



OPEN

Can we predict the unpredictable?

Abbas Golestani¹ & Robin Gras^{1,2,3}

SUBJECT AREAS:

EPILEPSY

PROJECTION AND PREDICTION

COMPUTER SCIENCE

Received
8 July 2014Accepted
9 October 2014Published
30 October 2014Correspondence and
requests for materials
should be addressed to
A.G. (golesta@
uwindsor.ca)

¹School of Computer Science, University of Windsor, ON N9B 3P4, Canada, ²Department of Biology, University of Windsor, ON N9B 3P4, Canada, ³Great Lakes Institutes for Environmental Research, University of Windsor, ON N9B 3P4, Canada.

Time series forecasting is of fundamental importance for a variety of domains including the prediction of earthquakes, financial market prediction, and the prediction of epileptic seizures. We present an original approach that brings a novel perspective to the field of long-term time series forecasting. Nonlinear properties of a time series are evaluated and used for long-term predictions. We used financial time series, medical time series and climate time series to evaluate our method. The results we obtained show that the long-term prediction of complex nonlinear time series is no longer unrealistic. The new method has the ability to predict the long-term evolutionary trend of stock market time series, and it attained an accuracy level with 100% sensitivity and specificity for the prediction of epileptic seizures up to 17 minutes in advance based on data from 21 epileptic patients. Our new method also predicted the trend of increasing global temperature in the last 30 years with a high level of accuracy. Thus, our method for making long-term time series predictions is vastly superior to existing methods. We therefore believe that our proposed method has the potential to be applied to many other domains to generate accurate and useful long-term predictions.

The prediction of future values in a complex time series is a major interest for scientists^{1–17}, with applications to various scientific fields^{18–20}. There are many natural phenomena that require a prediction algorithm for answering important questions, such as estimating future population variations, predicting the orbits of astronomical objects and predicting the occurrence of seismic waves. The prediction of twenty-first century global temperature rise would be a valuable resource for policy makers and planners⁶. Population projections may be used to predict species extinction before they reach a crisis point⁵. Moreover, prediction is an ongoing and pressing problem in the forecasting of economic time series⁸. In the medical sciences, there are also many applications for which an efficient prediction algorithm could save lives. A large number of time series obtained from monitoring the human body can be used as a basis for the decision-making process to treat or prevent grave diseases such as epilepsy or Alzheimer's disease^{1–3}.

It has been shown that data generated by such natural phenomena often behave chaotically¹¹. Although chaotic behaviours are deterministic, their complex properties make it difficult to distinguish them from random behaviour. Chaotic behaviours are known to be strongly dependent on initial conditions; small changes in initial conditions can possibly lead to immense changes in subsequent time steps and are particularly difficult to predict. Because the exact conditions for many natural phenomena are not known and the properties of a chaotic time series are very complex, it is difficult to model these systems.

Most of the existing methods for complex time series prediction are based on modelling the time series to predict future values, although there are other types of methods such as agent-based simulations that model the system generating the time series²¹. Model-based approaches can be classified into two main domains: linear modelling such as in ARIMA (autoregressive integrated moving average)¹⁸ and nonlinear modelling such as in MLP (multi-layer perceptron)²² and GARCH (generalised autoregressive conditional heteroskedasticity)²³ (for details, see Supplementary Information, section 1). However, other studies have concluded that there is no clear evidence in favour of nonlinear models over linear models in terms of forecasting performance⁷. Regardless, there is no robust procedure that can produce an accurate model for chaotic time series. For all of these methods, the prediction error increases dramatically with the number of time points predicted^{7,9,21}. Therefore, most of the existing methods focus on very short-term predictions to obtain a reasonable level of accuracy. None of the existing methods demonstrate an acceptable level of accuracy for long-term prediction¹². For example, for financial time series predictions, most methods can predict only one step ahead, which is not very helpful for acting against a financial crisis before it occurs^{10,12,21}. To address this deficiency in existing methods, we propose a novel approach to making long-term time series predictions (see Methods), GenericPred²⁴, with applications to financial time series, medical diagnosis and global temperature prediction.



