



An independent test of methods of detecting system states and bifurcations in time-series data

V.N. Livina^{a,*}, P.D. Ditlevsen^b, T.M. Lenton^c

^a School of Environmental Sciences, University of East Anglia, Norwich, UK

^b Centre for Ice and Climate, Niels Bohr Institute, Copenhagen, Denmark

^c College of Life and Environmental Sciences, University of Exeter, UK

ARTICLE INFO

Article history:

Received 12 April 2011

Received in revised form 1 July 2011

Available online 25 August 2011

Keywords:

Potential analysis

Degenerate fingerprinting

Time series analysis

ABSTRACT

We present an independent test of recently developed methods of potential analysis and degenerate fingerprinting which aim, respectively, to identify the number of states in a system, and to forecast bifurcations. Several samples of modelled data of unknown origin were provided by one author, and the methods were used by the two other authors to investigate these properties. The main idea of the test was to investigate whether the techniques are capable to identify the character of the data of unknown origin, which includes potentiality, possible transitions and bifurcations. Based on the results of the analysis, models were proposed that simulated data equivalent to the test samples. The results obtained were compared with the initial simulations for critical evaluation of the performance of the methods. In most cases, the methods successfully detected the number of states in a system, and the occurrence of transitions between states. The derived models were able to reproduce the test data accurately. However, noise-induced abrupt transitions between existing states cannot be forecast due to the lack of any change in the underlying potential.

© 2011 Elsevier B.V. All rights reserved.

1. Introduction

Developing new statistical techniques requires their testing on artificial data, which is usually performed on simulated surrogate series with known properties [1]. In this paper, we attempt an independent (or 'blind') test of new methods of time series analysis. Blind testing is well established in the geosciences (see for instance [2–4]) and is a rigorous form of testing new methods. The methods of time series analysis we examine are increasingly being used in important applied contexts, where their reliability needs to be well established. In particular, several groups are interested in detecting past transitions or bifurcations in climate data [5–9], with the ultimate aim of trying to forecast such features in the future [10,11]. It is vital that the methods used are tested to assess their reliability and limitations, because any positive results (e.g. a forecast of forthcoming climate bifurcation) could be extremely important to societies, but are also likely to be contested (in scientific and social arenas). Whilst those developing and applying statistical methods should always do a careful job of self-checking their reliability in a given context, past experience (in climate science at least) suggests this has not always been the case.

A transparent way to test methods is to apply them to artificial time series provided by an independent party who knows (but does not reveal) the underlying model used to generate the data. Here several sets of time series data were provided by one of the authors (P. D. Ditlevsen) without revealing their origin. The methods to be tested were then run by the other authors, and from the results of their statistical analysis and visual observations, they described the expected properties

* Corresponding author.

E-mail address: vlivina@gmail.com (V.N. Livina).

and behaviour of the underlying systems and attempted to model the corresponding data. The provider of the original data then commented on the results and disclosed the origin of the series and the extent to which the methods used were able to detect and describe the dynamics of the simulated systems. In Section 2, we describe the methods tested. Section 3 presents the results. Section 4 describes the performance of the methods. Section 5 concludes.

2. Methodology

For detecting and forecasting critical behaviour in time series (by which we mean climate temperature transitions and bifurcations), two techniques have been developed; potential analysis and degenerate fingerprinting. We distinguish transitional and bifurcational behaviour: in the case of bifurcation, the number of states in the system is changing (structural change), whereas in the case of transition, the underlying system potential may remain the same but the system state may be shifting with some trend (linear, parabolic, etc.). The transitions may be noise-induced, when the system oscillates between known states of a fixed potential, or forced by external forcing, as in the case of increasing temperature due to fossil fuel emissions.

2.1. Potential analysis

Following several early studies that proposed multistability of the climate system [12–14], it was suggested [15,5] that a stochastic equation with double-well potential can be appropriate for modelling some aspects of the Earth's climate. In particular, a climatic potential function was introduced in Ref. [16], where its time-dependent properties (probability flux and variance) and most probable states of the stationary probability distribution were studied. A stochastic resonance in a system with periodically oscillating potential was studied in Ref. [17], where a dynamical system was subject to both periodic forcing and random perturbation. However, these studies considered only bistable potential, whereas the approach we use allows for a higher number of states. Moreover, the present method provides a novel presentation of the dynamics of a time series in a potential contour plot, where bifurcations are visualised [8,18].

Following the potential approach, Kwasniok and Lohmann [7] studied a proxy for paleotemperature in Greenland during the last ice age using the stochastic differential equation

$$\dot{z}(t) = -U'(z) + \sigma\eta, \quad (1)$$

with a double-well potential

$$U(z) = a_4 z^4 + a_3 z^3 + a_2 z^2 + a_1 z,$$

and proposed a numerical method to derive the coefficients of the potential from an observed time series using the unscented Kalman filter (UKF) for its parameter estimations (see [19,7,20]).

Based on this approach, we have developed the method of potential analysis [8,18], which provides a dynamical portrait of time series (see Figs. 3–5). The number of system states is estimated by means of a polynomial fit of the probability density function of the data. The results of the analysis of sliding windows of variable size are plotted in a contour potential plot [8]. As an adopted colourmap standard, we denote detected one-well potential by red colour, double-well potential by green colour, triple-well potential by cyan colour, four-well-potential by purple colour; the number of detected wells bigger than 4 is denoted by yellow colour. If the change of the number of states (change of colour in the potential contour plot) is observed along all time scales (varying time windows), this denotes a bifurcation in the series. If there is only a sporadic change at smaller time scales, this may be due to small-size effects or nonstationarities, and does not indicate a global bifurcation.

2.2. Degenerate fingerprinting

The method of 'degenerate fingerprinting' was proposed by Held and Kleinen [10], who used the lag-1 auto-correlation function to estimate the decay rate in a time series from its short-term memory. For a system approaching a bifurcation, a nearly universal property is that this decay rate will tend to infinity – a phenomenon known as 'critical slowing down'. Subsequently, we modified the method of degenerate fingerprinting [11] to use Detrended Fluctuation Analysis (DFA) to estimate the decay rate. DFA is a widely used tool for the study of statistical scaling properties of nonstationary time series which was introduced in Ref. [21]. It has been applied successfully to DNA sequences [22], heart-rate dynamics [23–25], and climate dynamics [26,27].

Here we employ and compare both methods for estimation of the decay rate in time series. In the case of the autocorrelation approach, any linear trend is first removed from the data.

