A&A 386, 313–318 (2002)
DOI: 10.1051/0004-6361:20020246

Astronomy & Astrophysics

Forecasting the solar cycle with genetic algorithms

A. Orfila[1], J. L. Ballester[2], R. Oliver[2], A. Alvarez[1,⋆], and J. Tintoré[1]

[1] Instituto Mediterráneo de Estudios Avanzados (UIB–CSIC), Campus UIB, 07071 Palma de Mallorca, Spain
e-mail: vdfsaof8@uib.es, alvarez@saclantc.nato.int, dfsjts0@uib.es
[2] Departament de Física, Universitat de les Illes Balears, 07071 Palma de Mallorca, Spain
e-mail: dfsjlb0@uib.es

Received 12 November 2001 / Accepted 25 January 2002

Abstract. In the past, it has been postulated that the irregular dynamics of the solar cycle may embed a low order chaotic process (Weiss 1988, 1994; Spiegel 1994) which, if true, implies that the future behaviour of solar activity should be predictable. Here, starting from the historical record of Zürich sunspot numbers, we build a dynamical model of the solar cycle which allows us to make a long-term forecast of its behaviour. Firstly, the deterministic part of the time series has been reconstructed using the Singular Spectrum Analysis and then an evolutionary algorithm (Alvarez et al. 2001), based on Darwinian theories of natural selection and survival and ideally suited for non-linear time series, has been applied. Then, the predictive capability of the algorithm has been tested by comparing the behaviour of solar cycles 19–22 with forecasts made with the algorithm, obtaining results which show reasonable agreement with the known behaviour of those cycles. Next, the forecast of the future behaviour of solar cycle 23 has been performed and the results point out that the level of activity during this cycle will be somewhat smaller than in the two previous ones.

Key words. Sun: activity – methods: numerical

1. Introduction

From a rigorous point of view, the prediction of solar activity should rely on an accurate knowledge of the mechanism and operation of the solar dynamo. A solar activity prediction method using such an approach was developed two decades ago (Schatten et al. 1978) but it is based on only two solar cycles (i.e. about 22 years) of magnetic data, which prevents this technique from being tested with enough confidence. Nowadays, current techniques used to predict the level of solar activity can be divided in two categories: regression and precursor techniques (Hathaway et al. 1999). Regression techniques extrapolate from past levels of solar activity to the future and information of solar activity three years past the minimum is needed to characterise the next solar cycle. Precursor techniques, based on the concept of an extended solar cycle, allow one to make an estimation of the amplitude of the next solar cycle well before it starts. However, a common drawback of regression and precursor techniques is the lack of any physical or mathematical background, being completely empirical and unable to reproduce the dynamics underlying the solar cycle time series.

On the other hand, it has been postulated that the apparent randomness and complex behaviour of solar activity may be due to a low order chaotic process of non-linear but deterministic dynamics (Weiss 1988, 1994; Spiegel 1994). If true, it would be possible to exploit this determinism to obtain accurate predictive models of the solar cycle, a task that can be accomplished using recent results of a non-linear time series analysis (Casdagli 1989). In particular, Takens' theorem (Takens 1981) constitutes a valuable tool to build a predictive model of the solar cycle. The theorem establishes that, given a time series of observations $\{x(t_i)\}, i = 1, \ldots, N$, of a deterministic dynamical system, there exists a map $P(\cdot)$ satisfying

$$x(t) = P(x(t-\tau), x(t-2\tau), \ldots, x(t-m\tau)), \quad (1)$$

where m is called the embedding dimension, obtained from a state-space reconstruction of the time series (Ababarnel et al. 1993), and τ is a time lag unit.

Here we explore the presumed predictability of the solar cycle by solving the model $P(\cdot)$ in Eq. (1) for a filtered time series of sunspot data. Filtering is required because it is expected that the observed variability of a given experimental time series consists of a deterministic (predictable) and a truly random (unpredictable) part. Noisy

Send offprint requests to: R. Oliver,
e-mail: ramon.oliver@uib.es
⋆ *Present address:* North Atlantic Treaty Organization, SACLANT Undersea Research Centre, Viale San Bartolomeo 400, 19138 La Spezia, Italy.

data would lead to a worse performance of any predictor $P(\cdot)$, since the forecast system would attempt to predict the noise (i.e., to find a dynamical law for a random effect) at the expense of predicting the true underlying dynamics.