Results

For the DJIA (Dow Jones Industrial Average) time series between 1993 and 2001, when markets were stable with no major changes and no financial crisis, our proposed GenericPred method predicts the trend with very high accuracy, whereas the predictions based on other methods strongly and rapidly diverge from the real data (see Fig. 1a and Supplementary Table 1). GenericPred is the only method able to discover the decreasing trend corresponding to the financial crisis, while the three other methods predicted a growth in the stock market (see Fig. 1b and Supplementary Table 1). Although the 2009 financial crisis data were used for training in the third time period, when the recession was in the middle of the considered range, the GenericPred method successfully discovered the general trends for the next 500 days, with a particularly high accuracy for the first 300 days, effectively predicting the increase in the stock market (see Fig. 1c and Supplementary Table 1). The other three methods failed to predict the trend.

The GenericPred method's prediction errors were significantly lower than the other methods for all three periods with respect to both short-range and long-range predictions (see Supplementary Table 1). Moreover, the GenericPred predictions were more stable with a constant lower standard deviation, regardless of whether the target data lie before the recession, during the recession, or after the recession. The prediction error for the first 200 time steps is especially smaller than that of the other methods. For the three considered periods, GenericPred also significantly outperformed L-FABS²¹ and the MLP²² model for short-term predictions (see Supplementary Table 1). GenericPred is also stable given that it has a low sensitivity to the starting point of the prediction (see Supplementary Figure 2).

For the EEG time series, the P&H chaoticity values²⁵ were predicted using GenericPred (see Supplementary Figures 2 and 3) on a constant-length (20 minutes) sliding window (the window moves every 20 seconds) of the five EEG time series for all the patients. During a seizure, a peak in the P&H values obtained from the EEG time series appeared. Based on the analysis of all 21 patients, a threshold for the prediction of seizures (P&H value equal to 2.4) was determined based on patient data (see Supplementary Information). Using this threshold, the GenericPred method can predict the epileptic seizure with 100% sensitivity and specificity up to 17 minutes in advance (see Table 1). The same results have been obtained by considering the data of any five electrodes independently. This represents a considerable improvement over one of the best current predictive methods, which only achieves accuracy levels of 73% sensitivity and 67% specificity for 10 patients within a 1–10 minute range²⁶.

For the global temperature anomalies prediction, unlike ARIMA, GenericPred accurately predicted the increasing trend in the last 30 years (Fig. 2a). Moreover, most of the successive peaks and depressions were predicted with a high degree of precision. The mean square error for GenericPred is 0.64, whereas for ARIMA, it is 1.6. The GARCH and VAR methods were unable to make predictions due to insufficient data.

The existing dedicated forecasting models for global temperature anomalies predict that, as the world consumes more fossil fuel, greenhouse gas concentrations will continue to increase and the Earth's average surface temperature will continue to rise²⁷. Based on recent predictions, average surface temperatures could rise between 2°C and 6°C by the end of the 21st century^{6,28}. We predicted the global temperature anomaly until the end of the 21st century (2014–2100) (Fig. 2b). GenericPred predicted an average anomaly of 2.5°C for the years 2085–2100, which is in accordance with the predictions of the dedicated models.

Our approach demonstrates a significant gain in accuracy over traditional methods with respect to predicting different DJIA time series for both short- and long-term predictions. Its ability to predict the evolutionary trend of the stock market is vastly superior to the

predictive ability of existing methods. Further, the ability of our GenericPred method to predict epileptic seizures could be a major breakthrough in terms of managing epilepsy, providing, for the first time, a robust method for predicting an epileptic seizure long before its occurrence. Finally, the success of GenericPred's predictions of the trend of increasing global temperature shows that this method has the potential to be a powerful tool for controlling the global warming phenomenon. Another advantage of our approach is that it does not rely on a complex model of the original time series and is therefore highly general and highly computationally efficient.

This method provides a first step towards accurate and comprehensive time series long-term predictions. Although, with respect to long-term predictions, it is impossible to predict exact values, GenericPred's performance shows great potential for predicting time series' trends.