In the degenerate fingerprinting techniques, a sliding window of fixed size is moving along a time series and lag-1 correlation or DFA exponent is estimated for the segment of data. This provides the so-called ACF- and DFA-propagators, whose tending to a value 1 denotes critical behaviour. In principle, if a system is approaching a bifurcation, this should be forewarned by an upward trend in the propagators towards 1. Comparing propagators allows us to analyse the influence of trends and variability in the short-term memory.

To illustrate the propagator technique, in Fig. 1 we show artificial data, in which a sigmoid function is superimposed with red noise with fluctuation exponent $\alpha = 0.7$, which was generated using a perturbation in the Fourier domain [28]. This dataset simulates a forced transition, where the pattern of fluctuations does not change but the underlying trend drives a

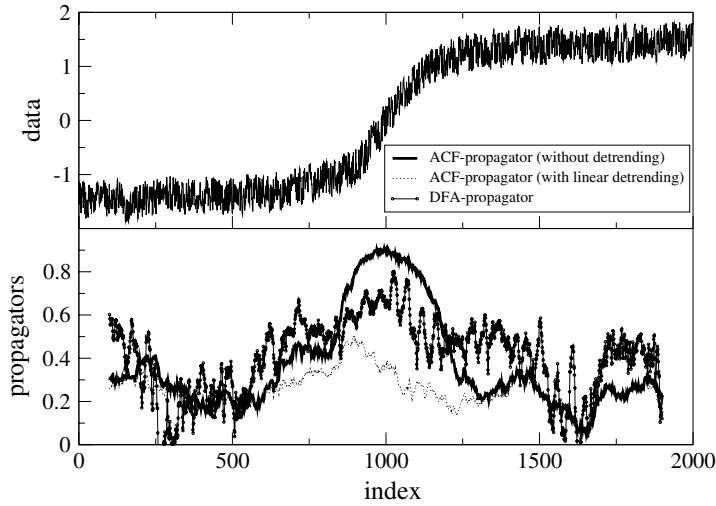


Fig. 1. Sigmoid function superimposed with colour noise with fluctuation exponent $\alpha = 0.7$ (upper panel). ACF-propagators with and without detrending: the influence of the transitional trend is indicated in the area where these two curves of ACF-propagators differ. DFA-propagator, similarly to the ACF-propagator with detrending, shows no presence of a genuine bifurcation.

system to another state. When applying the ACF-propagator without detrending, we see clear indication of critical behaviour, with values approaching 1. However, if we apply ACF-propagator with detrending or DFA-propagator, we can see that there is no real criticality in the data, and thus we can conclude that there is no genuine bifurcation or changing number of states in the system.

To detect transitions between states via low order moments of the probability distribution, one needs to be near criticalities where some of these states coalesce, as happens in pitchfork and limit point bifurcations. Otherwise, the transition occurs through finite jumps that have no clear signature in the autocorrelation function or variance levels. In these cases, degenerate fingerprinting can only be used for detecting (not forecasting) critical behaviour.

Recently, Ref. [29] suggested that autocorrelation alone is not enough to detect bifurcations, and in addition changes in variance should be monitored. Moreover, Ref. [30] showed that in finite-size time series of length N with power-law correlation exponent γ , variance $\Delta_N^2(s)$ and autocorrelation function $C_N(s)$ satisfy $\Delta_N^2(s) = C_N(s/s_x) + O(s^{-1})$, where

$$s_x = \left(\frac{2}{(2-\gamma)(1-\gamma)} \right)^{1/\gamma}.$$

This confirms our observation that the DFA-propagator is meaningful in monitoring changes in the data alongside the ACF-propagator. For example, most climatic time series possess power-law correlations, and in the presence of memory, especially with high fluctuation exponent and nonstationarities, lag-1 autocorrelations are not informative. In Fig. 2, we show data constructed of 11 chunks of red noise with increasing fluctuation exponent from 0.5 to 1.5, and one can see how the data becomes nonstationary, which is indicated by the trend in both propagators, as well as in the variance. Note that for the last two chunks, when ACF-propagator reaches its maximal value 1, DFA-propagator is able to detect further variability in the data, providing additional information about the changes in the variance.

3. Analysis of the test data

3.1. Data overview

In Figs. 3–5, we show the nine samples of artificial data provided for the test, their potential contour plots, and propagators (ACF- and DFA-propagators were calculated with sliding window of fixed size 500). Samples 1–7 demonstrate similar types of behaviour, with abrupt jumps between two main states, whereas samples 8 and 9 are structurally different. In addition, samples 4–7 have pronounced periodic variability of fluctuations within each state. Sample 8 demonstrates quasi-periodic behaviour, and sample 9 resembles ice-sheet dynamics with accumulation and ablation stages.

In the first sample (Fig. 3), the major abrupt transitions at 1500 and 2800 time units are detected by both propagators and by the potential plot, and also, because the potential plot considered various time scales (whereas propagators were calculated for the fixed window 500), it is able to detect the shorter-scale transition at 3100. In addition, the multi-scale potential plot shows, at bigger scales, that the data are quite homogeneous with pronounced double-well-potential behaviour.

Propagators can detect the noise-induced transitions when the background behaviour is non-critical (there are no nonstationarities, changes in memory, etc.). This happens, for example, in sample 1. Both propagators are far from critical value 1 in the beginning of the series, but then the abrupt change (noise-induced) kicks their values to 1, and they detect this noise-induced change.

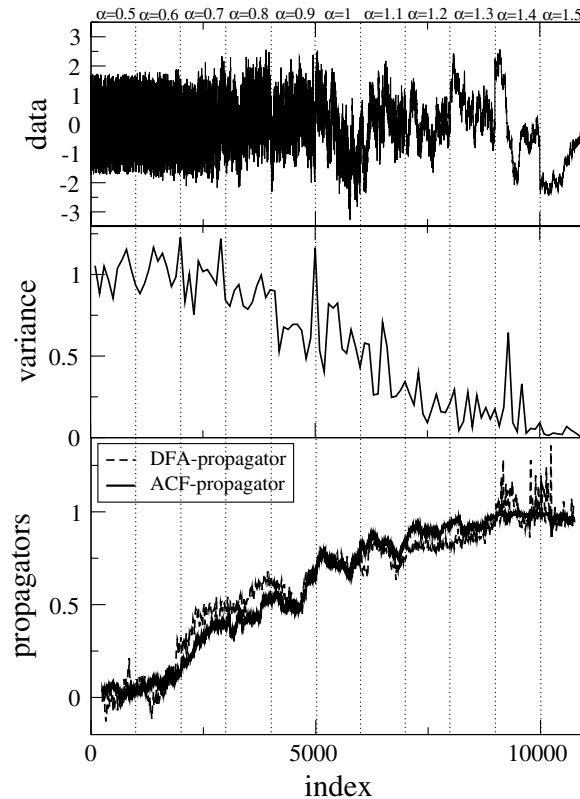


Fig. 2. Artificial data combined of eleven chunks of 1000 datapoints each with increasing fluctuation exponent α from 0.5 to 1.5 with step 0.1. Upper panel shows data with increasing nonstationarities due to increasing memory. Middle panel shows the variance calculated along the series in non-overlapping windows of size 100; due to the increasing memory the variance is decreasing. The bottom panel shows ACF- and DFA-propagators correctly detecting the critical behaviour of the data. Note that when ACF-propagator reaches critical value 1 and stabilises there, DFA-propagator is still able to detect variability in the data, and its fluctuations correspond to changes in variance in the middle panel.