2. Data and methods

In this paper, two different time series have been used: the yearly and monthly Zürich sunspot number (1834–1999), which have been downloaded from `http://www.astro.oma.be/SIDC`. The data cover 166 years, or 1992 months, which correspond to solar cycles 8 to 22 and the rising part of cycle 23. Data prior to 1834 has not been included in the present study since they are not too reliable.

An adequate approach to remove noise without loosing a significant portion of the deterministic signal is the Singular Spectrum Analysis (Penland et al. 1991), which is equivalent to the Singular Value Decomposition (SVD) technique widely used in Astrophysics. Briefly, for the sunspot time series, a trajectory matrix X is formed whose ith row is a d-dimensional vector of the form

$$[x(t - i\Delta t), x(t - (i+1)\Delta t), \ldots, x(t - (i+d-1)\Delta t)]\,,$$

Δt being the time step between observations and d the window size of the filter. Then, the covariance matrix $X^T X$ is computed and diagonalised to obtain the orthogonal singular vectors. The corresponding eigenvalues are the average root-mean-square projection of the d-dimensional vectors that constitute the trajectory matrix. A new, noise-free time series can then be reconstructed considering only the eigenvectors with eigenvalues above the noise level. It has been shown that previous SSA processing of an experimental time series considerable improves the forecast skill of linear stochastic predictors (Elsner & Tsonis 1994).

Both sunspot time series have been filtered using the SSA technique and the window width, d, has been chosen as $d = \frac{N}{3}$, with N the time series length (Vatuard et al. 1992). The time series reconstruction has been made by considering only the most representative eigenvalues, avoiding those which only add noise to the reconstruction. In the case of the monthly Zürich sunspot number the comparison between the original signal and the reconstructed one using SSA can be seen in Fig. 1.

An evolutionary algorithm, based on Darwinian theories of natural selection and survival (Koza 1992) is used to approximate the equation, in symbolic form, that best describes the sunspot number (Alvarez et al. 2001). The algorithm considers an initial population of randomly generated equations $\{P_j(\cdot)\}$ derived from random combinations of state variables $\{x(t_i)\}$, random numbers and the four basic operators $(+, -, \times, \div)$. This population of potential solutions is next subject to an evolutionary process, by which those individuals that best fit the data are selected from the initial population. This process is carried out as follows. First, part of the time series is selected as the "training set", i.e. the set of data that will be used

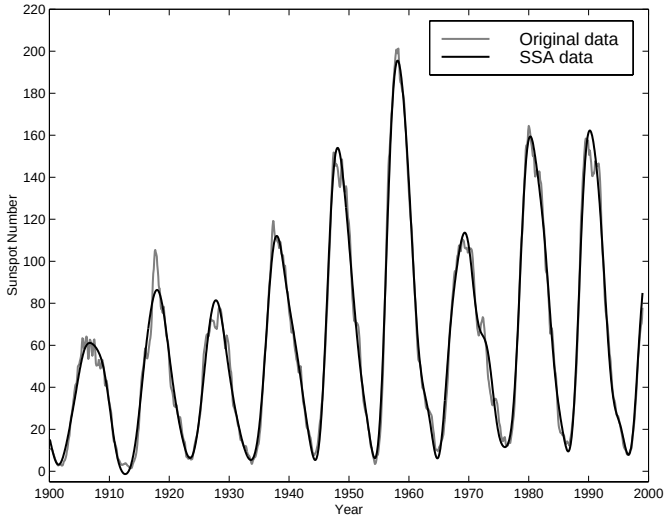

Fig. 1. Comparison between the 1900–1999 monthly Zürich sunspot number (gray line) and the SSA reconstruction (black line) using the 25 most representative eigenvalues. The 20th century covers solar cycles 14 to 22 and part of cycle 23.