Methods

Several researchers emphasise the potential of market predictions to improve important financial decisions²⁹, from helping businesses make sounder investment decisions to helping governments make more efficient fiscal and monetary policy decisions⁵. These time series are amongst the most complex time series because of the number of parameters involved. Our results are compared with respect to long-term predictions with ARIMA, GARCH, and VAR³⁰, which are the most widely used and most efficient methods for making long-term time series predictions. We also compared our results for short-term predictions with those obtained by two existing methods: the Learning Financial Agent Based Simulator (L-FABS)²¹ and the MLP model²².

For the first period, we considered the DJIA (Dow Jones Industrial Average) time series between 1993 and 2001, when markets were stable with no major changes and no financial crisis. In the second period considered, the US stock market peaked in October 2007, but by March 2009, the Dow Jones average had reached its minimum, which reflects the most serious effects of a financial crisis. In the third period (August 2004–August 2012), the recession was in the middle of the considered range.

Another important application of time series predictions is in medical science. Approximately 1% of the world population suffers from epilepsy³¹. Epileptic seizures are the result of unusual and irregular neuronal activity in the brain^{32,33}. Many recent methods have been proposed for predicting epileptic seizure^{26,34,35} but none of them as shown their ability to perform accurate predictions more than 10 minutes in advance on a large number of patients. To evaluate the performance of our new method for predicting epileptic seizures, we examined the EEG time series measured by five electrodes, generating five different time series, for 21 patients. For each EEG time series, the exact time of the seizure is known.

Predicting the monthly records of global temperature anomalies is currently one of the most pressing and controversial environmental concerns³⁶. As a third experiment, we used the global temperature anomaly data from 1880 to 1983 to train for the prediction of global temperatures during 1983–2013. Global temperature anomaly data come from the Global Historical Climatology Network-Monthly (GHCN-M) data set and International Comprehensive Ocean-Atmosphere Data Set (ICOADS), which have data from 1880 to the present. These two datasets are blended into a single product to produce the combined global land and ocean temperature anomalies.

Our new method for complex time series prediction is based on the concepts of chaos theory and an optimisation process. The general idea is to extract a unique characteristic from an existing time series that somehow represents the behaviour of the time series and to subsequently generate successive new values that continue the time series, each value minimising the difference between the characteristic of the new time series and the initial one. The details of the GenericPred method for long-term time series prediction are as follows. We consider a time series S_N :

$$S_N = \{x_1, x_2, \dots, x_N\} \quad (1)$$

A nonlinear measure $V()$ is computed on S_N . The fractal dimension³⁷ and the Lyapunov exponent³⁸ are examples of such nonlinear measures that return a single value for a time series. A possible mapping may be required, forming a new time series $S_N^m = \{y_L, y_{L+1}, \dots, y_N\}$, for different applications as follows:

$$y_i = V(S_{i-L+1,i}), L \leq i \leq N \quad \text{where} \quad S_{i-L+1,i} = \{y_{i-L+1}, y_{i-L+2}, \dots, y_i\} \quad (2)$$

otherwise, $S_N^m = S_N$, where $0 < L < N$ is the size of a sliding window used to compute the local level of chaos measured by $V()$. Therefore, when the mapping is applied, the new considered time series S_N^m corresponds to the variation in time of the local nonlinear measure in the initial time series S_N .

We consider $V(S_N^m)$ as a reference value that will be used for predicting the next k values of the time series:

$$y_{N+i}, 1 \leq i \leq k \quad (3)$$

The parameter σ of a normal distribution $N(y_i, \sigma^2)$ is estimated by computing the variation between every two consecutive values (y_i to y_{i+1}) of the time series S_N^m . This