In the sample 4, where there is irregularity at 2300–3000, this is detected by the potential plot. In sample 5, the detection of the pattern is also consistent in all three techniques.

Cyan in the potential colour plot denotes three wells in the potential, and this colour is observed at larger time scales in samples 4 and 6, where the time window aggregates large amount of data with several types of variability. In these samples, double-well-potential data also have high-frequency periodic oscillations, and in the histogram this additional variability appears as a third mode—this explains the higher number of states in that case. In sample 5, the waiting time of jumps between wells is much longer, and therefore the first half of the data is identified as one-well-potential (with periodic oscillations).

Analytically, proving non-potentiality would require proving non-existence of non-trivial solutions to particular equations for the Lyapunov function (which is also difficult to find). However, the potential analysis allows us to detect experimentally a sample without globally existing potential, which is indicated by variable patches in the potential contour plot (see sample 9, Fig. 5). This intermittent pattern in the contour plot is caused by varying potential or by complicated noise or both. In fact, the series 9 may be non-potential with damping.

3.2. Suggested models to reproduce the test data

We attempted to reproduce datasets with estimated properties using the Langevin equation with varied parameters and forcings in order to test our initial guesses about the provided data.

In Figs. 3–5, we show samples of the modelled data reconstructed from our analysis of the data provided for the test. We used four main types of the models:

- Langevin equation with white noise (samples 1–2).
- Langevin equation with white noise and periodic forcings (samples 4–7).
- Signal with three sine waves superimposed — analogous to the insolation forcing (sample 8)
- Non-potential model of ice-accumulation with damping (sample 9).

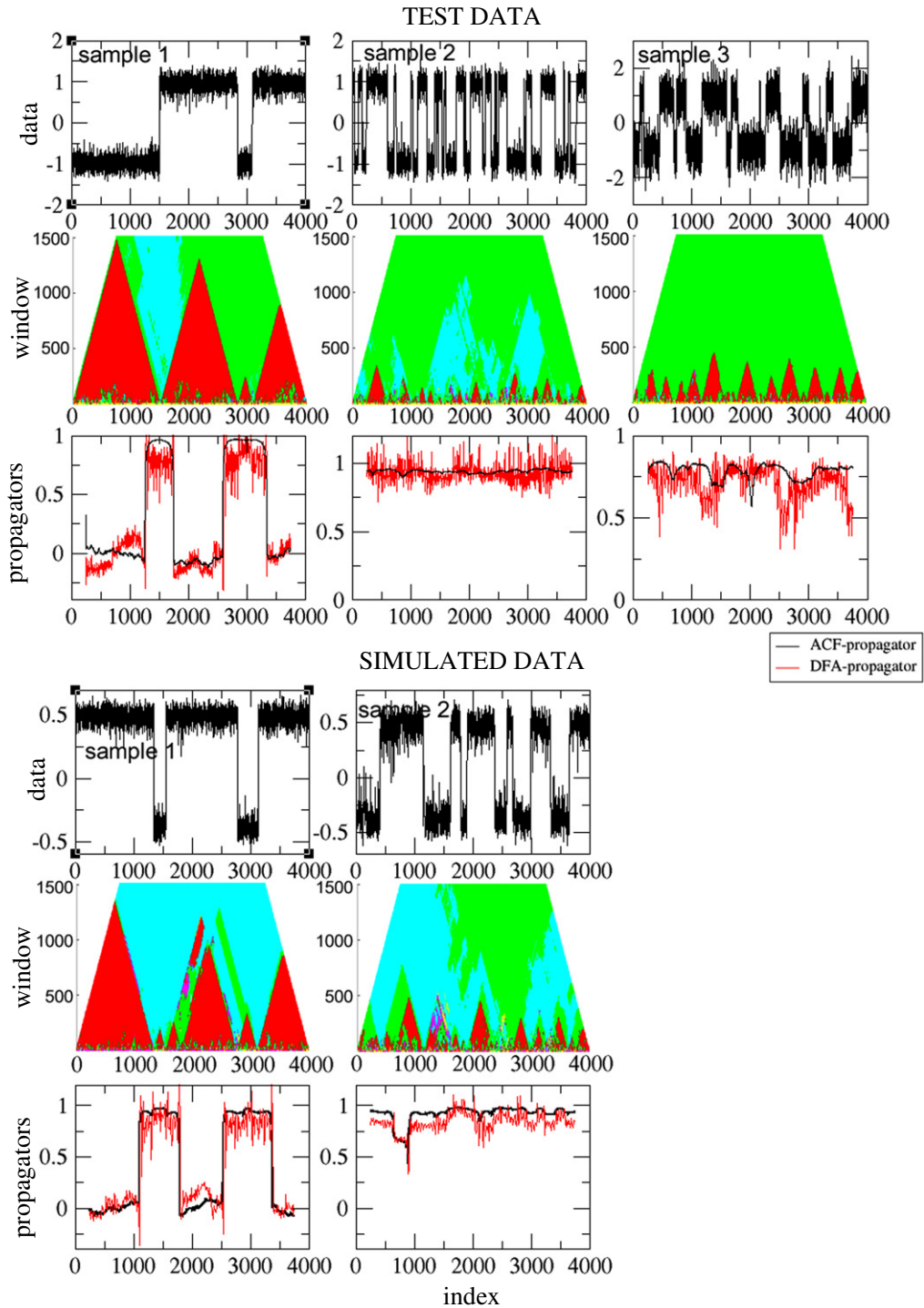


Fig. 3. Three samples (1–3) of artificial data provided for the test, and their potential colour plots and propagators (upper 9 panels). Artificial data reconstructed from the analysis of the test data, their potential plots and propagators (lower 6 panels). Since we were unable to reproduce the data presented in sample 3 of the test, the relevant panels are omitted. In the contour plots, detected one-well potential is denoted by red colour, double-well potential by green colour, triple-well potential by cyan colour, four-well potential by purple colour; the number of detected wells bigger than 4 is denoted by yellow colour. ACF- and DFA-propagators were calculated with a fixed sliding window of size 500.

