Statistics and Its Interface Volume 10 (2017) 145–150

Assessment of SSA predictions of Earth temperature records

Vladimir Kornikov

In this paper, the forecasts of Earth temperature records made by A. Pepelyshev and A. Zhigljavsky in 2009 are compared with the data actually observed during 2010–2014. It is demonstrated that the forecasts made in 2009 are quite accurate. In the second part, the SSA-based change-point detection algorithm proposed by Moskvina and Zhigljavsky in 2003 is applied to the same temperature records data. The results show that the data does not have essential structural breaks except perhaps a small rise of general level of temperatures at around 1998.

AMS 2000 subject classifications: Primary 62M20.
Keywords and phrases: Singular spectrum analysis, Stability of forecasts, Long-horizon forecasting, Earth temperature, Change-point detection.

1. INTRODUCTION

An important paper [2] by A. Pepelyshev and A. Zhigljavsky published in 2009 in the Statistics and Its Interface contains an interesting discussion concerning assessing stability of time series forecasts with no model available and many forecasts of the Earth temperature data. The data were taken from http://vortex.nsstc.uah.edu/, the web-site of the National Space Science and Technology Center, USA, NASA. This web-site contains the so-called monthly temperature anomalies since December 1978. In our study, we use the data from the same website, see the file http://vortex.nsstc.uah.edu/data/msu/t2lt/uahncdc_lt_5.6.txt. Therefore, the data plotted for 1978-2009 coincides with the data plotted in [2].

The forecasts in the paper [2] are made at different points of time using using the retrospective principle. The latest time point used in [2] is December 2009. One of purposes of this paper is to compare forecasts made in [2] from December 2009 onwards with the actual temperature records for 2010-2014. The method used in [2] is the so-called Singular Spectrum Analysis (SSA). Description of the SSA can be found in many publications; we refer here to the fundamental books [3] and [4]. In Section 2 we provide a very brief introduction to SSA and in Section 3 we present the comparison of forecasts.

Another purpose of this paper is to apply the SSA change-point detection algorithm for discovering structural breaks in temperature records. This algorithm proposed by Moskvina and Zhigljavsky in 2003 has the ability to find changes in mean, variance of noise, the amplitude, frequency and phase of periodic components. The description of the algorithm and it's application is given in Section 4.

2. THE METHOD

Singular spectrum analysis (SSA) is a technique of time series analysis and forecasting combining elements of classical time series analysis, multivariate statistics, multivariate geometry, dynamical systems and signal processing. SSA aims at decomposing the original series into a sum of a small number of interpretable components such as a slowly varying trend, oscillatory components and a 'structureless' noise. It is based on the singular value decomposition (SVD) of a specific matrix constructed upon the time series.

Neither a parametric model nor stationarity-type conditions have to be assumed for the time series. This makes SSA a model-free technique and hence enables SSA to have a very wide range of applicability.

SSA is a well-known technique for analyzing and forecasting time series in climatology in general and the analysis of temperature records in particular, see e.g. [1, 4, 6, 7]. The version of SSA used in [2] is the so-called Basic SSA. Below we give a short description of it following [8]. For more details, see [3, Chapter 1] and [4].

The Basic SSA. Let $x_1, \ldots, x_N$ be a time series of length N. Given a window length L $(1<L<N)$, we construct the L-lagged vectors $X_i = (x_i, \ldots, x_{i+L-1})^T,\ i = 1, 2, \ldots, K = N-L+1$, and compose these vectors into the matrix $\mathbf{X} = (x_{i+j-1})_{i,j=1}^{L,K} = [X_1 : \ldots : X_K]$. This matrix has size $L \times K$ and is often called 'trajectory matrix'. It is a Hankel matrix, which means that all the elements along the diagonal $i + j =$const are equal.