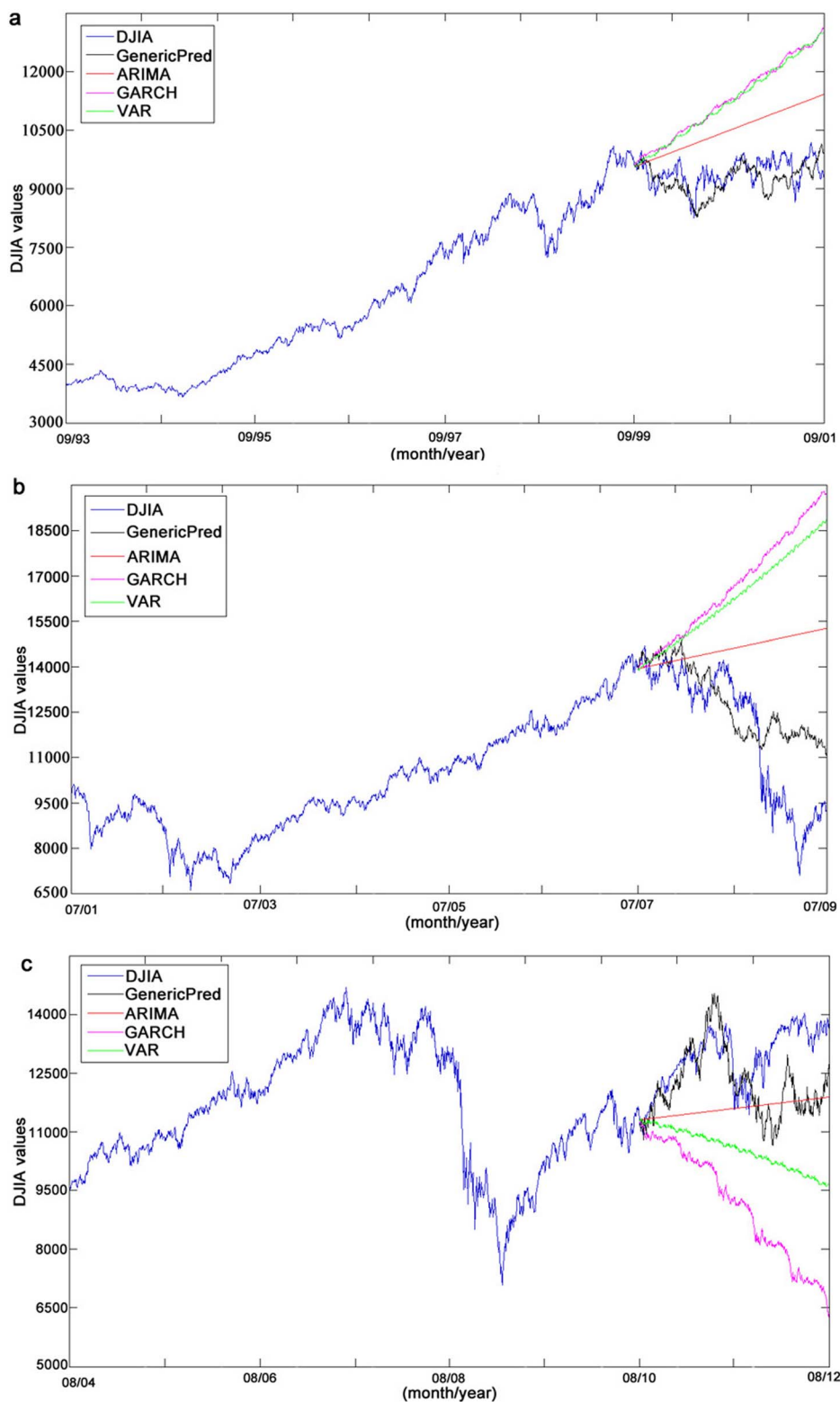


Figure 1 | Dow Jones Industrial Average stock market index prediction. We examined the DJIA time series with respect to the daily closing values of the DJIA for three time periods: (a) September 1993-September 2001 for the prediction of DJIA values before the 2009 financial crisis, (b) July 2001-July 2009 for the prediction of the financial crisis in 2009 and (c) August 2004- August 2012 for the prediction of DJIA values after the financial crisis in 2009. For each time series, 1500 time steps (approximately 6 years) were analysed to predict the next 500 time steps (approximately two years).

distribution represents the probability distribution $P(y_i|y_{i-1})$ (see Fig. 3). Several data sets have been considered to determine that a normal distribution is a good approximation of the real distribution. However, the same method has been applied using other distributions without significant degradation in the prediction.

