then a concatenation of N letters $\mathbf{l}^N=\{l_0,l_1,\ldots,l_i,l_{N-1}\}$, where there are L types of such letters, ranging from 0,1,2, etc., up to $L-1$. In order to maximize the prior probable information content of the processed sequence, digitization should normally be performed such that there are equal numbers of each letter present. For example, in the case of binarization (where $L=2$), this would mean that the threshold for letter 1 would be the median value of the data. It should be noted, though, that the best way to digitize the sequence is that which *actually* maximizes the information content of the result; but that cannot usually be guessed. Another approach that has been suggested [6] is to use the formalism of maximum entropy.

The next step is to parse the sequence. One begins by composing words from each group of n consecutive letters; the ith word, $\mathbf{W}_i$, is defined by

$$\mathbf{W}_i=\mathbf{l}_i^n=\{l_i,l_{i+1},\ldots,l_{i+n-1}\}. \tag{1}$$

Thus there are L^n possible words, each represented by a unique scalar, W_i:

$$W_i=\sum_{j=0}^{n-1} L^{(n-1)-j}\times l_{i+j}. \tag{2}$$

The total number of words generated from the sample is $N-(n-1)$. We now introduce some terminology; any word $\mathbf{W}_i$ may be called a *proword*, $\mathbf{W}_P$ when followed by any word $\mathbf{W}_{i+1}$. This latter is called the *epiword* $\mathbf{W}_E$. For this sentence we digress slightly to note that it may sometimes be beneficial to perform the initial digitization on each separate block of data $2n$ letters long rather than the entire dataset.

We now proceed to capture causal structure in the word sequence by compiling a tally of epiwords following each proword. This means going through the sequence incrementing an array $\mathbf{T}_{(W_P,W_E)}$ accordingly. Representing summation over an index by its omission, we see that the total tally is $T=N-(2n-1)$. Thus, contracting over epiwords gives a tally of prowords only:

$$\mathbf{T}_{(W_P)}=\sum_{W_E=0}^{L^n-1}\mathbf{T}_{(W_P,W_E)} \tag{3}$$

and the fractional prevalence of each proword in the sequence is therefore contained in the vector

$$\mathbf{P}_{(W_P)}=\frac{\mathbf{T}_{(W_P)}}{T}. \tag{4}$$

Finally, the fractional *profile* of each proword by epiword is given by the array

$$\mathbf{P}_{(W_E|W_P)}=\frac{\mathbf{T}_{(W_P,W_E)}}{\mathbf{T}_{(W_P)}}, \tag{5}$$

where the repeated indices in the division are not summed over. Given a particular proword, this tells us the likelihoods of transitions to the various epiwords.

The crux of the technique now lies in identifying prowords with equivalent epiword profiles. Such prowords are said to belong to the same ''equivalence class'' or ''causal state''—i.e., they share statistically equivalent probabilistic

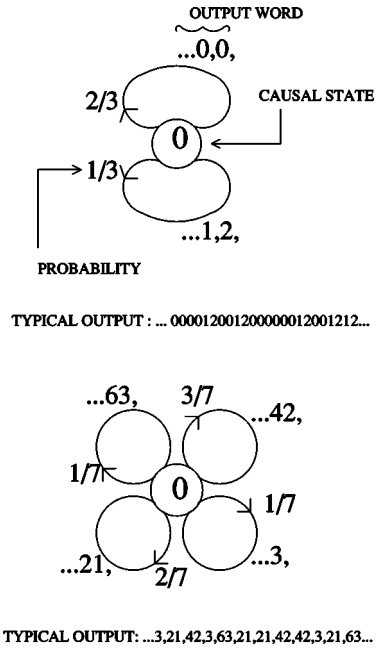

FIG. 1. Two labeled diagraphs representing minimal models with statistical complexities of zero.

futures (at the level of analysis one has been pursuing). The identification is made via an equivalence relation, denoted by $\sim$. For an infinite sequence, $\sim$ can demand exact correspondence between profiles, in which case it is always transitive (meaning $A\sim B, B\sim C \Rightarrow A\sim C$). In a practical situation, where even the finite length of the sequence introduces fluctuations in the calculated profiles [12], it is not possible to be so exact. We therefore introduce a tolerance parameter τ within the bounds of which the profiles of words in the same equivalence class are allowed to vary: Two prowords, A and B, are in the same equivalence class if, $\forall W_E$;

$$|\mathbf{P}_{(W_E|W_P=A)} - \mathbf{P}_{(W_E|W_P=B)}| \leqslant \tau, \tag{6}$$

where the large vertical bars signify absolute magnitude. Although this destroys the formal transitive property of $\sim$, because now $A\sim B, B\sim C$ no longer implies $A\sim C$, a practical way to reenforce it is to group equivalence classes that share at least one word.

Having identified the words lying within each equivalence class, a model which outputs a series of letters statistically equivalent to the original can be constructed. It is a particular strength of the technique that the model generated is always a minimal representation of the data's statistical structure for the amount of memory the analysis employs [5]. By ''statistically'' equivalent, we mean that the model reproduces the same profiles and statistical complexity (see below) as the original data, rather than just reproducing those statistical measures which discard phase information, e.g., autospectra/autocorrelation functions [13]. The model is easiest to describe in terms of its representation as a labeled ''diagraph.'' Two very simple labeled diagraphs, with extracts from their outputs, are presented in Fig. 1. A more complicated labeled diagraph, also representing a minimal model, is shown in

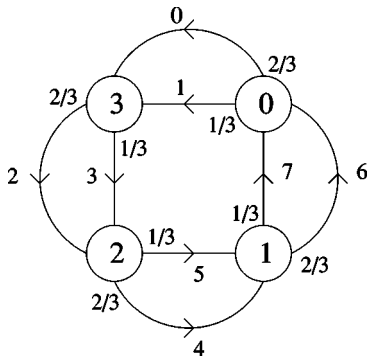

FIG. 2. A more complicated minimal model than that shown in Fig. 1.

Fig. 2. Each diagraph comprises a node or nodes, indicated by a circle with a number in it, and lines joining one node to another or to itself. Each numbered node of a diagraph represents a causal state corresponding to each of the model's equivalence classes, while each line (unidirectionally) joining two nodes is labeled with the string of letters (the word) that is output when that line is followed. In addition, each line is associated with a probability. The word output on going from one causal state to another is in the equivalence class of the future state. The probability of each word's output may therefore be trivially given by $\mathbf{P}_{(W_E|W_P)}$. It should be held in mind that only a subset of all possible labeled diagraphs represent minimal models. Even so, an arbitrary diagraph's output can naturally be used to construct the appropriate minimal model.

Models like this are useful for three reasons.

(i) Their minimality allows the structure of two sets of data to be directly compared.

(ii) Once a model has been synchronized with current data it optimizes one's ability to forecast the behavior of the system in the future.

(iii) The information concerning scales of causal structure in the data can be used to optimize the performance of more physically plausible models.

If a recursive decomposition is employed [4,5,14–18], diagraphs labeled with outputs that are words can be reformed into equivalent diagraphs labeled with single letters. If such a decomposition is performed on a topological basis rather than the statistical one described here, then the resultant diagraph is referred to as an ''epsilon machine'' [4]. This is perhaps a mathematically aesthetic thing to do. However, because any real analysis is performed with an effective memory of only n symbols, only the last n symbols are of any use for prediction, making it necessary to synchronize a single-letter diagraph to an input stream of data. This fact is not manifest in the single-letter diagraph obtained through the decomposition of the transition matrix $P_{(W_E|W_P)}$. In this paper we concentrate only upon the identification of proword equivalence classes, because it is a powerful tool for pattern discovery in its own right.