in computing the best approximation to the map $P(\cdot)$ in Eq. (1). For μ-step ahead prediction, the fitness of the j-candidate equation string $P_j(\cdot)$ is computed as

$$\Delta_j^2 = \sum_{t=L+1}^{T} [x(t) - P_j(x(t-\mu\tau), x(t-(\mu+1)\tau), \ldots, x(t-(\mu+m-1)\tau))]^2,$$

where $L = (\mu + m - 1)\tau$ and T is the total length of the training set. The fitness to the data, Δ_j^2, establishes the strength of each individual in the selection process. Specifically, the strength index for the j-equation string is expressed as

$$R_j = 1 - \frac{\Delta_j^2}{\sum_{t=L+1}^{T}(x(t) - \bar{x})^2},$$

where $\bar{x}$ is the mean value of the training data. R_j can be interpreted as the percentage of the training set's total variance explained by the j-equation string and so, for a given individual, the higher this quantity, the better the data are represented by the corresponding equation.

The strongest individuals, i.e. the equation strings which best fit the training set, are then selected to exchange parts of the character strings between them (which mimics the reproduction and crossover processes found in the natural world). On the other hand, the individuals less fitted to the data are discarded. Finally, a small percentage of elements in the equation strings are mutated at random. As a result of this process a new set of equations with better fitness properties is obtained. The evolutionary steps are repeated with the new generation and are finished after a number of generations determined by the user. Consequently, the equation that best fits the data is obtained. Finally, a validation can be carried out using the data not included in the training set to test the

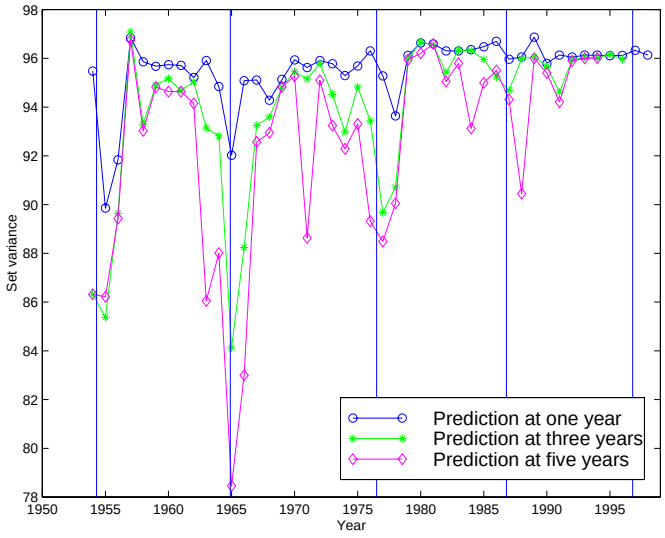

Fig. 2. Forecast skill of the algorithm for predictions at 1, 3 and 5 years. The abscissae mark the end year of the subsets (all of which start in 1834) used in testing the performance of the genetic algorithm. For each subset, the equation that best fits its dynamics has been used to make predictions at 1, 3 and 5 years and the results are compared with the original sunspot values by computing the total set variance between the two series. This variance is plotted here against the end year of each data subset. Vertical lines mark the occurrence of solar activity minima.

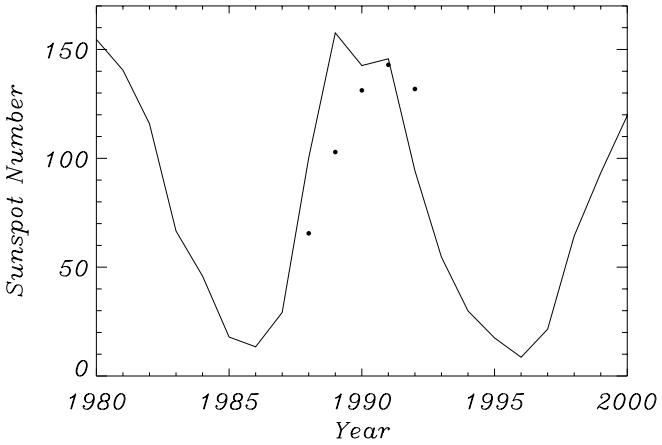

Fig. 3. Yearly Zürich sunspot number (solid line) and prediction starting in 1988 (filled circles). The predicted values correspond to years 1988–1992 and display a maximum yearly sunspot number of 143 ± 29 in 1991.

goodness of the candidate equation for reproducing the deterministic signal.