To simulate the potential data in samples 1–2, 4–7, we used the Euler scheme with step size h , sampling every 5000 time steps. The potential model equations were as follows (potential $U(z) = z^4 - 0.16z^3 - 0.4z^2 + 0.016z$ for six simulated datasets; see Figs. 3–5):

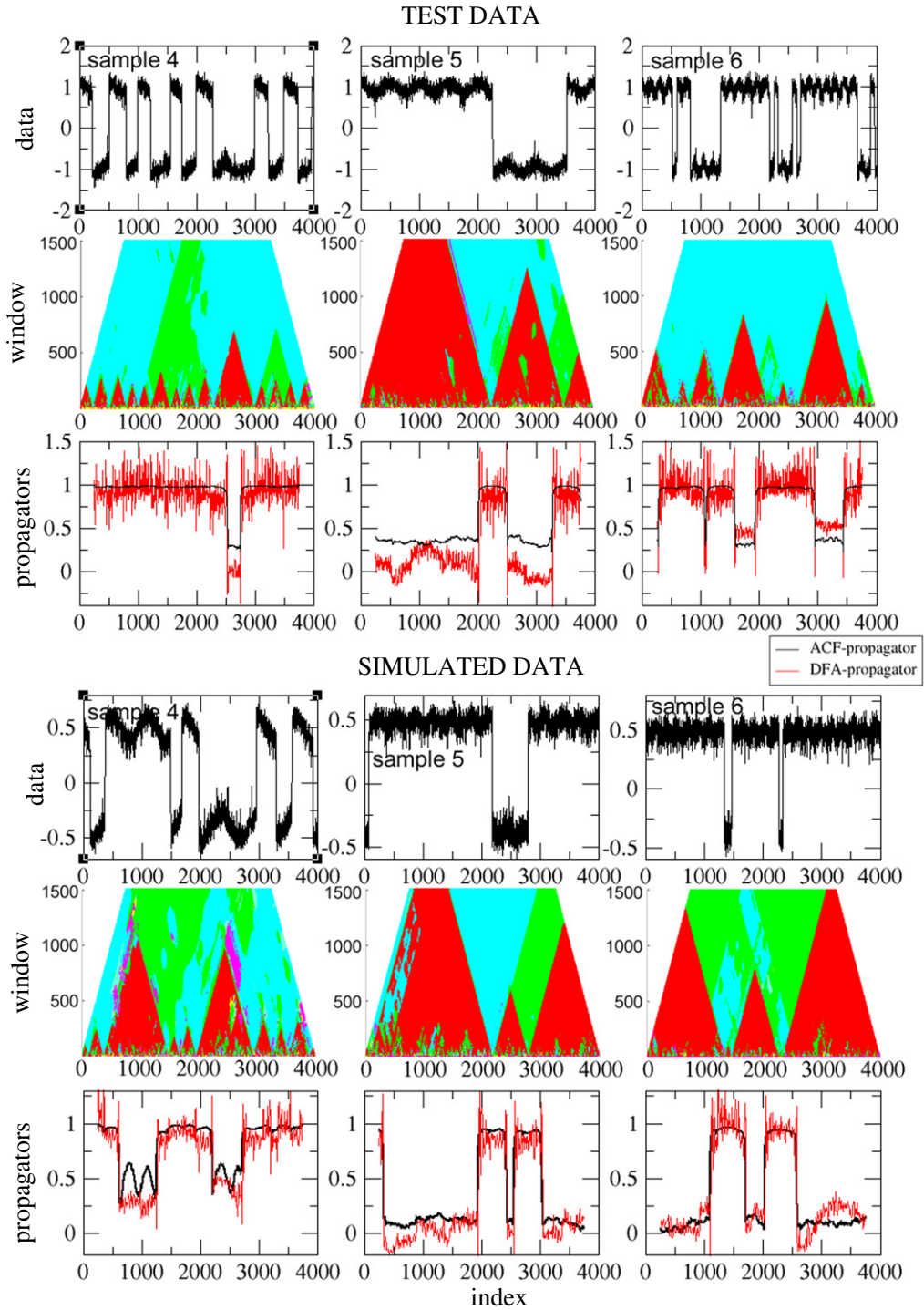


Fig. 4. Three samples (4–6) of artificial data provided for the test, and their potential colour plots and propagators (upper 9 panels). Artificial data reconstructed from the analysis of the test data, their potential plots and propagators (lower 9 panels). In the contour plots, detected one-well potential is denoted by red colour, double-well potential by green colour, triple-well potential by cyan colour, four-well potential by purple colour; the number of detected wells bigger than 4 is denoted by yellow colour. ACF- and DFA-propagators were calculated with a fixed sliding window of size 500.