A measure of the structure of such models is given by the information entropy of the equivalence classes. The information that is retained about the conditional probabilities of states following a given state differentiates this approach

from more traditional information entropies, such as the Kolmogorov entropy [1]. When the equivalence classes are defined statistically, as in this paper, this measure is called statistical complexity C_ϕ and is given by [17]

$$C_\phi \equiv -\sum_i P(C_i)\log_2 P(C_i), \tag{7}$$

where logarithms are canonically taken to base 2 and the prevalence $P(C_i)$ of equivalence class i is given by the sum of the prevalences of the words in that class. When equivalence classes are defined topologically, the entropy measure is instead denoted by C_μ [2]; it is still called statistical complexity but it might be better termed topological complexity. For example, the models represented by the labeled diagraphs in Fig. 1 both have a statistical complexity $C_\phi = 0$ because they only have one causal state (and therefore one equivalence class) each. This is sensible because they both output noise. The model represented in Fig. 2, though, has four causal states with equal prevalences and a correspondingly higher statistical complexity of two bits:

$$C_\phi = -\sum_{i=1}^{4} P(C_i)\log_2 P(C_i) = -4\times\frac{1}{4}\times\frac{\ln\left(\frac{1}{4}\right)}{\ln(2)} = 2. \tag{8}$$

C_ϕ is extremely important, not only because it reflects the complexity of the system, but also because it does not converge until the data have been fully characterized. It is a hard fact that if the sequence length N is too small, full characterization will not be possible. This is because fluctuations in the proword profiles will corrupt the identification of equivalence classes. In this text the resultant unresolvable structure is called *noise*. In contrast, resolvable but as yet unresolved structure is described as *hidden* (see the discussion in Ref. [19]). Such hidden structure is likely to be encountered in analyzing data sets with correlation lengths comparable to or exceeding the maximum word length. Making this distinction is very important, even though it is not possible to discern whether unresolved structure is noisy or hidden until further computation has resolved it. In other words, the data appear to be noisy until the series is found to be *effectively sofic*, at which point C_ϕ attains its correct value and the model is complete. Sofic sequences are those which still have a finite number of equivalence classes when N is infinite and n is semi-infinite (see Badii and Politi [1], p. 80 for a longer explanation). *Effective* soficity is here defined to mean that a sequence has equivalence classes that are stable to an increase in word length. Thus, a sequence could be effectively sofic at one range of word lengths but not at another where either more or less structure is in the process of being identified.

Both hidden structure and noise will redistribute the original tally from what would be expected if only resolved structure was present, raising C_ϕ from the value corresponding to resolved structure alone and increasing the complexity of its model. A simple one-parameter model for the corruption process is to assume that the probability that any letter is corrupted to any other letter is χ. Then the probability any letter stays as it is $\sigma = 1-\chi$ and the corruption will have been governed by the redistribution function

$$\mathbf{T}^{\text{corrupt}}_{(W_P^C, W_E^C)} = \sum_{W_P=0}^{L^n-1}\sum_{W_E=0}^{L^n-1} P(\$\mathbf{W}_P = \mathbf{W}_P^C, \$\mathbf{W}_E = \mathbf{W}_E^C) \times \mathbf{T}^{\text{pure}}_{(W_P, W_E)}, \tag{9}$$

where $\$\mathbf{A} = \mathbf{B}^C$ reads, *the pure word* $\mathbf{A}$, *when corrupted by noise in a certain way, is identical to the corrupt word* $\mathbf{B}^C$. The label ''pure'' implies effective soficity. Thus, assuming the corruption of prowords and epiwords are independent we have

$$P(\$\mathbf{W}_P = \mathbf{W}_P^C, \$\mathbf{W}_E = \mathbf{W}_E^C) = P(\$\mathbf{W}_P = \mathbf{W}_P^C)P(\$\mathbf{W}_E = \mathbf{W}_E^C), \tag{10}$$

where

$$P(\$\mathbf{W} = \mathbf{W}^C) = \prod_{i=1}^{n} \{\sigma\delta(w_i = w_i^c) + \chi\delta(w_i \neq w_i^c)\}, \tag{11}$$

where δ is a Kronecker delta and the corruption of letters is assumed to be independent.

It happens that arbitrarily corrupted distributions can be uniquely deconvolved as long as one knows χ, but this is not usually the case in an experimental situation. We have two alternative options. The first is to scan through χ, deconvolving the proword prevalences each time. This will produce a drastic decrease in the statistical complexity at some point, signifying correct parametrization of χ. A good guess for χ might be the first value which results in a single proword having a prevalence of zero.

Whilst the assumption of independent corruption of letters is likely to be a good model of noise, it is unlikely to be a good model of the uncharacterized correlated structure that we call hidden. Consequently, a second option is to ignore the details of any corruption and simply assume that the prevalence of any expletive (corrupted word) is below a certain expletive prevalence x. We scan through x, eradicating any prowords whose prevalence is less than x, and recalculate C_ϕ each time. We choose as valid ranges for x those within which C_ϕ is constant, and therefore locally stable to variation of this parameter. This procedure can alternatively be performed after the identification of preliminary equivalence classes, to eradicate expletive equivalence classes. In any case, the approach can only work when the actual structure-to-noise ratio (SNR) is high enough to ensure that expletives are eradicated before meaningful words are. If the pure proword prevalence distribution is very uneven this method cannot work. In general, a combined method would probably be most successful—that is, where one first attempts the deconvolution and then removes the resulting low-prevalence words completely. It is always possible to determine all the resolvable structure of a sequence for which the SNR is arbitrarily small, so C_ϕ is independent of

SNR. Of course though, if the SNR is zero, so is C_ϕ, because the model suddenly collapses to a single equivalence class.

Note that deconvolution can always be achieved by inversion of an assumed convolution matrix, but that this is not always easy. In particular, if one knew the actual matrix then the ''noise'' would not be noise at all, but resolved structure. The only deconvolution that is strictly necessary is that which removes the noise (unresolvable structure) from the signal. It should therefore assume that the redistribution is Gaussian. In practice though, some hidden (resolvable) structure may be so computationally difficult to identify that a messy deconvolution is required to remove it, allowing the analysis of more easily resolvable structure to proceed. It is admissible to remove expletives from the prevalence distribution because they destroy the effective soficity of the data.

It is instructive at this point to go through the uncertainties present in the profile and prevalence distributions. When the sequence length is large compared to L^n the probability that any individual word has been corrupted is approximately $\Delta = n\chi$. Following the definition of the prevalence distribution, we find that the uncertainty in the prevalence of a proword $\Delta P_{(W_P)}$ is governed by an inequality:

$$\frac{\Delta}{\sqrt{T}} < \Delta P_{(W_P)} < \Delta, \tag{12}$$

where the lower limit corresponds to uncorrelated errors and the upper limit to systematic errors. We indeed expect the uncertainty to be somewhere in this range because the errors are due to unresolved structure. The uncertainty in the prevalence of a single epiword within a particular proword's profile is expected to be greater:

$$\frac{\Delta}{\sqrt{\mathbf{T}_{(W_P)}}} < \Delta P_{(W_E|W_P)} < \Delta. \tag{13}$$

These inequalities go some way to justifying the use of the blanket tolerance τ to identify the equivalence classes, because we know nothing about the nature of the errors. In some cases it is conceivable that τ would have to be scaled by $1/\mathbf{T}_{(W_P)}{}^m$, where $0 < m < \frac{1}{2}$ in order to correctly identify equivalence relations between profiles. In such cases m is an extra parameter.

III. EXAMPLES

We now turn to the analysis of test sets of data by the algorithm described in detail above. The test data represent signal types thought to be present in time series measurements of the geomagnetic field that we shall study in the following section.

A. White noise

A white noise (temporally uncorrelated) signal was generated by a sequence of 5000 independent samples from a uniform distribution and converted to binary by setting those values above the median to unity and those below the median to zero. Figure 3 shows the variation of statistical complexity C_ϕ versus word length n and tolerance τ for this signal. The absence of a plateau in this graph indicates that, for the range of memories (word lengths) tested, the analysis does not discern any structure at all in the signal. The linear variation of C_ϕ with n for $\tau \approx 0$ represents models with as much arbitrariness as possible at each level of memory used in the analysis. These models collapse to a single equivalence class as the tolerance parameter is increased. An increasing amount of tolerance is required for this collapse for increasing word length, as expected from Eq. (13). Thus, no complex models at all were constructed for this noisy sequence at any time during this analysis. This was expected; we would have been disappointed with the random number generator that was used to construct the sequence (the IDL ''randomu'' function, see, also, Ref. [20]) if we had easily found correlations.