This algorithm is particularly useful when the dynamical model underlying the time series is non-linear, such as is expected in the case of the solar cycle. In our computations, the last $\frac{N}{3}$ data have been used as the training set and each generation consists of a population of 120 equations. After 10 000 generations have been performed, the equation that yields the smallest residual against the training set is kept for forecasting purposes.

The present algorithm is a stochastic optimization method, so if two simulations of the same length using the same parameters are performed, different best-fit equations are obtained. Nevertheless, if the noise in the original time series has been adequately removed before running the genetic algorithm, similar equations, although not necessarily the same, are always derived.

3. Results and conclusions

First of all, an estimation of the forecast skill of the evolutionary algorithm for the yearly data set has been performed. To this end, the algorithm has been applied to different subsets of the reconstructed data series (always retaining the same number of eigenvalues in the SSA) and the equation that best describes the dynamics of each of these subsets has been used to make predictions from one to five years. These predictions can then be used to assess the goodness of the method. This procedure has been performed for 45 different subsets, all of them starting in 1834 and finishing in 1954 (first subset), 1955 (second subset), ... and 1998 (45th subset). Note that different data subsets possess their own predictive equation. The reason to consider these subsets is to check the performance of the algorithm in predicting solar cycles of quite different strengths: cycle 19 (which begins in 1954) displays the largest recorded activity, cycle 20 is of moderate intensity and cycles 21 and 22 are more active than the average. The comparison between prediction and original unfiltered data is made by computing the total set variance between the two series (see Fig. 2). As expected, the results show that the prediction worsens as the number of years being forecasted is increased. It is also noticeable that the evolutionary algorithm yields better results when the solar cycle is well underway than at the beginning or the end of the cycle. However, even for the five year prediction the minimum variance accounted for by the predicted signal is around 80%, which indicates that the Darwinian algorithm is suitable for making reliable predictions of the yearly sunspot number a few years in advance. From Fig. 2, we estimate a ∼20% uncertainty when making predictions at five years of future yearly sunspot number values. The above is just an estimation of the error based on the 5-year prediction in Fig. 2 whose performance is the worst. This is a first approach towards the error assessment and, probably, more sophisticated methods, such as Montecarlo simulations of multiple runs, should be implemented.

Once the forecast skill has been established, we next give an example of the performance of the evolutionary algorithm. We consider the time interval 1834–1987, which includes the beginning of solar cycle 22, for which the best equation that best maps its underlying dynamics is

$$x(t) = x(t-1) + [x(t-9) - 1.069590 * x(t-3)]/ \{3.927562 + (3.420568 + x(t-31)) * \frac{-2.326474}{x(t-32) * x(t-3)}\},$$

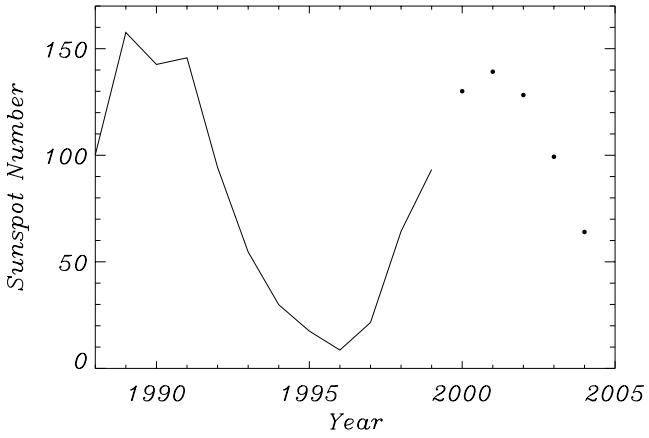

Fig. 4. Yearly Zürich sunspot number (solid line) and prediction starting in 2000 (filled circles). The predicted values correspond to years 2000–2004 and display a maximum yearly sunspot number of 143 ± 29 in 2001.