For predicting y_{N+i} a set $Pos(y_{N+i})$ of N_{rand} random values are generated following the distribution $N(y_{N+i-1}, \sigma^2)$ (Fig. 3):

$$Pos(y_{N+i}) = \{y_{N+i}^j, 1 \leq j \leq N_{rand}\} \tag{4}$$



Table 1 | Sensitivity and specificity of epileptic seizure prediction for 21 patients for different lengths of prediction. For each patient, one positive and 10 negative samples were constructed. The positive sample contains one epileptic seizure event, and the 10 negative samples are seizure-free. Therefore, there are 21 positive and 210 negative samples in total that were used to compute the specificity and the sensitivity accuracy levels

Length of prediction before seizure	Sensitivity	Specificity
16 minutes ± 7 seconds	100%	100%
17 minutes ± 7 seconds	100%	100%
18 minutes ± 13 seconds	85%	100%
19 minutes ± 13 seconds	57%	100%
20 minutes ± 43 seconds	43%	100%

N_{rand} is a parameter that can impact the quality of the prediction because having more values will increase the chance of finding an optimal value. However, no significant improvement was observed for the data considered when N_{rand} was greater than 10. For this reason, we chose 10 as the value of N_{rand} for each experiment. y_{N+i} is then computed by selecting the y_{N+i}^j that makes the new nonlinear measure the closest to $V(S_N^m)$:

$$j_{\min} = \arg \min_j \left(\left| V(S_{N+i-1}^m + y_{N+i}^j) - V(S_N^m) \right| \right),$$

$$y_{N+i} = y_{N+i}^{j_{\min}} \quad (5)$$

where $(S_{N+i-1}^m + y_{N+i}^j = \{y_1, y_2, \dots, y_{N+i-1}, y_{N+i}^j\})$

The value y_{N+i}^j is chosen to make $V(S_{N+i-1}^m + y_{N+i}^j)$ as close as possible to $V(S_N^m)$.

The important point is that the reference value is always $V(S_N^m)$, which is the calculated nonlinear measure from the original time series. Therefore, the GenericPred method uses two basic rules:

R1: Always endeavour to keep the value of a nonlinear measure as steady as possible during prediction (Fig. 3).

R2: The new value must be chosen from a set of potential values generated from a probability distribution.

The prediction has to be pursued one step at a time because the predicted value in the current step is needed for determining the valid range of change for the next step. For those problems for which a binary prediction ('yes' or 'no') is required, (e.g., the epileptic seizure prediction), a threshold t is computed from the learning data.

Whenever the value y_{N+i} is greater than the threshold t , the prediction is positive. For example, yes there is an epileptic seizure at time $N+i$ if $y_{N+i} > t$; otherwise, there will be no seizure at time $N+i$.

Classical model-based prediction approaches consider a unique value for the next step, whereas in the GenericPred method, several points are considered simultaneously. Our method is also able to constantly adjust the information regarding the current time series, whereas classical predictive methods apply the model without