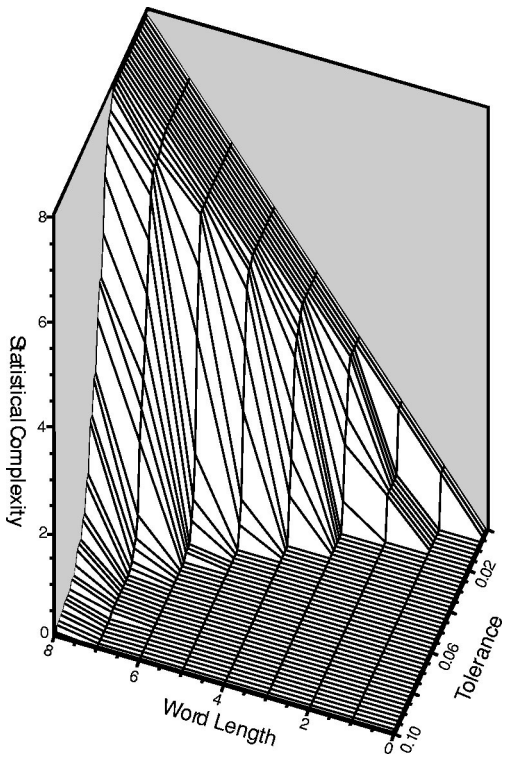

FIG. 3. Statistical complexity of a noisy binary sequence, 5000 symbols long over a range of model construction parameters.

B. Periodic signal with white noise corruption

Figure 4 shows the result of the analysis on a binary period four sequence (i.e., 00110011$\cdots$) of length 5000, where 10% of the bits have been randomly flipped. This graph has a stable, but rather jagged, plateau at $C_\phi \approx 2.8$ which begins at word length 4 for tolerances in the range $0.1 \leqslant \tau \leqslant 0.24$. This plateau corresponds to a group of models that capture the essential structure in the signal. In the absence of noise the statistical complexity of a binary period four signal should be $C_\phi = 2$. The apparently anomalously high level of the plateau is caused by both the noise and the

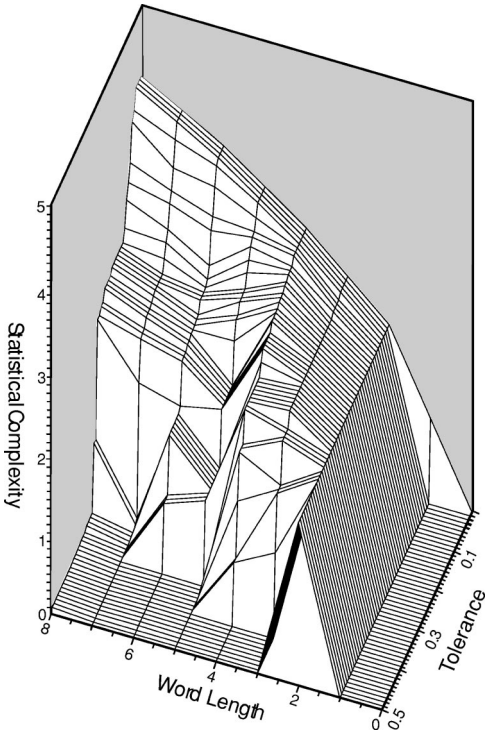

FIG. 4. Statistical complexity of a binary period 4 sequence, 5000 symbols long, 10% flipped at random, over a range of model construction parameters.

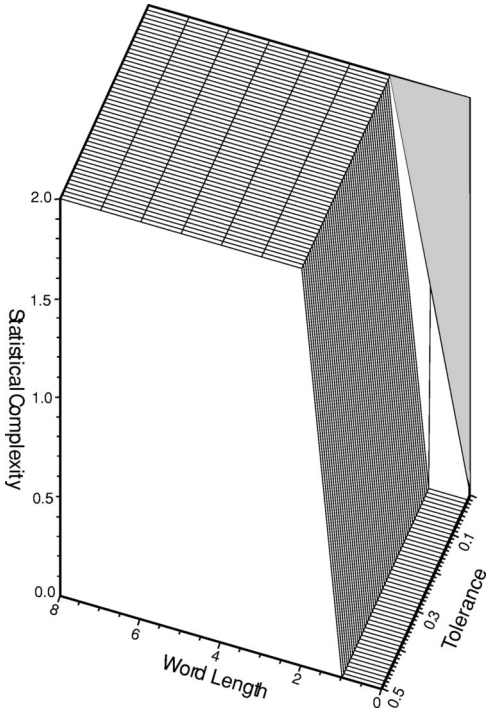

FIG. 5. Analysis of the same sequence as in Fig. 3, with an assumed expletive prevalence of $x=0.075$.

finite sequence length corrupting the identification of the equivalence classes. It is not entirely flat because the corruption is different at each value of word length and tolerance. In fact, there is a gentle downward trend which would converge to $C_\phi=2$ in the limit of the extra, spurious, states decreasing in prevalence at longer and longer word lengths, if the sequence was long enough. Note that the gradient of the increase of statistical complexity with word length changes at a memory equal to half the period of the structure in this signal. This is the point at which the structure is first discovered: It is important to note that the convergence of C_ϕ is not immediate, suggesting an analog of the Nyquist sampling theorem for CM. Note also steep drops in C_ϕ where previously distinguishable equivalence classes have suddenly collapsed together as the tolerance parameter τ exceeds some critical value.

Figure 5 shows results of a similar analysis on the same sequence, excepting that this time, words of prevalence less than x (the expletive prevalence parameter) were eradicated from the probability distributions. x was chosen to be 0.075 for this graph. As we can see, this approach was entirely successful in the respect that C_ϕ converges to a plateau for a broad range of the tolerance parameter τ. A minimal model that was capable of outputting sequences with statistical structure identical to that characterized from the input was effectively constructed at every point on this plateau. The value of C_ϕ for periodic sequences was always found to directly reflect the amount of memory required by the system to produce such data: A sequence with a sole period Q has a statistical complexity of $\log_2(Q)$ bits (in this case the period four signal has a statistical complexity of 2.0 bits). Moreover, we can appreciate that the analysis only yields a convergent value after the word length has exceeded at least half the period of the sequence. More generally; convergence begins when an analysis first has greater memory than a system. If the system has certain structure with greater memory than it may be feasible to analyze, for example, a red noise signal that consists of many Fourier modes with a power law distribution of amplitudes and random phases, C_ϕ will not ever truly converge. However, there may be stages where the analysis has enough memory to identify *some* structure, and this is indicated by approximately flat regions, or, at the very least, dips in the gradient of C_ϕ with increasing word length.

C. Biased Poisson switch

We next turned to more detailed analyses of two other illustratively important diagraph's outputs. The first we con-

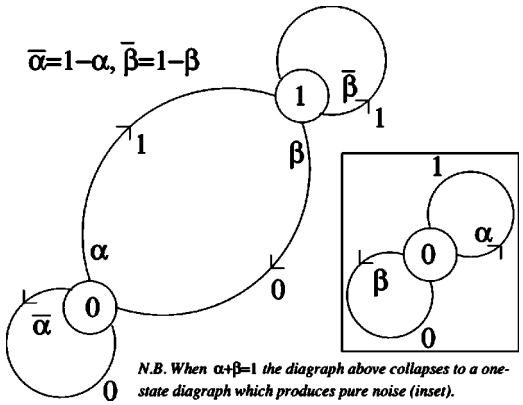

FIG. 6. The minimal model of biased Poisson switches.

sidered was the biased Poisson switch [21], represented as a labeled diagraph in Fig. 6. The circled states, 0 and 1 may each generate either a one or a zero with the probabilities shown.