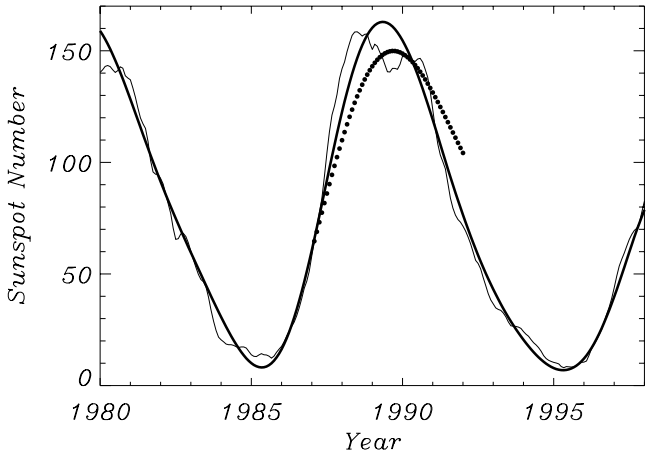

Fig. 5. Monthly sunspot number (thin line), SSA smoothed monthly sunspot number (thick line) and monthly prediction for sixty months after January 1987 (filled circles).

with t the time in years. Using this equation, and starting in 1988, we have predicted the behaviour of solar cycle 22 for the following five years, finding a solar activity maximum in 1991 (with one year uncertainty) with an amplitude of 143 ± 29 (Fig. 3). This value is quite similar to the actual one in that year (145.7) and is just 9% smaller than the maximum Zürich sunspot number of this cycle (157.6), that took place in 1989. This example illustrates the strengths and weaknesses of the genetic algorithm: predicted values from the yearly data series are reasonable but the time of occurrence of maximum activity is wrong by one or two years. Moreover, one should bear in mind that in this particular case the prediction is started at the beginning of an activity cycle, the time for which the performance of the method is found to be worst (cf. Fig. 2).

Next, we have used the evolutionary algorithm to compute the symbolic equation for the whole yearly time series (1834–1999). Now, the symbolic expression that best maps this set is

$$x(t) = x(t-1) + \frac{-3.460}{x(t-33)} + \frac{1.390 + x(t-9) - x(t-3)}{3.4141},$$

with t the time in years. Using this equation we have forecasted the behaviour of solar cycle 23 (1996–) for five years starting from 2000, finding that it predicts, with one year uncertainty, a maximum of solar activity during the year 2001 with an amplitude of 143 ± 29 (Fig. 4).

To reduce the uncertainty about the date on which the maximum or minimum of solar cycles will take place we have resorted to the monthly data set. Our first aim has been to assess the performance of the Darwinian method for making predictions based on monthly sunspot data. To this end we have made predictions using four data sets covering the period from 1834 to half the rising part of solar cycles 19, 20, 21 and 22, respectively. Predictions at 12, 24, 36, 48 and 60 months have next been constructed for each of these four sets and a comparison with the SSA time series has been made using the total set variance (Table 1). The results point out that, even for predictions at five years, the variance accounted for by the predicted signal is above 82% of the variance of the SSA time series. Thus, the genetic algorithm seems quite powerful at forecasting the non-linear dynamics underlying solar activity, although obviously it fails to predict the noisy component present in the time series that shows up as variations with short time-scales (see Fig. 1). From Table 1 we estimate an error around 15% in predicting, five years in advance, the deterministic component of the monthly sunspot number.

We again show a comparison between the predicted behaviour of cycle 22 and the true sunspot number data. First of all, we have obtained the equation that best fits the monthly sunspot number data from January 1834 to December 1986,

$$\begin{aligned} x(t) \;=\; & x(t-1) - \{x(t-20) + 7.769775 \\ & - [7.572632 + 9.569702 * (x(t-2) - x(t-10))] \\ & - x(t-40)\ \}\ /(9.583740 * 8.582764) \end{aligned}$$

with t in months. Now, we have made a forecast of the behaviour of solar cycle 22 for sixty months starting in January 1987, finding that the maximum of this cycle is predicted to occur in 1989.7 while its true date of occurrence was 1989.3. Moreover, the predicted maximum sunspot number value (149.9) is only 8% smaller than the actual one (162.9). The comparison between the predicted and known behaviour of the rising phase and maximum of solar cycle 22 can be seen in Fig. 5, which shows a reasonable agreement between reality and forecasting.