The columns X_j of $\mathbf{X}$ can be considered as vectors in the L-dimensional space $\mathbb{R}^L$. The singular value decomposition of the matrix $\mathbf{X}\mathbf{X}^T$ yields a collection of L eigenvalues and eigenvectors. A particular combination of a certain number r of these eigenvectors determines an r-dimensional subspace $\mathbb{L}$ in $\mathbb{R}^L$, $r < L$. The L-dimensional data $\{X_1, \ldots, X_K\}$ is then projected onto the subspace $\mathbb{L}$ and the subsequent averaging over the diagonals yields some Hankel matrix $\tilde{\mathbf{X}}$ which is considered as an approximation to $\mathbf{X}$. The time series $\tilde{x}_1, \ldots, \tilde{x}_N$, which is in the one-to-one correspondence with the matrix $\tilde{\mathbf{X}}$, provides an approximation to the original series $x_1, \ldots, x_N$.

Forecasting. The r-dimensional subspace $\mathbb{L}$ constructed by the Basic SSA yields a linear recurrent formula which may be used for forecasting. As an alternative to this method of forecasting we may use the so-called 'vector forecasting'. The main idea of the vector forecasting algorithm is in the consecutive construction of the vectors $X_i = (x_i, \ldots, x_{i+L-1})^T$, for $i = K+1, K+2, \ldots$ so that they lie as close as possible to the subspace $\mathbb{L}$ created by the chosen r eigenvectors.

Choice of parameters in the Basic SSA. There are two parameters to choose in the Basic SSA which we use: the window length L and the number r of largest eigenvectors chosen for the approximation and forecasting. A rational or even optimal choice of these parameters should depend on the task we are using SSA for. A detailed discussion on this topic can be found in [4] and [3, Section 1.6]. There are versions of the Basic SSA where given the window length L, the group of r indices determining the subspace $\mathbb{L}$ (for extraction of either trend or periodic components) is chosen automatically; see e.g. [4]. In the present paper we compare the forecasts from [2] for their main parameter choices; in most cases, this choice was $L = 50$ and $r \in \{5, 7\}$.

3. ASSESSMENT OF ACCURACY OF PREDICTIONS

In our study, we have used the same method as in [2] and exactly the same software, the programm Caterpillar-SSA downloaded from the website http://www.gistatgroup.com/cat/programs.html.

In Figure 1 we can see that during the period from Jan 2010 to Sep 2014 the global Earth temperatures roughly followed the most typical forecast given in [2]. Figures 2, 3, 4, 6, 5, 7 and 8 demonstrate the same for the temperature records for Northern hemisphere, Southern hemisphere, Tropics, Earth land, Earth ocean, North Pole and South Pole, respectively.

We can make the following conclusions from observing the patterns shown in Figures 1–8:

(a) global Earth temperatures did neither significantly increase nor significantly decrease anywhere;
(b) these temperatures continued to be volatile;
(c) the key SSA forecasts for the period 2010-2015 published in [2] have captured the main pattern of the temperature records for this period very well.

In Figures 1–8 we have compared the actual data against a selected SSA forecast, namely, the forecast with parameters $L = 50$ and $r = 7$. We can see a remarkable overlap of the actual temperatures and forecasts for the period 2010-2015. Unlike Figure 1, Figure 2 shows that the forecast made in Dec 2009 indicates a very small increase in the level of temperatures in the North pole but the actual temperature does not grow. Note also that in Figure 3 the forecast for

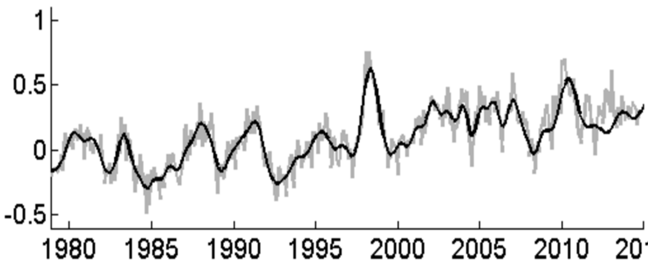

Figure 1. Global Earth temperature. Gray: the original time series for the period from Dec 1978 to Sept 2014. Black: the SSA approximation until Dec 2009 and the forecast from Jan 2010 onwards. SSA parameters: $L = 50$ and $r = 7$.