In the figure, $\overline{\alpha}=1-\alpha$ and $\overline{\beta}=1-\beta$. Note that when $\alpha=\beta$ the output sequence is no longer biased. It turns out that the values of C_ϕ we can derive for different values of α and β provide some nice insights into the nature of information and the optimization of measurement processes. The measure has two distinct regimes: where $\alpha+\beta=1$, and where they do not. Since the diagraph only has two states, it is clear that as far as prediction of the next epiword is concerned, only the last bit of any proword can ever matter. Therefore, all words usually separate into two equivalence classes (corresponding to odd and even words). If, however, $\alpha+\beta=1$ then $\alpha=1-\beta=\overline{\beta}$ and $\beta=\overline{\alpha}$. This always results in the two equivalence classes collapsing into one, giving a statistical complexity of zero, corresponding to pure noise. This is appropriate because in this degenerate situation the possible outcomes of node 0 in Fig. 6 are identical to those of node 1 and the diagraph collapses to a single state too (see inset), and can only produce noise anyway. If the diagraph does not collapse in this way there will always be two equivalence classes. Their prevalences are found to be $1/(\alpha/\beta+1)$ and $1/(\beta/\alpha+1)$. Thus, when $\alpha+\beta\neq 1$, we have

$$C_\phi=-\sum_i P_i\log_2(P_i)=\frac{1}{\left[\frac{\alpha}{\beta}+1\right]}\log_2\left(\frac{\alpha}{\beta}+1\right)+\frac{1}{\left[\frac{\beta}{\alpha}+1\right]}\log_2\left(\frac{\beta}{\alpha}+1\right) \tag{14}$$

and if $\alpha+\beta=1$, C_ϕ is always zero. A graph of this function is shown in Fig. 7. Note that it always evaluates to unity when $\alpha=\beta$, except when $\alpha=\beta=\frac{1}{2}$. If α does not equal β (and $\alpha+\beta\neq 1$) then it is less than unity. In fact, as the switch becomes more and more biased the statistical complexity goes down and down, reaching zero when only one digit is ever output. This is to be expected because a biased data set (e.g., more ones than zeros) is a symptom of an inefficient measurement apparatus: If one symbol is more prevalent than any other then the system is undercharacterized by the alphabet in use. In the parlance of Shannon's theory of communication this statistical complexity is equivalent to the maximum rate of information.

Given that the collapse discussed above takes a slice out of the graph in Fig. 7, we would expect sequences generated by certain Poisson switches to be more difficult to characterize. For example, sequences produced by a switch with $\alpha=\beta=0.49$ have a statistical complexity of unity, but it is difficult to distinguish them from noise (where $C_\phi=0$) because they are so close to the collapse at $\alpha=\beta=\frac{1}{2}$. Such a sequence, one million symbols long, was analyzed up to a word length of $n=7$ at 100 equal intervals between $\tau=0.00$ and $\tau=0.01$, without assuming any noise was present

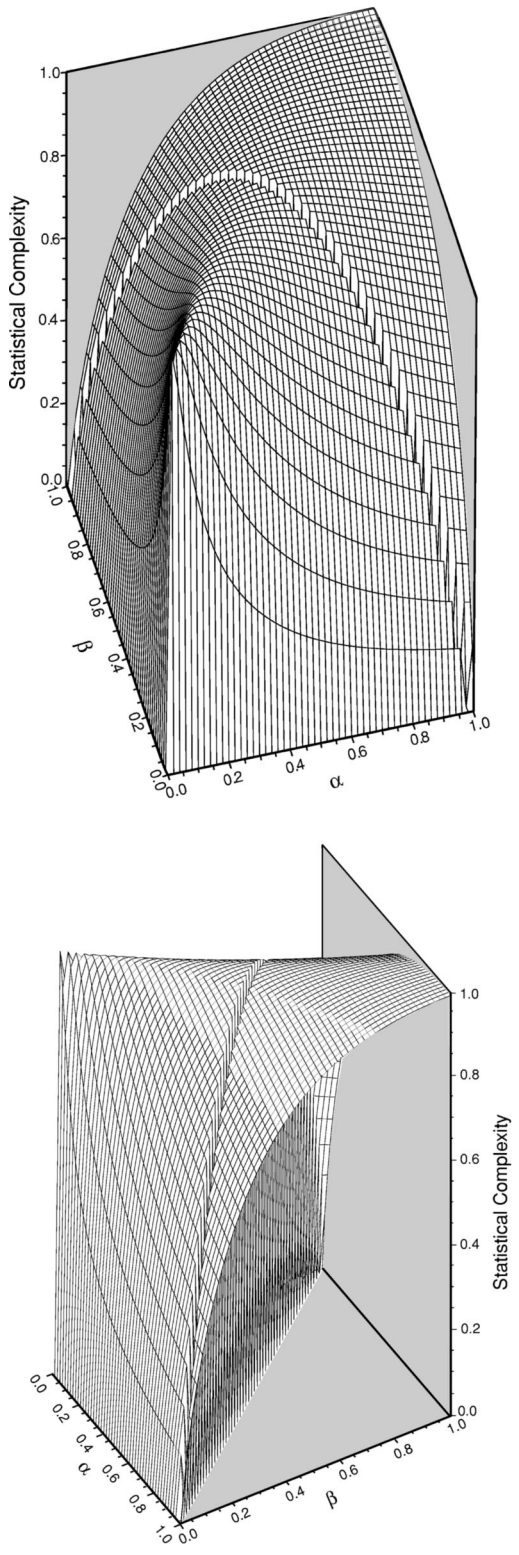

FIG. 7. Two views of the variation of statistical complexity of the biased Poisson switch versus up-switching bias α and down-switching bias β.

(i.e., $\chi=0$ and $x=0$). The results are presented in Fig. 8. The plateau corresponding to the optimal model is that which has a statistical complexity of unity. We can see that it is difficult to construct this model because the plateau is radi-

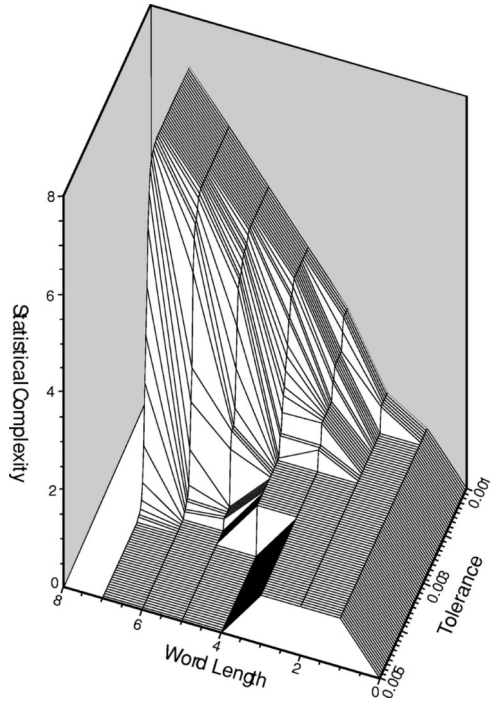

FIG. 8. Statistical complexity of a one million binary symbol sequence produced by a Poisson switch with $\alpha=\beta=0.49$, over a range of model construction parameters.

cally constricted at higher word lengths. On one side of it τ is too small to identify the equivalence classes, so every proword occupies its own equivalence class and $C_\phi=n$, its maximum value at any word length n. On the other side, τ is too large, so the two equivalence classes collapse together, producing degenerate models that would output noise. The $C_\phi=1$ plateau has a distinct end at $n=5$ because the sequence is not long enough to support analysis at a word length of $n=6$. At the latter word length statistical fluctuations in every proword profile mean that the correct classification of equivalence classes is no longer possible at any range of τ. We are not too concerned about this here because we have already identified the optimal model which was stable from $n=1$ to $n=5$. In fact, a ''more optimal'' model would be able to predict the flipping of the switch itself to some extent. The construction of such a model would probably need a lot of computation and would probably require N to be very large. These things depend on how random the switch is and on the signal-to-noise ratio.

If a set of data is very complicated, no stable model might be identified before the word length becomes too large to be statistically supportable by the sequence length. The only solution is to gather more data. The alternative is to settle with models that are either inadequate or arbitrarily complicated. Although the latter models reproduce structure well (and are therefore most useful to engineers), studying them can reveal little about underlying processes. They are scientifically unaesthetic. In contrast, one can tell a lot about the intricacies of a system from the minimal adequate model associated with it at a certain level of analysis. This is the concern of scientists.

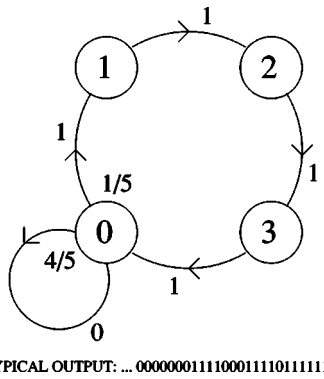

FIG. 9. The minimal model of an unbiased switch which sustains for 4 symbols.