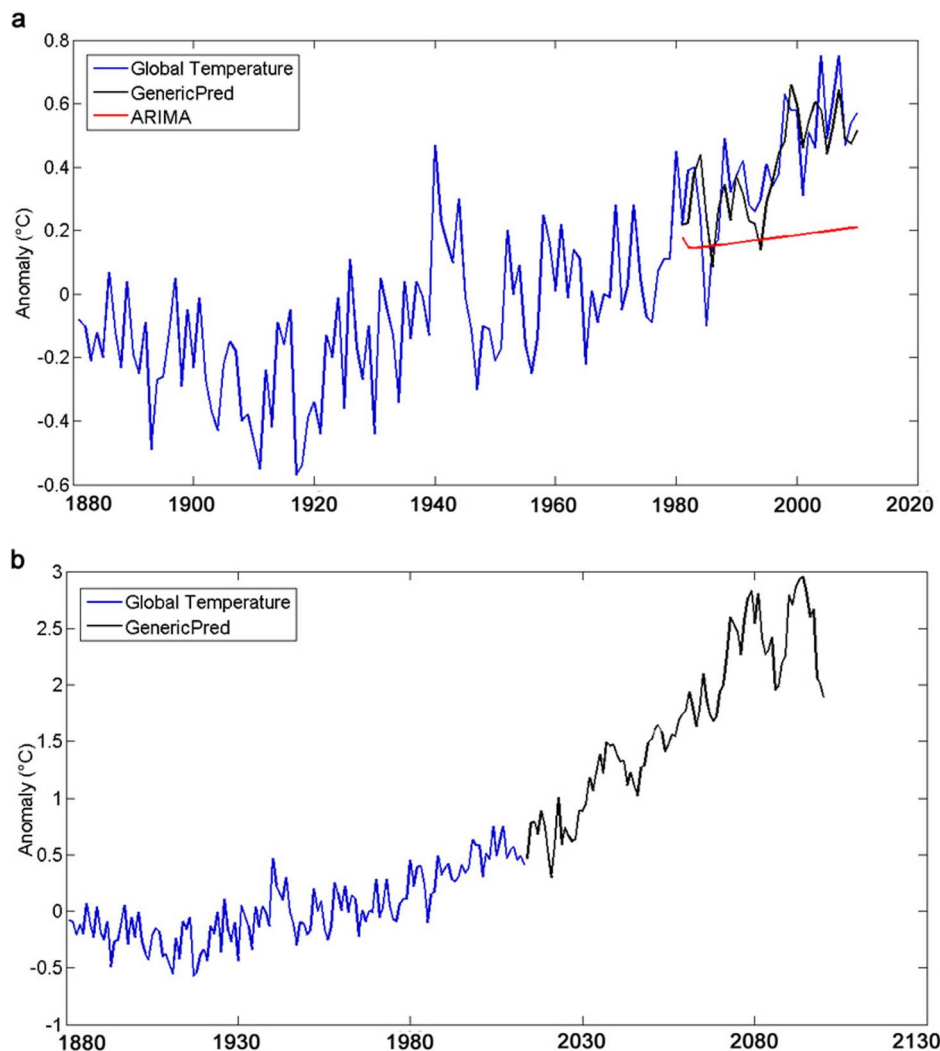


Figure 2 | Predicting the annual records of global temperature anomaly (a) for 30 years (1983–2013) and (b) until the end of the 21st century (2014–2100).

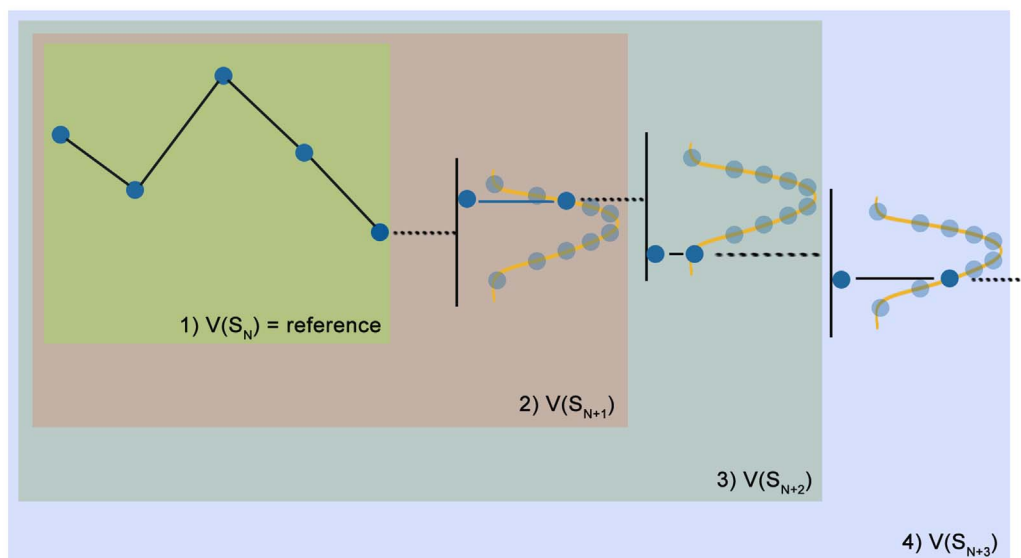


Figure 3 | Successive steps of the GenericPred method for time series prediction.

taking into account the concordance between the original time series and the predicted ones. Technically, any nonlinear measure could be used for the time series characterisation. However, here, we used the P&H method²⁵ because it has been shown that this method can efficiently discriminate between different types of nonlinear behaviour³⁹.