Sample 1

$$dz = -U'(z)dt + \sigma dW, \quad \sigma = 0.1, \quad h = 0.008.$$

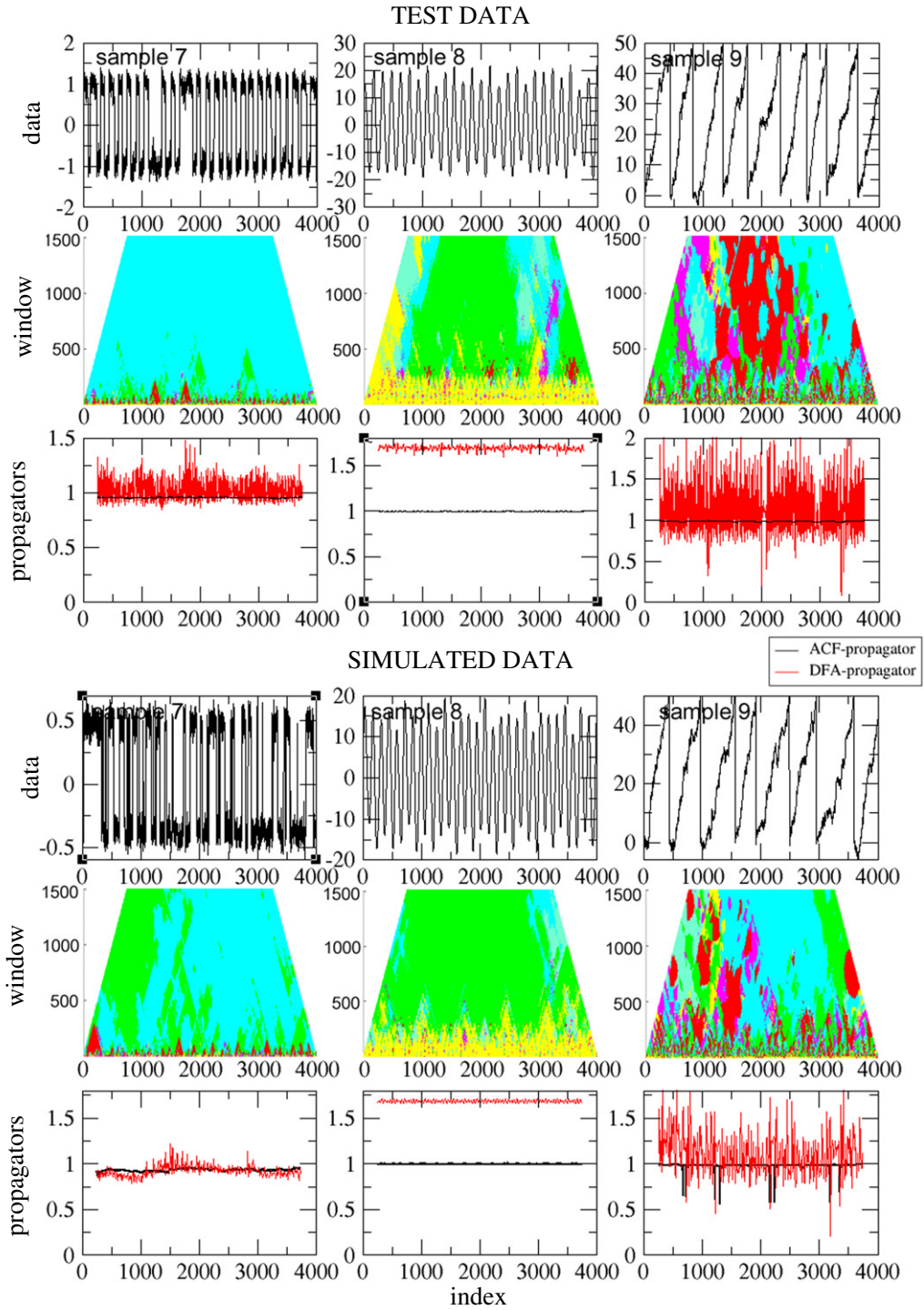


Fig. 5. Three samples (7–9) of artificial data provided for the test, and their potential colour plots and propagators (upper 9 panels). Artificial data reconstructed from the analysis of the test data, their potential plots and propagators (lower 9 panels). In the contour plots, detected one-well potential is denoted by red colour, double-well potential by green colour, triple-well potential by cyan colour, four-well-potential by purple colour; the number of detected wells bigger than 4 is denoted by yellow colour. ACF- and DFA-propagators were calculated with a fixed sliding window of size 500.

Sample 2

$$dz = -U'(z)dt + \sigma dW, \quad \sigma = 0.15, \quad h = 0.0001.$$

Table 1
Parameter values used for generating test samples 1–7.

Sample	σ	σ_1	A	τ
1	1.11	–	–	–
2	1.26	–	–	–
3	1.26	0.40	–	–
4	0.79	–	4.0	500
5	0.73	–	4.0	500
6	0.79	–	4.0	150
7	0.95	–	4.0	150

Sample 3

Although the potential analysis clearly detected the sample as double-well-potential, we were unable to find model parameters of the Langevin equation that would produce a similar time series. This may be due to a more complex model used to generate the series.

Sample 4

$$dz = -U'(z)dt + \sigma dW + A \sin(2\pi z/\tau), \quad \sigma = 0.1, \quad h = 0.0003, \quad A = -0.1, \quad \tau = 630.$$

Sample 5

$$dz = -U'(z)dt + \sigma dW + A \sin(2\pi z/\tau), \quad \sigma = 0.1, \quad h = 0.003, \quad A = -0.03, \quad \tau = 320.$$

Sample 6

$$dz = -U'(z)dt + \sigma dW + A \sin(2\pi z/\tau), \quad \sigma = 0.11, \quad h = 0.0032, \quad A = -0.03, \quad \tau = 200.$$

Sample 7

$$dz = -U'(z)dt + \sigma dW + A \sin(2\pi z/\tau), \quad \sigma = 0.05, \quad h = 0.93, \quad A = -0.03, \quad \tau = 125.$$

Two other datasets (samples 8 and 9) were simulated as follows

Sample 8

$$f(x) = K [\sin(2\pi x/\tau_1) + \sin(2\pi x/\tau_2) + \sin(2\pi x/\tau_3)], \quad K = 3, \quad \tau_1 = 340, \quad \tau_2 = 140, \quad \tau_3 = 80.$$

Sample 9

$$\frac{dz}{dt} = \begin{cases} (C - B)dt + \sigma dW, & C = A_0, \quad z < Z_{\max}, \\ z = Z_{\min}, & C = 0.1 \cdot A_0, \quad z \geq Z_{\max}, \end{cases}$$

$$A_0 = 0.1, \quad z_0 = 0, \quad Z_{\max} = 50, \quad Z_{\min} = 0, \quad B = 0.01, \quad \sigma = 0.5.$$

In Figs. 3–5, we show the propagators of the simulated data, and we note that, similarly to the test data, the propagators are able to detect the abrupt transitions. The main patterns of the simulated potential plots are the same as for the test data, and the non-potential character of the last sample is confirmed by the same patchy behaviour as in the test sample.

3.3. Actual modelled data

The generated data are the generic non-trivial models of bimodal behaviour, which is seen in many real measured data series—in much the same way as an Ornstein–Uhlenbeck process is generic for a noisy linear process.

The nine samples of time series generated for the independent test were as follows:

The samples 1, 2 were generated by a Langevin equation:

$$dz = -U'(z)dt + \sigma dW, \quad (2)$$

where the drift term is generated by a symmetric double-well potential

$$U(z) = 5z^4 - 10z^2. \quad (3)$$

Sample 3 is the sum of a process generated as sample 2 and a white noise with intensity σ_1 . samples 4–7 are stochastic resonance processes generated by:

$$dz = (-U'(z) + A \sin(2\pi z/\tau))dt + \sigma dW.$$

The parameters are listed in Table 1.