D. Fixed pulse duration Poisson switch

The next class of labeled diagraphs we consider produce binary sequences that are simple models of a process with bursts. These sequences have the structure of sustained switches—that is, when the switch is down it has a constant probability of switching up, and when up, it stays up for a fixed count U. When the sequence is unbiased the up-switching probability is $1/(U+1)$. See Fig. 9 for an example of this kind of labeled diagraph. The exact statistical complexities of such unbiased sustained switches are given by

$$C_\phi=-\left[\left(\frac{U+1}{2U}\right)\log_2\left(\frac{U+1}{2U}\right)+(U-1)\left(\frac{1}{2U}\right)\log_2\left(\frac{1}{2U}\right)\right]$$
$$=\frac{1}{\ln(2)}\left[\ln(2U)-\left(\frac{U+1}{2U}\right)\ln(U+1)\right]. \quad (15)$$

We now investigate the practical analysis of a sequence one million binary symbols long that was produced by a sustained switch with $U=4$. The statistical complexities of the models constructed by the analysis are shown in Fig. 10. It can be seen that the first convergent values are at word lengths one greater than U. That is to say, good models can be constructed when the analysis first has a greater memory than the system. The plateau identifiable with a model of the form shown in Fig. 9 begins at a word length of 4 and extends laterally from $\tau\approx0.02$ to $\tau\approx0.06$. The remarkable thing about this plateau is that, although it is very flat, it is not *entirely* flat. It begins at $C_\phi^5\approx1.770\,69$ which is significantly higher than the theoretical statistical complexity of: $C_\phi=1/\ln(2)[\ln(8)(\frac{5}{8})\ln(5)]\approx1.548\,79$ and subsequently oscillates around this value while it converges to it (e.g., $C_\phi^{10}\approx1.521\,60$). This behavior is caused by the phase ambiguity due to the absence of information concerning the synchronization of a burst when a word is composed entirely of ''up'' symbols. For example, at word length six, the profile of word 63, (i.e., 111111 in binary), is a superposition of the profiles of sequences like 10[111111], 01[111111], and 11[111111], from each of which it cannot be distinguished at that level of analysis. Therefore, in this case, the profile of word 63 does not match that of any other word, and is allocated its own equivalence class. Although the prevalence of

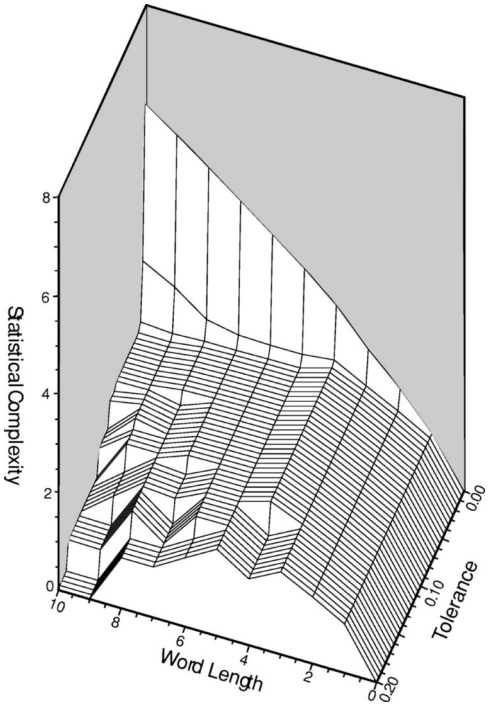

FIG. 10. Statistical complexity of a one million binary symbol sequence produced by an unbiased switch which sustains for four symbols, over a range of model construction parameters.

this word, and thence its class, is very low, it is sufficient to distort the statistical complexity.

In an analysis with recourse to infinite memory, the prevalence of an infinite sequence of ''up'' symbols is zero. Thus, the U causal states of such a sequence would be correctly identified, and the statistical complexity of the model constructed would match exactly with the theoretical value. Of course, in practice no analysis can have infinite memory. If one wishes to retain optimal predictability of future data then it is necessary to accept whatever model is actually constructed by an analysis with finite memory.

IV. ANALYZING GEOMAGNETIC DATA

The test data examples analyzed in the preceding section represent signal types thought to be present in time series measurements of the geomagnetic field. If this is true, we may expect to see similar structure emerging from a CM analysis of a real geomagnetic time series.

The CM analysis detailed in Sec. II was performed on 3-h averaged measurements of the variation of the East-West component of the geomagnetic field D at Halley, Antarctica, from three separate years: 24 February–16 December, 1995, 26 January–28 December, 1998, and 2 January–30 December 2000. A graph of the data from 26 January to 28 December 1998 is shown in Fig. 11. It can be seen that the magnetic deflections have both a linear trend and a high frequency signal with an annual amplitude modulation that maximizes in the austral summer. The linear trend is caused by the movement of the ice shelf upon which Halley is situated and was removed by subtracting the result of a linear regression for each of the three years. The detrended time series was

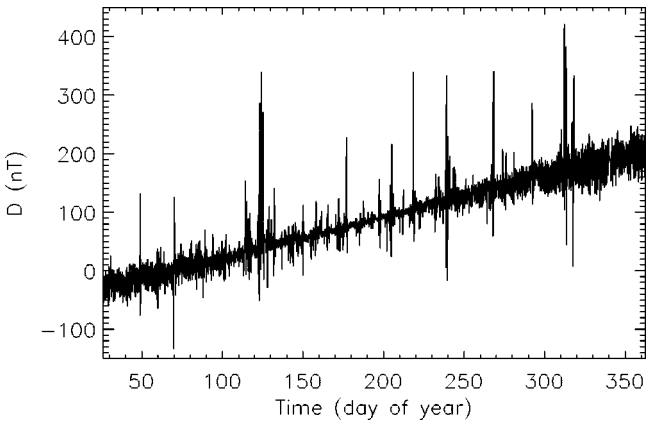

FIG. 11. Eastward D component of the magnetic deflection at Halley during 1998.

then binarized with respect to the median, giving three sequences of 2352, 2688, and 2896 symbols, respectively. These series were then analyzed up to a word length of 10 and with tolerances varying in 80 equal steps from 0.05 to 0.25. Words with prevalences less than $x=0.004$ were eradicated. The graph of statistical complexity C_ϕ is shown in Fig. 12. Two plateaus are evident, one at $C_\phi \approx 0.9$, covering a wide range of tolerances and between word lengths of 1 and 3, and the other plateau at $C_\phi \approx 5.0$, at the highest corner

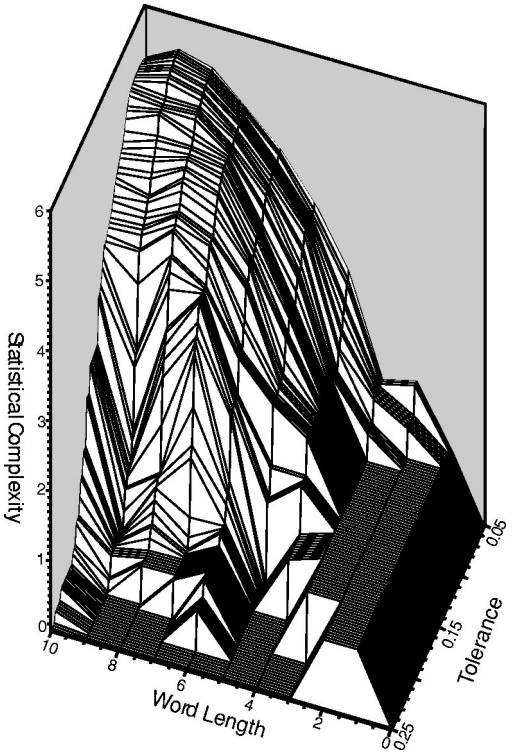

FIG. 12. Statistical complexity of a binary symbol sequence from about three years' worth of 180-min time-averaged readings of the positive eastward component of the magnetic deflection at Halley, over a range of model construction parameters. The plateau at a word length of eight indicates major correlation at a period of 24 h. All the plateaus in this diagram were stable to variation of the assumed expletive prevalence, here set at 0.4%.

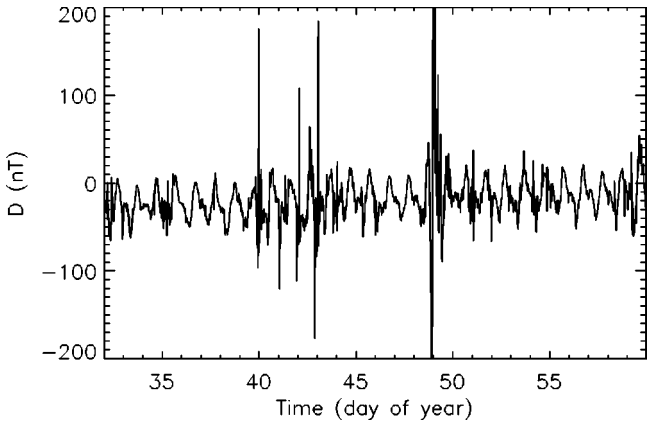

FIG. 13. Eastward *D* component of the magnetic deflection at Halley during February 1998.