- Lehnertz, K. & Elger, C. E. Can epileptic seizures be predicted? evidence from nonlinear time series analyses of brain electrical activity. *Phys. Rev. Lett.* **80**, 5019–5023 (1998).
- Jeong, J. Nonlinear dynamics of EEG in Alzheimer's disease. *Drug. Dev. Res.* **56**, 57–66 (2002).
- Dauwels, J., Vialatte, F., Latchoumane, C., Jeong, J. & Cichocki, A. EEG synchrony analysis for early diagnosis of Alzheimer's disease: a study with several synchrony measures and EEG data sets. *Paper presented at Engineering in Medicine and Biology Society, Minneapolis*. US-New York: IEEE, 2224–2227 (2009 Sep).
- Newman, W. I., Gabrielov, A. M. & Turcotte, D. L. *Nonlinear Dynamics and Predictability of Geophysical Phenomena* [Newman, W. I., Gabrielov, A. M. & Turcotte, D. L. (ed.)] [15–35] (American Geophysical Union, Washington, 1994).
- Sugihara, G. *et al.* Detecting causality in complex ecosystems. *Science* **338**, 496–500 (2012).
- Stott, Peter, A. & Kettleborough, J. A. Origins and estimates of uncertainty in predictions of twenty-first century temperature rise. *Nature* **416**, 723–726 (2002).
- Clements, M. P., Franses, P. H. & Swanson, N. R. Forecasting economic and financial time-series with non-linear models. *Int. J. Forecast.* **20**, 169–183 (2004).
- Arrow, K. J. *et al.* The promise of prediction markets. *Science* **320**, 877–878 (2008).
- Zhang, G. P. & Kline, D. Quarterly time-series forecasting with neural networks. *IEEE. Trans. Neural. Netw.* **18**, 1800–1814 (2007).
- Gooijer, J. D. D. & Hyndman, R. 25 Years of time series forecasting. *Int. J. Forecast.* **22**, 442–473 (2006).
- Romanelli, L., Figliola, M. A. & Hirsch, F. A. Deterministic chaos and natural phenomena. *J. Stat. Phys.* **53**, 991–994 (1988).
- Zanin, M. Forbidden patterns in financial time series. *Chaos* **18**, 013119 (2008).
- Boettiger, C. & Hastings, A. Tipping points: From patterns to predictions. *Nature* **493**, 157–158 (2013).
- Scheffer, M. Complex systems: Foreseeing tipping points. *Nature* **467**, 411–412 (2010).
- Podobnik, B. *et al.* Network Risk and Forecasting Power in Phase-Flipping Dynamical Networks. *Phys. Rev. E* **89**, 042807 (2014).
- Drake, J. M. & Griffen, B. D. Early warning signals of extinction in deteriorating environments. *Nature* **467**, 456–459 (2010).
- Becks, L., Hilker, F. M., Malchow, H., Jurgens, K. & Arndt, H. Experimental demonstration of chaos in a microbial food web. *Nature* **435**, 1226–1229 (2005).
- Box, G. E. P., Jenkins, G. M. & Reinsel, G. C. *Time Series Analysis: Forecasting and Control* (Prentice-Hall, Englewood Cliffs, 1994).
- Montgomery, D. C., Johnson, L. A. & Gardiner, J. S. *Forecasting and Time Series Analysis* (McGraw Hill, New York, 1990).
- Gottman, J. M. *Time Series Analysis* (Cambridge University Press, Cambridge, 1981).
- Neri, F. Learning and predicting financial time series by combining natural computation and agent simulation. *EvoApplications* **6625**, 111–119 (2011).
- Zirilli, J. S. *Financial prediction using Neural Networks* (International Thompson Computer Press, London, 1997).
- Bollerslev, T. Generalized autoregressive conditional heteroskedasticity. *J. Econometrics.* **31**, 307–327 (1986).
- Filed under US Patent Application Serial Number 61/882863 (granted Sep. 26 2013).
- Golestani, A., Jahed Motlagh, M. R., Ahmadian, K., Omidvarnia, A. H. & Mozayani, N. A new criterion for distinguish stochastic and deterministic time series with the Poincaré section and fractal dimension, *Chaos* **19**, 013137 (2009).