Sample 8 is a realisation of the x component of the chaotic Rössler system:

$$\dot{x} = -y - z$$

$$\dot{y} = x + ay$$

$$\dot{z} = b + -z(x - c)$$

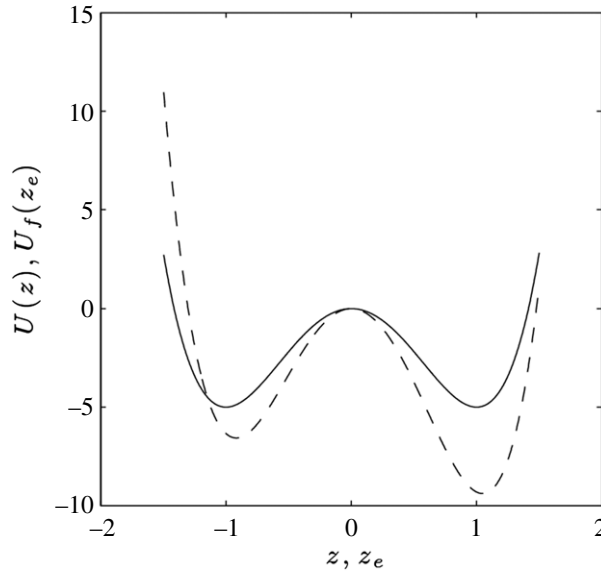


Fig. 6. Comparison between the potential used for generating the test samples (full) and the estimated potential used for generating the 'simulated' samples (dashed). All samples of potential test data were simulated using double-well-potential model (1).

with $(a, b, c) = (0.1, 0.1, 14)$. Finally, sample 9 is generated by a Langevin equation with a constant drift and a jump when a threshold is crossed:

$$dz = \begin{cases} Cdt + \sigma dW, & z < Z_{\max}, \\ -(Z_{\max} - Z_{\min}), & z \geq Z_{\max}, \end{cases}$$

where $C = 0.1$, $Z_{\max} = 50$, $Z_{\min} = 0$, and $\sigma = 0.5$.

4. Discussion: comparing test and estimated generators

The tested methods identified the governing equations for test data samples 1 and 2 to be generated using a constant potential and test data samples 4–6 using the same potential and an additional time varying component, which are thus stochastic resonances. The generating equations for these 6 first test samples, except for sample 3, were essentially reconstructed correctly.

In order to compare the suggested models with those actually used to produce data from Eq. (2), the following transformation is used,

$$\begin{aligned} dz &= -\lambda U'(z) d(t/\lambda) + \sqrt{\lambda} \sigma d(W/\sqrt{\lambda}) \\ &= -\tilde{U}'(z) d\tilde{t} + \tilde{\sigma} d\tilde{W}, \end{aligned}$$

where $\tilde{t} = t/\lambda$. The potential used above thus rescales to $U_e(z) = \lambda(z^4 - 0.16z^3 - 0.4z^2 + 0.016z)$.

Furthermore, a further rescaling of the estimated parameters is necessary. Defining $U(z) = 5z^4 - 10z^2$ and $U_e(z) = z^4 - 0.16z^3 - 0.4z^2 + 0.016z$, the rescaling is done by using the variable $z_e = \mu(z - z_0)$, such that we approximately have $\tilde{U}'_e(z_e) \equiv U_e(z_e/\mu + z_0) = 0 \Leftrightarrow U'(z) = 0$. The minima for $U_e(z)$ are $z_- = -0.402$ and $z_+ = 0.502$, while the local maximum is at $z_0 = 0.020$. Since the corresponding extrema for the potential $U(z)$ are in ± 1 and 0, respectively, we can use $\mu = 1/0.46 = 2.17$. By this change of variables the Langevin equation for the variable z_e becomes $dz_e = -\mu^2 \tilde{U}'(z_e) dt + (\mu\sigma) dW \equiv -U_f(z_e) dt + \sigma_f dW$.

Finally, we have for sample 1

$$U_f(z) = 8.5z^4 - 1.5z^3 - 16.3z^2.$$

The “true” potential (full) and the estimated potential (dashed) are shown in Fig. 6.

The sample 7 has a step size too large to be well represented by the equation stated above in Section 3.3, it must rather be considered a difference equation generated by the map:

$$z_{n+1} = f^{(5000)}(z_n),$$

where $f(z) = z - h(4z^3 - 0.48z^2 - 0.8z + 0.016) + \sqrt{h}\sigma\eta$ and η is a Gaussian random variable with zero mean and unit variance. In this case, the harmonic component $A \sin(2\pi z/\tau)$ is completely negligible, since $|A/h| \approx 10^{-6}$, the estimated period for the harmonic $\tau = 125$ time units is, even though close to the correct stochastic resonance period, not relevant for their test sample 7, which does not have a periodic component.

Table 2
Estimated parameter values used for samples 1–6.

Estim.	σ_e	$\sigma_e\sqrt{\lambda}$	$\sigma_f = \sigma_e\sqrt{\lambda}\mu$	$-A$	$-A\lambda$	τ
1	0.10	0.63	1.37	–	–	–
2	0.15	0.11	0.23	–	–	–
4	0.10	0.12	0.27	0.1	4.0	630
5	0.10	0.39	0.84	0.03	1.2	320
6	0.11	0.44	0.96	0.03	1.2	200