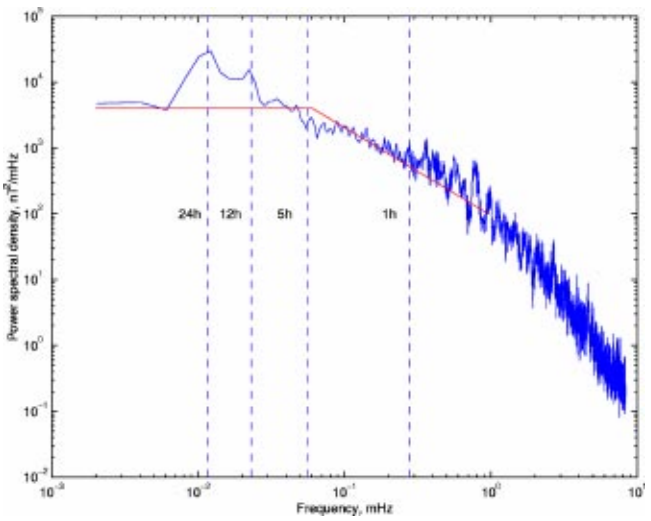

FIG. 14. Fourier spectrum of the Eastward *D* component of the magnetic deflection at Halley for 1998. Two spectral slopes of f^0 below 0.06 mHz and $f^{-4/3}$ above 0.06 mHz are shown for reference. Vertical dashed guidelines are also shown at key frequencies referred to in the manuscript.

of the graph, between tolerances of about 0.05 and 0.07 and at word lengths of 8 or more. The convergence to this second plateau is slow, so it is not as obvious as the lower plateau. However, this does not mean that it is any less significant: The only criterion that these features need to satisfy is that the statistical complexity does not vary in their vicinity in parameter space. This is to ensure that each parameter has a definite and well-defined value for the particular regime of structure that has been identified.

The convergence of statistical complexity at a word length of 8 corresponds to a time scale of $8\times3=24$ hours. Such a diurnal variation is well known and is primarily caused by the rotation of the observing station with the Earth under the so-called SQ ionospheric current system that is driven by pressure gradients caused by solar heating and is thus fixed in the Sun-Earth frame [22]. The variation can be seen in the raw data, as illustrated by plotting a typical month of Halley geomagnetic data in Fig. 13. The associated ground magnetic variation has neither a pure sinusoidal shape nor a fixed period of exactly 24 h, and this is likely to contribute towards the higher observed statistical complexity of $C_\phi\approx5.0$ compared to the $C_\phi=3$ that would be expected for a pure binary period 8 signal.

Of course, it is possible to detect a simple 24 h periodicity using Fourier techniques: Fig. 14 shows the Fourier power spectrum of the eastward *D* component of the magnetic deflection over Halley for the year 1998. The power spectrum shown is the average of 59 power spectra calculated from 8192 min-long intervals of 1-min averaged data. Each interval was linearly detrended and a Hanning window applied before calculating the power spectrum. The average power spectrum shows a clear peak at 0.012 mHz (24 h), and a second harmonic, but there are no other obvious peaks below 1 h^{-1} (0.28 mHz). In contrast, the CM analysis detected structure at 3–9 h time periods, as evidenced by the plateaus at word lengths of 1–3 in Fig. 12. The reason why such structure is not seen as a peak in the Fourier power spectrum must be because the signal is not periodic, only recurrent. Instead, such recurrent structure could explain why there is a break in the gradient of the Fourier power spectrum at a frequency of ≈0.06 mHz (a period of about 5 h). For example, it is possible to generate a Lorentzian-like power spectrum from a random sequence of pulses [23], but the measured spectral slope above 0.06 mHz in Fig. 14 is not exactly f^{-2}. Alternatively, the break could be a low-pass filter effect. Such ambiguity illustrates how Fourier methods are excellent tools for the analysis of signals containing many distinct periodicities, but are harder to interpret for the more general class of stochastic signals.

Instead, Computational Mechanics provides a more appropriate formalism for the analysis of these signals. The plateau in Fig. 12 at word lengths of 1–3 indicates the presence of significant structure at 3–9 h time scales. This plateau has a statistical complexity of ≈0.9 and an overall structure similar to that of Fig. 8, suggesting the possibility of some random pulselike process. Such a possibility is intriguing because pulselike geomagnetic perturbations on hour time scales (known as magnetic bays) are particularly prominent during the night time at high (auroral zone) latitudes and are associated with magnetospheric substorms [24]

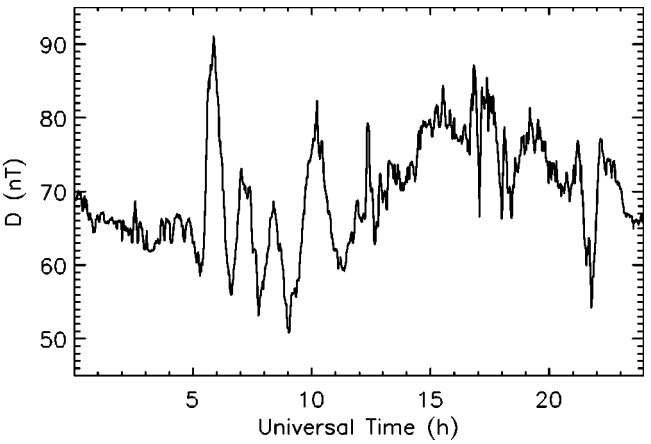

FIG. 15. Eastward *D* component of the magnetic deflection at Halley during 16 June 1998.

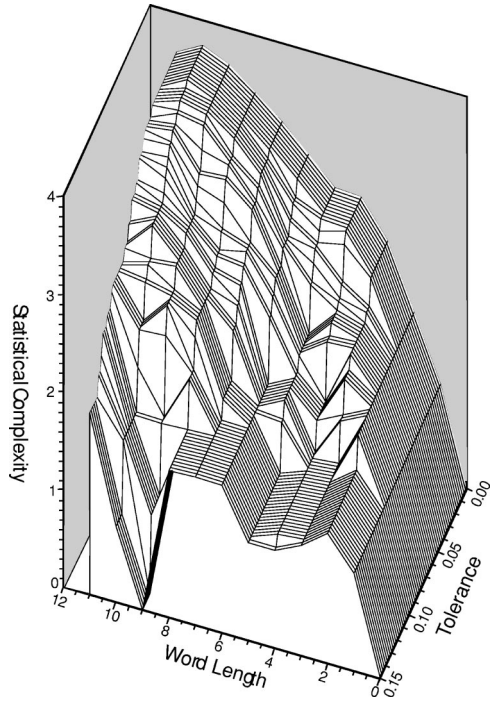

FIG. 16. Statistical complexity of a 12011 binary symbol sequence from 40-min time-averaged readings of the positive eastward component of the magnetic deflection at Halley over a range of model construction parameters.

whose occurrence has been argued to be a stationary Poisson process with mean recurrence time of 5 h [11]. Figure 15 shows a single day of Halley geomagnetic data that illustrates the presence of such pulselike disturbances on hour time scales sitting on top of the diurnal variation.

To investigate this further, an analysis was made of a 40-min averaged time series of the East-West component of the geomagnetic field at Halley from 00:00 UT, 25 January, 1998 to 00:00 UT, 26 December, 1998. After removing the linear trend in the data due to the movement of the ice shelf, the time series was binarized with respect to the median, giving a sequence of 12011 symbols. The series was analyzed up to a word length of 11 for tolerances in 60 equal steps between 0.00 and 0.15. Words with prevalences less than $x=0.015$ were eradicated. The graph obtained for C_ϕ is shown in Fig. 16. The plateaus in this graph are stable to variation of x. The higher plateaus have models that are more useful for prediction of future data, if they are stable to an increase in the amount of data available to the analysis. The lower plateaus have models that show the most dominant structures—and are easier to understand and interpret physically.