- Shouyi, W., Chaovalitwongse, W. A. & Wong, S. Online seizure prediction using an adaptive learning approach, *IEEE Trans. Knowledge Data Eng.* **25**, 2854–2866 (2013).
- Seneviratne, S. I., Donat, M. G., Mueller, B. & Alexander, L. V. No pause in the increase of hot temperature extremes. *Nature Clim. Change.* **4**, 161–163 (2014).
- Smith, D. M. *et al.* Improved surface temperature prediction for the coming decade from a global climate model. *Science* **317**, 796–799 (2007).
- Hahn, R. & Tetlock, P. Using information markets to improve public decision making. *J. Law Pub. Pol.* **28**, 213 (2005).
- Lütkepohl, H. *New Introduction to Multiple Time Series Analysis* (Springer-Verlag, Berlin, 2005).
- Meinardi, H., Scott, R. A. & Reis, R. The treatment gap in epilepsy: the current situation and ways forward. *Epilepsia* **42**, 136–149 (2001).
- Fisher, R. S., Boas, W. V. E., Blume, W., Elger, C. & Genton, P. Epileptic seizures and epilepsy: definitions proposed by the International League Against Epilepsy (ILAE) and the International Bureau for Epilepsy (IBE). *Epilepsia* **46**, 470–472 (2005).
- Jirsa, V. K., Stacey, W. C., Quilichini, P. P., Ivanov, A. I. & Bernard, C. On the nature of seizure dynamics. *Brain*, awu133 (2014).
- Park, Y., Luo, L., Parhi, K. K. & Netoff, T. Seizure prediction with spectral power of EEG using cost-sensitive support vector machines. *Epilepsia* **52**, 1761–1770 (2011).
- Feldwisch-Drentrup, H. *et al.* Joining the benefits: Combining epileptic seizure prediction methods. *Epilepsia* **51**, 1598–1606 (2010).
- Ye, L. M., Yang, G. X., Van Ranst, E. & Tang, H. J. Time-series modeling and prediction of global monthly absolute temperature for environmental decision making. *Adv. Atmos. Sci.* **30**, 382–396 (2013).
- Falconer, K. *Fractal Geometry: Mathematical Foundations and Applications* (Wiley, New York, 2007).
- Rosenstein, M. T., Collins, J. J. & De Luca, C. J. A practical method for calculating largest Lyapunov exponents from small data sets. *Phys. D.* **65**, 117–134 (1993).
- Golestani, A. & Gras, R. Regularity analysis of an individual-based ecosystem simulation. *Chaos* **20**, 043120 (2010).

Acknowledgments

We thank R. Scott and B. MacPherson for comments on the manuscript. This work was supported by the NSERC grant ORGPIN 341854, the CRC grant 950-2- 3617 and the CFI grant 203617. All financial data have been acquired from the website of the Federal Reserve Bank of St. Louis at <http://research.stlouisfed.org>. The EEG datasets of 21 patients were acquired from the Epilepsy Center of the University Hospital of Freiburg. Global temperature anomaly data come from the Global Historical Climatology Network-Monthly (GHCN-M) dataset and the International Comprehensive Ocean-Atmosphere Data Set (ICOADS),



Author contributions

A.G. designed the study and analysed and interpreted the data. R.G. supervised the study. A.G. and R.G. wrote the paper.

Additional information

Supplementary information accompanies this paper at <http://www.nature.com/scientificreports>

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Golestani, A. & Gras, R. Can we predict the unpredictable? *Sci. Rep.* 4, 6834; DOI:10.1038/srep06834 (2014).



This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder in order to reproduce the material. To view a copy of this license, visit <http://creativecommons.org/licenses/by-nc-sa/4.0/>