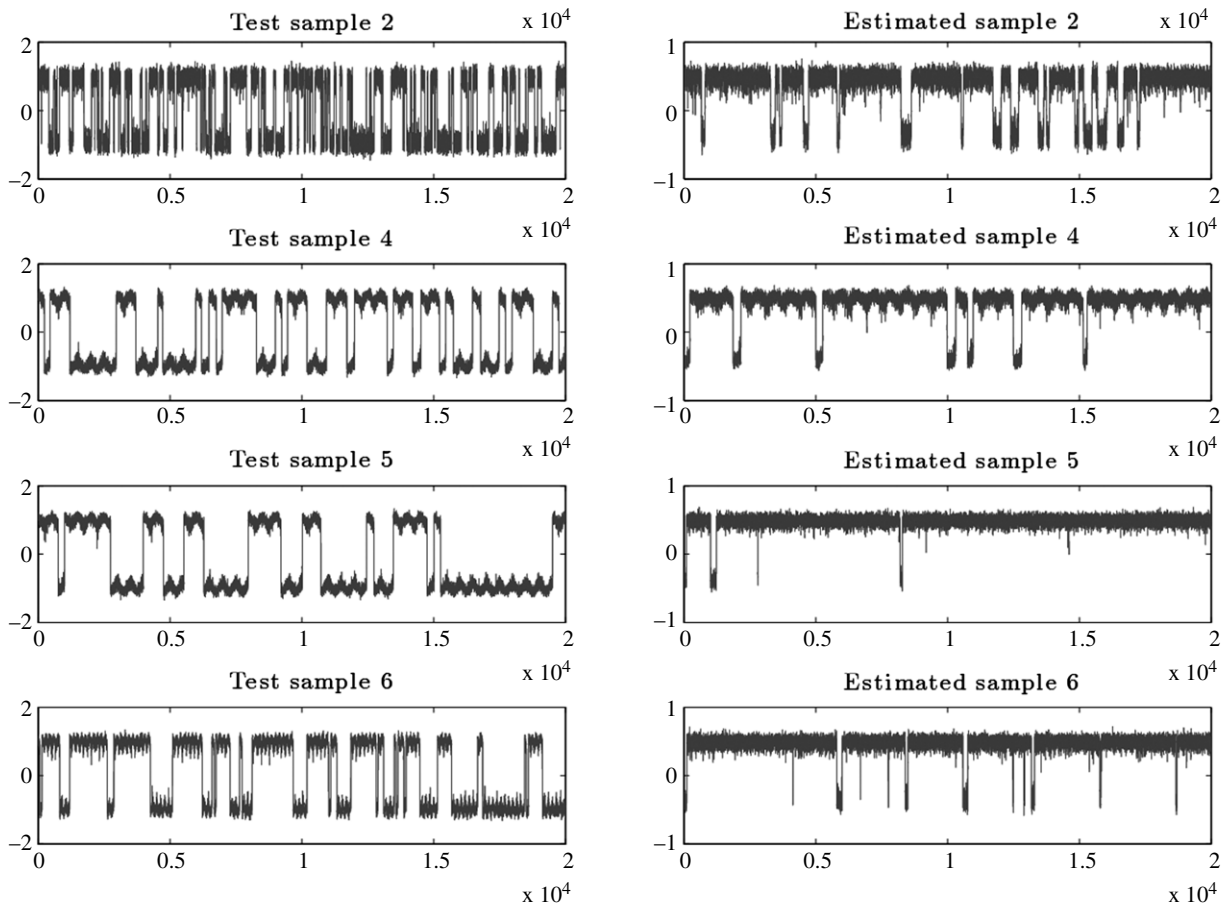


Fig. 7. Comparison between realisations of the test dynamics (left column) and the ‘simulated’ dynamics (right column). This is now for a time period five times longer than what was provided to my co-authors.

Sample 8 is chaotic, but even though the suggested model is not correct, the method identifies correctly that the data cannot be generated by an effective potential.

The analysis of sample 9 seems rather uninformative, suggesting that the correct guess on the very simple governing equation is more based on direct inspection of the data than on the result of the analysis.

Comparison between the actual and the estimated parameters (see Table 1, Table 2 and Fig. 6) shows that the method is not superior in quantitative estimation of parameters. This is always a delicate problem for short samples. To illustrate this data samples equivalent to samples 1, 2, 4, 5 and 6 but now five times longer than the ones provided for analysis, were generated. This was done using the same parameters as before for both the provided and the estimated samples. The realisations are shown in Fig. 7. From these, clear differences are seen. With the knowledge of the type of generating dynamics (2), which the analysis did not have, the uncertainty in estimation of parameters can be assessed by Monte Carlo simulation. To illustrate this, the (maximum likelihood) potential from different realisations of length 4000 time units of the dynamics generating test sample 1 has been calculated. The potential is calculated from the stationary probability density $p(z)$ by solving the Fokker–Planck equation with respect to the drift term:

$$U(z) = -\frac{\sigma^2}{2} \log p(z).$$

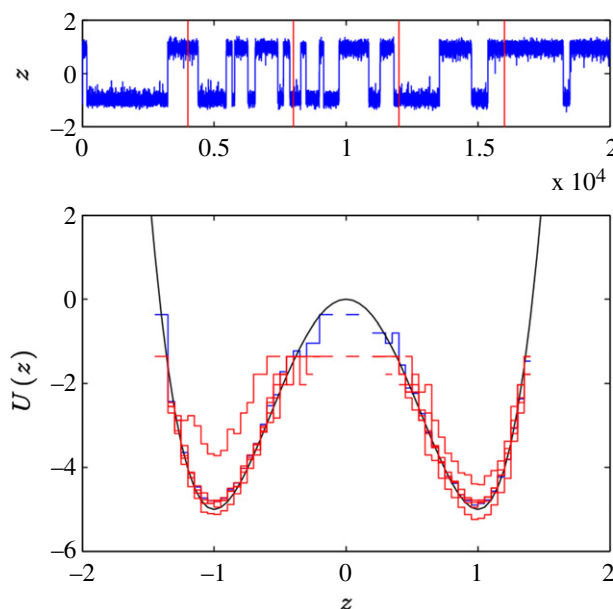


Fig. 8. A 20000 time unit realisation of the process generating test sample 1. The potential (black curve) is estimated from the probability density by solving the Fokker–Planck equation. The blue estimate is obtained by using the full time series, while the five red curves are obtained by using only data within the 4000 time unit slots indicated by vertical lines in the top panel.

For simplicity it is assumed that σ , which can be estimated independently, is known. In Fig. 8, the potentials estimated in five 4000 time unit sections of the 20000 time unit realisation (in red) are compared with the ‘true’ potential (black curve). The estimate in blue is obtained by using the probability density of the full 20000 time unit realisation. The estimate with a too shallow left well corresponds to the last 4000 time unit period, where by chance the system is mostly in the right well. The observation that the red curves in Fig. 8 are closer to the ‘true’ potential than the dashed curve in Fig. 6 is an indication that the method is mainly applicable for qualitative identification of time varying effective potentials generating time series.

5. Conclusion

The techniques tested here were able to successfully detect changes of state and abrupt potential jumps in the test time series. The potential analysis was able to detect the number of states and to recognise a sample with non-potential behaviour. The modified degenerate fingerprinting was able to detect but not forecast changes of state, due to their abrupt and noise-induced character in the test data. Based on this analysis and visual observation of the test series, statistical models were proposed that reproduced the behaviour of the test series with analogous properties.

The generating equations for the six first test samples, except for sample 3, were reconstructed correct. For sample 8, the model suggested was not correct, but the tested statistical methods identified correctly that the data cannot be generated by an effective potential. The guess on sample 9 was correct, but more based on direct inspection of the data than on the result of the analysis.

Overall, the potential analysis method is suitable for detecting potentials in cases where such potentials are generating the data. However, for the case (sample 3), where an additional uncorrelated noise masks the dynamics, the method fails in detecting the potential. Hence the method can provide false rejection if the null hypothesis is that a generating potential exists. Conversely, the method has (for the test datasets examined) not made any false rejections of the alternative hypothesis that the data was not generated by an effective potential.

We conclude that the methods of potential analysis and degenerate fingerprinting are suitable for analysing system potentials and are able to detect noise-driven jumps between system states. We note that the quantitative estimation of potential parameters may be inaccurate in the case of short data samples. We suggest that, although unconventional, independent testing allows one to objectively assess the reliability and limitations of methods of time series analysis. This is particularly important in the area of climate change research, where knowledge is highly contested (especially in a social context). A public web database detailing general rules and procedures could be set up if others are willing to test their methods in this manner.