In order to make a further check of the predictive capability of the method, we have computed the symbolic equation that best fits the monthly data set from January 1834 to December 1989,

$$\begin{aligned} x(t) \;=\; & x(t-1) - \left[\frac{x(t-24)}{35.5741} \right. \\ & \left. + \frac{-1.857910 * x(t-113) - 0.6655 * x(t-1)}{x(t-152) + x(t-91)} \right], \end{aligned}$$

Table 1. Forecast skill of the Darwinian algorithm for predictions up to 60 months. Total set variance between the prediction at 12, 24, 36, 48 and 60 months and the SSA reconstructed monthly time series. The results correspond to four data sets constructed with the monthly sunspot data from 1834 to a point in the rising phase of cycles 19, 20, 21 and 22, respectively.

Cycle Number	12 months	24 months	36 months	48 months	60 months
19	98.7	92.5	87.1	85.9	86.2
20	99.7	98.4	95.8	93.6	82.6
21	99.6	97.1	95.5	93.8	92.2
22	99.9	99.8	99.7	99.6	99.0

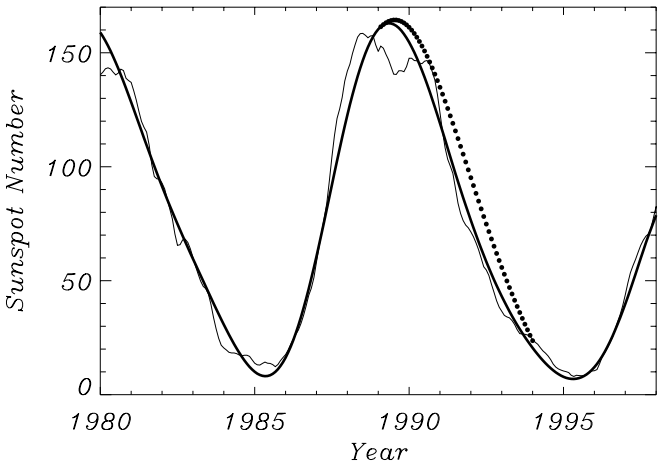

Fig. 6. Monthly sunspot number (thin line), SSA smoothed monthly sunspot number (thick line) and monthly prediction for sixty months after January 1990 (filled circles).

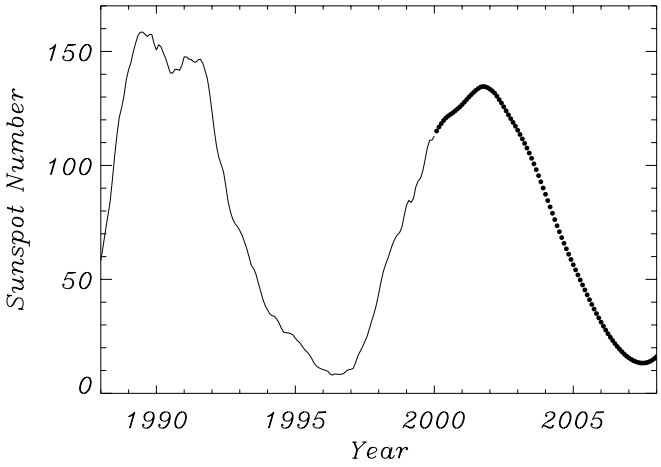

Fig. 7. Monthly sunspot number (solid line) and monthly prediction for 96 months after January 2000 (filled circles). These results suggest that solar activity in cycle 23 will peak around July 2001 with a maximum sunspot number of 135 ± 20 and will go through a minimum around June 2007.

with t in months, and have predicted the behaviour of solar cycle 22 for sixty months starting in January 1990, i.e. the declining phase of the cycle. The result is shown in Fig. 6 and displays a good agreement with the already known behaviour of the declining phase of this cycle, which further strengthens our confidence in the genetic algorithm.