It can be seen from the graph that, at a tolerance between $\tau=0.12$ and $\tau=0.15$, more structure is identified between word lengths six and eight than it was possible to resolve with a memory of only five symbols. The model which corresponds to this plateau is represented, for a word length of seven, in Fig. 17. For clarity, transitions with a probability less than 0.055 are not shown, which is why each node's branching probabilities do not quite sum to unity. Moreover, the output labels are binarized averages, weighted according

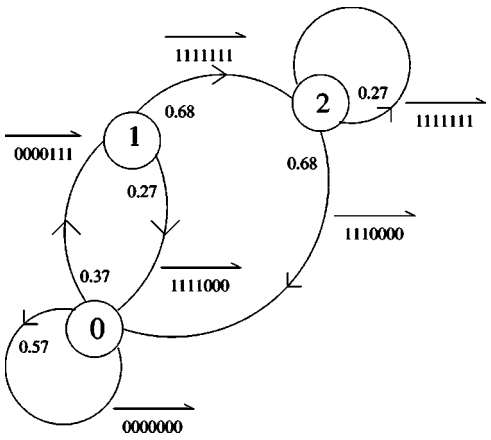

FIG. 17. Predominant structure of the diagraph constructed from the Halley data at word length seven ($\tau=0.14$, $x=0.015$). For clarity, transitions with a probability less than 0.055 are not shown, which is why each node's branching probabilities do not quite sum to unity.

to the probabilities of individual words. Thus, this diagraph represents the least detailed structure discovered in the time series. More complicated diagraphs were constructed that are more suited to prediction than easy interpretation. It is a coincidence of averaging over all words between nodes that a 0.68 and 0.27 split appears twice as the actual transitions differ in structure. The details of the model are in Table I. Comparing with Fig. 6, the transitions between states 0 and 1 of this diagraph are an approximately Poisson-switched process with a timescale of about 5 h. This value is given by the range of word lengths capable of resolving this structure from the sequence within this range of τ ($n=6,7$, and 8); at $n=7$ the characteristic timescale is 7×40 minutes≈4 h and 40 min.

It was thought that the other states and transitions in Fig. 17 would be caused by the diurnal variation of the data alone. This was investigated by analyzing, in exactly the same way, a pure binary sequence with a period of 36 symbols—corresponding to a period of one day if each symbol were to represent a 40-min average. The principal transitions of the model constructed for this sequence are shown in the diagraph drawn in Fig. 18. The structural similarities and differences between this diagraph and the one in Fig. 17 are obvious, and support the idea that the transitions between states 0 and 1 of Fig. 17 are due to substorm activity, rather than merely being an artifact of a partially characterized 24-h period.

V. DISCUSSION

In the preceding sections, we have demonstrated how CM can measure the statistical complexity of linear data sequences and construct the minimal model necessary to describe the data. The reader may have noticed that there are seven degrees of freedom in making such a model:

(a) digitization method (binary, trinary, etc.),

(b) coarse-graining scale s,

(c) sequence length L,

TABLE I. Details of the geomagnetic data model: Details of the simple stable model at word length seven, $\tau \approx 0.14$. $x = 0.015$, at which value 114 of 128 words are cut. Statistical complexity $= 1.51477$ bits.

Class	Equivalent surviving words
0	0 64 96 112 120 124 126
1	1 3 7 15 31 63
2	127

Word number (Class)	Word	Probability
	Transitions from Class 0	
0 (0)	0000000	0.481382
1 (1)	0000001	0.0759726
3 (1)	0000011	0.0729505
7 (1)	0000111	0.0641128
15 (1)	0001111	0.0540634
31 (1)	0011111	0.0546018
63 (1)	0111111	0.0532191
64 (0)	1000000	0.0338069
96 (0)	1100000	0.0234980
127 (2)	1111111	0.0514929
	Transitions from Class 1	
0 (0)	0000000	0.0423272
64 (0)	1000000	0.0291209
96 (0)	1100000	0.0209377
112 (0)	1110000	0.0209523
120 (0)	1111000	0.0293505
124 (0)	1111100	0.0442261
126 (0)	1111110	0.0822315
127 (2)	1111111	0.684994
	Transitions from Class 2	
0 (0)	0000000	0.0789801
64 (0)	1000000	0.0907960
96 (0)	1100000	0.0945274
112 (0)	1110000	0.103234
120 (0)	1111000	0.108831
124 (0)	1111100	0.105721
126 (0)	1111110	0.101990
127 (2)	1111111	0.268035

Class	Average word	Probability
	Average transitions from Class 0	
0	[0.161, 0.102, 0.061, 0.040, 0.027, 0.009, 0.000]	0.573587
1	[0.000, 0.142, 0.288, 0.432, 0.603, 0.797, 1.000]	0.374920
2	[1.000, 1.000, 1.000, 1.000, 1.000, 1.000, 1.000]	0.051493
	Average transitions from Class 1	
0	[0.843, 0.735, 0.657, 0.579, 0.470, 0.306, 0.000]	0.269146
1	[0.000, 0.309, 0.444, 0.590, 0.714, 0.906, 1.000]	0.045860
2	[1.000, 1.000, 1.000, 1.000, 1.000, 1.000, 1.000]	0.684994
	Average transitions from Class 2	
0	[0.885, 0.752, 0.614, 0.463, 0.304, 0.149, 0.000]	0.684080
1	[0.000, 0.117, 0.221, 0.325, 0.532, 0.701, 1.000]	0.047886
2	[1.000, 1.000, 1.000, 1.000, 1.000, 1.000, 1.000]	0.268035

Class	Prevalence
0	0.483737
1	0.269561
2	0.246701

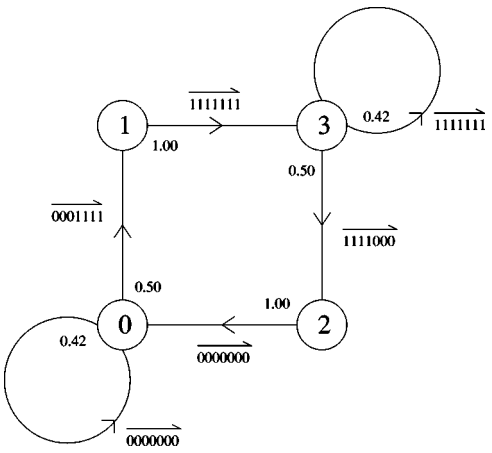

FIG. 18. The main transitions of the minimal model for a binary period 36 sequence at a word length of 7, $\tau=0.14$, $x=0.015$.

(d) word length n,
(e) tolerance τ,
(f) corruption frequency χ, and
(g) expletive frequency x.

These degrees of freedom express the level of information in the data and the depth of knowledge with which the model is probing the system from which the data are measured. For example, increasing the sequence length L, reducing the coarse-graining scale s, or increasing the digitization from binary to trinary, all provide increased information and thereby increased knowledge of the system that the data represent. Conversely, increasing the tolerance or the expletive frequency reduces information by admitting different states to be equivalent or to be omitted, respectively, thereby reducing knowledge of the system. Consequently, we might anticipate that the best model of the system is the model corresponding to the region of the multidimensional parameter space in which information is maximized. Whilst such a model is the most accurate description of the data sequence with the greatest information content, it is not necessarily the optimal model of the system. This is because any data sequence is not a complete representation of the system it is measured from. In particular, it is limited in two important respects: First, there is structure in a data sequence, that we have termed noise, that cannot be resolved under any amount of computation. This will create differences in the profiles of words that are statistically insignificant and should be ignored by allowing some nonzero value of tolerance, corruption frequency or expletive frequency. Second, there is structure in a data sequence, that we have termed hidden, that has not been resolved at a certain level of memory or word length but that is resolvable at a greater word length. In other words, meaningful models of the data can only be found within certain, usually finite, zones of the parameter space [25]. Within each zone, C_ϕ is constant and the model is both stable and minimal. Outside this zone, the model is either too degenerate or overly complicated. For example, it will be degenerate (and C_ϕ will be too low) if τ is set too large. This is because equivalence classes will collapse into one another. Similarly, the model will be unnecessarily complicated (and C_ϕ will be too high) if τ is set too small. This is because distinctions will be made between words on the basis of insignificant differences in their profiles.

An analogy is the construction of a vocabulary for the structure of speciation of feline animals. If one is too fussy about the tail, Manx cats cannot be classed as domestic cats. If one's sole criterion is purring or a meow, a lion cub may be misclassed as a domestic cat. The correct classification of feline animals needs a finite amount of information to fall within the boundaries of a finite number of provisos.