Acknowledgements

The idea of the blind test arose during the research visit of P. D. Ditlevsen and G. Lohmann to the University of East Anglia in September 2009. P. D. Ditlevsen proposed the test as a scientific challenge for the new techniques of time series analysis.

The research was supported by NERC through the project “Detecting and classifying bifurcations in the climate system” (NE/F005474/1) and by AXA Research Fund through a postdoctoral fellowship for VL. The research presented in this paper was carried out on the High Performance Computing Cluster supported by the Research Computing Service at the University of East Anglia.

References

- [1] T. Schreiber, A. Schmitz, Surrogate time series, *Physica D*, 142, 346–382.
- [2] T.Z. Liu, Blind testing of rock varnish microstratigraphy as a chronometric indicator: results on late Quaternary lava flows in the Mojave Desert, California, *Geomorphology* 53 (3–4) (2003) 209–234.
- [3] G. Parkin, G. O'Donnell, J. Ewen, J.C. Bathurst, P.E. O'Connell, J. Lavabre, Validation of catchment models for predicting land-use and climate change impacts. 2. Case study for a Mediterranean catchment, *J. Hydrol.*, 175 (1–4) 595–613.
- [4] J.C. Bathurst, J. Ewen, G. Parkin, P.E. O'Connell, J.D. Cooper, Validation of catchment models for predicting land-use and climate change impacts. 3. Blind validation for internal and outlet responses, *J. Hydrol.*, 287 (1–4) 74–94.
- [5] P.D. Ditlevsen, Observation of α -stable noise induced millennial climate changes from an ice-core record, *Geophys. Res. Lett.* 26 (10) (1999) 1441–1444.
- [6] D. Paillard, Glacial cycles: toward a new paradigm, *Rev. Geophys.* 39 (3) (2001) 325–346.
- [7] F. Kwasiński, G. Lohmann, Deriving dynamical models from paleoclimatic records: application to glacial millennial-scale climate transitions, *Phys. Rev. E* 80 (2009) 066104.
- [8] V.N. Livina, F. F. Kwasiński, T.M. Lenton, Potential analysis reveals changing number of climate states during the last 60 kyr, *Clim. Past* 6 (2010) 77–82.
- [9] D.J. Peavoy, C. Franzke, Bayesian analysis of rapid climate change during the last glacial using Greenland $\delta^{18}O$ data, *Clim. Past* 6 (2010) 787–794.
- [10] H. Held, T. Kleinen, Detection of climate system bifurcations by degenerate fingerprinting, *Geophys. Res. Lett.* 31 (2004) L23207.
- [11] V. Livina, T. Lenton, A modified method for detecting incipient bifurcations in a dynamical system, *Geophys. Res. Lett.* 34 (2007) L03712.
- [12] M. Budyko, The effect of solar radiation variations on the climate of the earth, *Tellus XXI* 5 (1969) 611–619.
- [13] W. Sellers, A global climatic model based on the energy balance of the earth-atmosphere system (1969), *J. Appl. Meteorol.* 8 (1969) 392–400.
- [14] K. Hasselmann, Stochastic climate models. part I. theory, *Tellus XXVIII* 6 (1976) 473–485.
- [15] K. Hasselmann, Climate change: linear and nonlinear signature, *Nature* 398 (1999) 755.
- [16] C. Nicolis, G. Nicolis, Stochastic aspects of climate transitions—additive fluctuations, *Tellus* 33 (1981) 225–234.
- [17] R. Benzi, A. Sutera, A. Vulpiani, The mechanism of stochastic resonance, *J. Phys. A: Math. Gen.* 14 (1981) L453–L457.
- [18] V.N. Livina, F. Kwasiński, G. Lohmann, J.W. Kantelhardt, T.M. Lenton, Changing climate states and stability: from Pliocene to present, *Clim. Dyn.* (2011) doi:10.1007/s00382-010-0980-2.
- [19] S.J. Julier, J.K. Uhlmann, Unscented filtering and nonlinear estimations, *Proc. IEEE* 92 (3) (2004) 401.
- [20] H. Voss, J. Timmer, J. Kurths, Nonlinear dynamical system identification from uncertain and indirect measurements, *Internat. J. Bifur. Chaos* 14 (6) (2004) 1905–1933.
- [21] C.-K. Peng, S.V. Buldyrev, S. Havlin, M. Simons, H.E. Stanley, A.L. Goldberger, Mosaic organization of DNA nucleotides, *Phys. Rev. E* 49 (1994) 1685–1689.
- [22] S.V. Buldyrev, A.L. Goldberger, S. Havlin, R.N. Mantegna, M. Matsa, C.-K. Peng, S. Simons, H.E. Stanley, Long-range correlations properties of coding and noncoding DNA sequences: GenBank analysis, *Phys. Rev. E* 51 (1995) 5084–5087.
- [23] C.-K. Peng, S. Havlin, H.E. Stanley, A.L. Goldberger, Quantification of scaling exponents and crossover phenomena in nonstationary heartbeat time series, *Chaos* 5 (1995) 82–87.
- [24] A. Bunde, S. Havlin, J.W. Kantelhardt, T. Penzel, J.H. Peter, K. Voigt, Correlated and uncorrelated regions in heart-rate fluctuations during sleep, *Phys. Rev. Lett.* 85 (2000) 3736–3739.
- [25] Y. Ashkenazy, P.Ch. Ivanov, S. Havlin, C.-K. Peng, A.L. Goldberger, H.E. Stanley, Magnitude and sign correlations in heartbeat fluctuations, *Phys. Rev. Lett.* 86 (2001) 1900–1903.
- [26] E. Koscielny-Bunde, A. Bunde, S. Havlin, H.E. Roman, Y. Goldreich, H.-J. Schellnhuber, Indication of a universal persistence law governing atmospheric variability, *Phys. Rev. Lett.* 81 (1998) 729–732.
- [27] J.W. Kantelhardt, E. Koscielny-Bunde, D. Rybski, P. Braun, A. Bunde, S. Havlin, Long-term persistence and multifractality of precipitation and river runoff records, *J. Geophys. Res. Atmos.* 111 (D1) (2006) D01106.
- [28] H. Makse, S. Havlin, S. Schwartz, M. E. Stanley, Method for generating long-range correlations for large systems, *Phys. Rev. E* 53 (5) (1996) 5445–5449.
- [29] P.D. Ditlevsen, S.J. Johnsen, Tipping points: early warning and wishful thinking, *Geophys. Res. Lett.* 37 (2010) L19703. doi:10.1029/2010GL044486.
- [30] S. Lennartz, A. Bunde, Eliminating finite-size effects and detecting the amount of white noise in short records with long-term memory, *Phys. Rev. E* 79 (2009) 066101.