Finally, we have moved again to solar cycle 23 and have computed the symbolic equation that best fits the whole monthly data set (January 1834 to December 1999), which is given by

$$x(t) = x(t-1) - 0.0264 * \left[\frac{x(t-27) * x(t-29)}{5.002 + x(t-37)} - x(t-117) \right],$$

where t is the time in months. Using this equation we have predicted the behaviour of the present solar cycle, obtaining a maximum monthly sunspot number of 135 ± 20 in July 2001 (Fig. 7). Hence, the yearly and monthly predictions agree quite well and suggest that the present cycle will be weaker than previous cycles 21 and 22. This forecast refers to the deterministic part of solar activity but not to its inherent noise and so this equation does not provide a detailed description of future Sunspot Numbers, only of their main behaviour. Furthermore, and despite that the last equation is not suitable for predictions more than five years in advance, we have used it to forecast the behaviour of solar cycle 23 during all its declining phase. In spite of the inadequacy of the starting equation, it can be appreciated the correct form of the declining phase which suggests that the next minimum of solar activity will occur around June 2007.

One last comment about the various best-fit equations must be made. The two predictive equations for the yearly sunspot number look rather different, but nevertheless they incorporate similar time-delay parameters, namely 3, 9 and ∼32 years. Similar delay times also show up in the model equation for the 1834–1999 monthly sunspot number (namely ∼30 months and ∼120 months) and in other equations used but not shown in this paper. This is a potentially interesting result that should be explored further in the future.

In summary, in this paper, an evolutionary algorithm that is particularly suitable when the dynamical model underlying a time series is non-linear, such as seems to be the case with the solar activity cycle, is presented. The method has been validated by comparing its predictions with existing data and the future behaviour of solar cycle 23 has been forecast. The obtained results are reasonable, suggesting that this approach can be useful to forecast solar activity without the need to resort to other techniques based on empirical rules. It is also worth pointing out that,

even restricting ourselves to predictions at five years, it is possible to start forecasting at the beginning of a solar cycle and still be able to reproduce most of the variance of the true signal, i.e. to forecast the amplitude of the future maximum, with a 15–20% error.

Acknowledgements. J.L.B. and R.O. acknowledge the financial support received from MCyT under grant BFM2000-1329.

References

Ababarnel, H. D. I., Brown, R., Sidorowich, J., & Tsimring, L. S. 1993, Rev. Mod. Phys., 65, 1331
Allen, R. M., & Smith, L. A. 1996, J. Climate, 9, 3373
Alvarez, A., Orfila, A., & Tintoré, J. 2001, Comp. Phys. Comm., 136, 334
Casdagli, M. 1989, Phys. D, 35, 335
Elsner, J. B., & Tsonis, A. A. 1994, Singular Spectrum Analysis (Plenum Press)
Hathaway, D. H., Wilson, R. M., & Reichmann, E. J. 1999, J. Geophys. Res., 104, A10, 22375
Koza, J. R. 1992, Genetic Programming (The MIT Press)
Penland, C., Ghil, M., & Weickmann, K. M. 1991, J. Geophys. Res., 96, 22659
Schatten, K. H., Scherrer, P. H., Svalgaard, L., & Wilcox, J. M. 1978, Geophys. Res. Lett., 5, 411
Spiegel, E. A. 1994, in Lectures on Solar and Planetary Dynamos, ed. M. R. E. Proctor, & A. D. Gilbert (Cambridge University Press), 245
Szpiro, G. G. 1997, Phys. Rev. E, 55, 2557
Takens, F. 1981, in Dynamical systems and Turbulence, ed. D. A. Rand, & L. S. Young, Lecture Notes in Math., vol. 898 (Springer–Verlag), 365
Vatuard, R., Yiou, P., & Ghil, M. 1992, Phys. D, 58, 95
Weiss, N. O. 1988, in Secular Solar and Geomagnetic variations in the last 10 000 years, ed. F. R. Stephenson, & A. W. Wolfendale (Kluwer Academic Publishers), 69
Weiss, N. O. 1994, in Lectures on Solar and Planetary Dynamos, ed. M. R. E. Proctor, & A. D. Gilbert (Cambridge University Press), 59