In the case of computational mechanics we interpret effectively sofic models to be optimal. Thus we seek plateaus in the multidimensional parameter space. Generally, this space can contain many plateaus, the heights of which are the corresponding models' statistical complexities. If we want to forecast the data most accurately, we are looking for the highest plateau, which has the most stringent conditions [26]. More physically understandable models may exist on some lower plateaus where only the more dominant causal structures are preserved.

Thus, in the end, the success of the analysis depends upon the existence of effectively sofic plateaus of statistical complexity in the multidimensional parameter space and our ability to discover them. This is contingent upon the data that are supplied and how much computing power is available. It is important to bear in mind that the data are not only a function of the physical system's behavior, but also of the measurement apparatus and any preprocessing. There are four main pitfalls (represented by corresponding model parameters).

(i) Mischaracterization of the system by the measurement apparatus $(\ldots\chi,x)$.

(ii) Degradation of data prior to the analysis by processing (s).

(iii) Insufficient data to resolve all structure present (L), and

(iv) insufficient computing power to resolve hidden structure (n,τ).

The apparatus may easily mischaracterize the system, either by introducing structure to the data which is foreign to the system's behavior, or by neglecting to transcribe structure that should be present. This situation is most apparent when the apparatus is clearly only taking measurements from a cross section of the system. Nevertheless, if it is reasonable to assume in a particular case that the apparatus is capable of providing a good representation, then identified structure can be attributed to the system. In such cases we would also expect the statistical complexity to scale with the system's true complexity. Naturally, this is not valid when the cross-section happens to be an exact subsystem.

Although all processing degrades data, it may still be possible to correctly characterize all the structure present. This is because the degradation will usually produce noise, which can be ignored. A graver problem is when (uncharacterizable) noise represents some of the system's structure. The only solution may be to collect more data, but other preliminary approaches are to use a finer scale when coarse graining and/or to digitize more finely. However, it is always necessary to choose sensible margins for the parameter search because some regions of the parameter space are computation-

ally very costly to explore. For example, a trinary sequence is about seven hundred times as hard to fully analyze at a word length of eight than a binary sequence. You must have a good reason not to use binary.

An alternative approach may be useful when the data have resolvable structure at two widely separated scales; it may be more computationally efficient to construct higher-level equivalence classes than to persist with using longer and longer words. Classes on the next highest level are found by applying the same analysis method to the sequence expressed in terms of a set of primary level causal states for which C_ϕ has not yet converged. All information between the scales $n_1 s$ and $n_2 s$ is lost in this process. Even so, it is a more preferable approach than simply further coarse graining the data to intervals of $n_1 s$ if one has reason to believe that the system's degrees of freedom at the two scales are coupled. The total statistical complexity is the sum of those calculated at each level, so it is in fact possible to test for such coupling by comparing the coarse-grained C_ϕ with the hierarchical value.

There is actually no reason why the prowords and epiwords should not come from different sequences, enabling the direct causal correlation of two systems, such as the solar wind and the magnetosphere.

VI. CONCLUSION

Computational mechanics is an intuitive and powerful way to study complicated nonlinear sequences derived from physical systems. This is because the analysis identifies causal structure from data presented to it and constructs the minimal adequate model that fits these data. The information about this structure, and in particular its scales, can then be used to optimize more physically plausible models. In this paper, we have discussed in detail how the original formalism has to be used when applied to noninfinite sequences. The main conclusion is that models constructed by computational mechanics are good if, and only if, they are stable to the variation of the parameters used to construct them from the data. In addition, two quite general definitions are made. These concern the general constructibility of models from a set of observations:

(i) Structure which cannot be resolved from a set of data under any amount of computation is most usefully called *noise*.

(ii) Structure which has not been resolved at a certain level of computation or memory, but which is resolvable from the set of data is usefully called *hidden*.

The prior undecidability of whether unresolved structure is noise or hidden is a direct parallelism of Gödel's famous theorem. For a proof relating the two fields, but in a slightly different context, see G.J. Chaitin [27].

The method developed in this paper was applied to magnetometer measurements of ionospheric currents for the years 1995, 1998, and 2000. The technique successfully constructed models, the simplest of which comprised a diurnal component and a Poisson-switched process with a timescale of about 5 h that likely relates to the occurrence of magnetic substorms. The most complicated model could be used to forecast space weather.

A similar method was also proposed to characterize the causal relationship of any two systems, such as the solar wind and the magnetosphere.

ACKNOWLEDGMENTS

We are grateful to Tom March and Sandra Chapman for a very thorough reading of the manuscript resulting in several valuable suggestions for improvement, to Cosma Shalizi for helpful discussions.

[1] R. Badii and A. Politi, *Complexity-Hierarchical Structures and Scaling in Physics* (Cambridge University Press, Cambridge, UK, 1999).

[2] J.P. Crutchfield and K.A. Young, Phys. Rev. Lett. **63**, 105 (1989).

[3] C.R. Shalizi and J.P. Crutchfield, J. Stat. Phys. **104**, 819 (2001).

[4] J.P. Crutchfield, Physica D **75**, 11 (1994).

[5] C.R. Shalizi and J.P. Crutchfield, e-print arXiv.org/abs/cs.LG/0001027.

[6] A.J. Palmer, C.W. Fairall, and W.A. Brewer, IEEE Trans. Geosci. Remote Sens. **GE-38**, 2056 (2000).

[7] J.D. Pelletier and D.L. Turcotte, Adv. Geophys. **40**, 91 (1999).

[8] D. P. Feldman, Ph. D. thesis, University of California, Davis, 1998.

[9] S.B. Lowen and M.C. Teich, Phys. Rev. E **47**, 992 (1993).

[10] M.B. Weissmann, Rev. Mod. Phys. **60**, 537 (1988).

[11] J.E. Borovsky, R.J. Nemzek, and R.D. Belian, J. Geophys. Res. **98**, 3807 (1993).

[12] Even if the profiles are roughly equally occupied then a sequence length of about L^{2n} is required to prevent statistical fluctuations from dominating the identification of equivalence classes.

[13] A. Witt, J. Kurths, F. Krause, and K. Fischer, Geophys. Astrophys. Fluid Dyn. **77**, 79 (1994).

[14] J.P. Crutchfield and K.A. Young, in *Complexity, Entropy and the Physics of Information*, edited by W.H. Zurek (Addison-Wesley, Reading, MA, 1990), p. 223.

[15] J.P. Crutchfield, in *Modeling Complex Phenomena*, edited by L. Lam and V. Naroditsky (Springer-Verlag, Berlin, 1992), p. 66.

[16] J.E. Hanson, Ph. D. thesis, University of California, Berkeley, 1993.

[17] N. Perry and P.-M. Binder, Phys. Rev. E **60**, 459 (1999).

[18] K.A. Young, Ph. D. thesis, University of California, Santa Cruz, 1991.

[19] H. Poirier, Science et Vie **1003**, 56 (2001).

[20] W.H. Press, S.A. Teukolsky, W.T. Vetterling, and B.P. Flannery, *Numerical Recipes*, 2nd ed. (Cambridge University Press, Cambridge, 1996).

[21] For the unbiased Poisson switch, see the discussions of the ''random telegraph'' in J. S. Bendat, *Principles and Applications of Random Noise Theory* (Wiley, New York, 1958); and of the Poisson switch in H.J. Jensen, *Self-Organized Criticality* (Cambridge University Press, Cambridge, 1998).

[22] S.-I. Akasofu and S. Chapman, *Solar-Terrestrial Physics* (Oxford University Press, New York, 1972).

[23] See H.J. Jensen, K. Christensen, and H.C. Fogedby, Phys. Rev. B **40**, 7425 (1989), and also Ref. [21].

[24] A.J. Smith, M.P. Freeman, M.G. Wickett, and B.D. Cox, J. Geophys. Res. **104**, 12 351 (1999).

[25] If we had been considering infinite sofic sequences this zone would be always be infinitely big because all structure would be resolved after a certain word length, and would continue to be resolvable. This does not usually apply to finite sofic sequences because beyond a certain threshold word length statistical fluctuations in the profiles are able to destroy the equivalence classes.

[26] In an engineering context, one would usually settle for that model constructed by the highest level of analysis, whether or not it is on a plateau. This is not acceptable from a scientific perspective because there is no way to justify that such models are not totally arbitrary.

[27] G.J. Chaitin, IEEE Trans. Inf. Theory **IT-20**, 10 (1974).