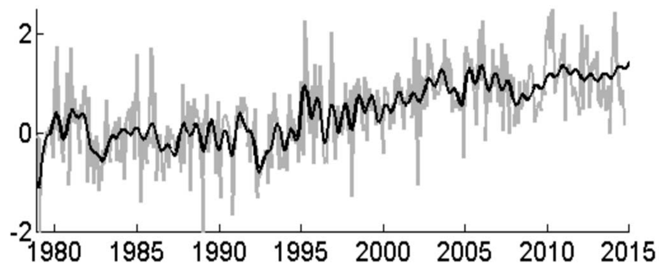

Figure 2. North Pole temperature. Gray: the original time series for the period from Dec 1978 to Sept 2014. Black: the SSA approximation until Dec 2009 and the forecast from Jan 2010 onwards. SSA parameters: $L = 50$ and $r = 7$.

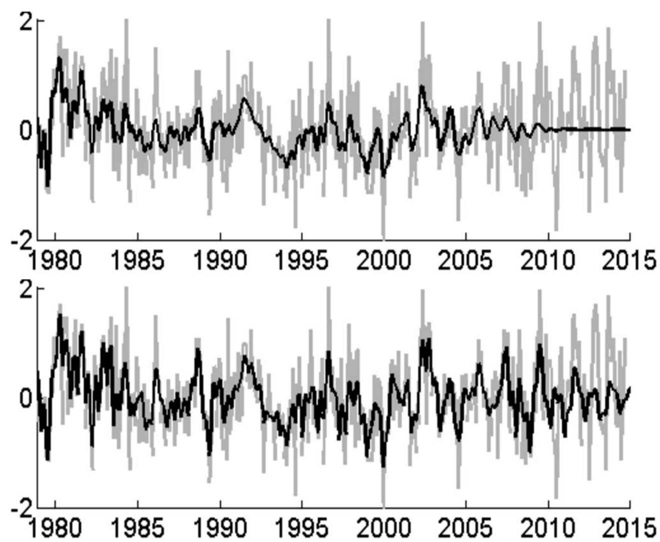

Figure 3. South Pole temperature. Gray: the original time series for the period from Dec 1978 to Sept 2014. Black: the SSA approximation until Dec 2009 and the forecast from Jan 2010 onwards. SSA parameters: top: $L = 50$ and $r = 7$; bottom: $L = 24$ and $r = 7$.

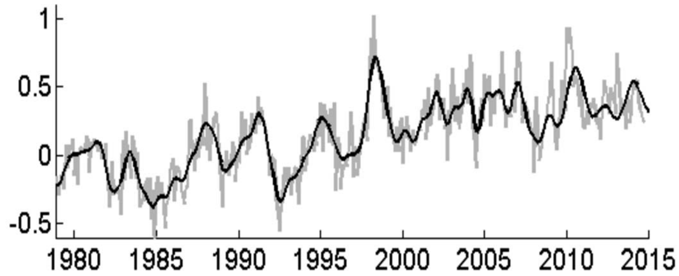

Figure 4. Northern hemisphere temperature. Gray: the original time series for the period from Dec 1978 to Sept 2014. Black: the SSA approximation until Dec 2009 and the forecast from Jan 2010 onwards. SSA parameters: $L = 50$ and $r = 7$.

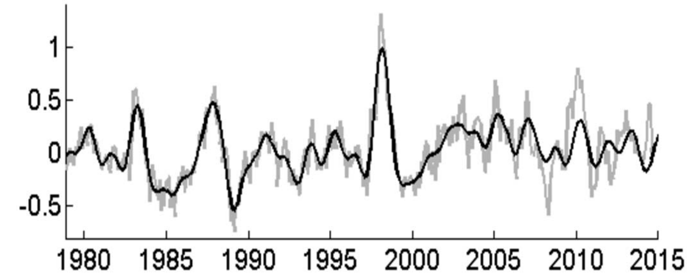

Figure 6. Tropics temperature. Gray: the original time series for the period from Dec 1978 to Sept 2014. Black: the SSA approximation until Dec 2009 and the forecast from Jan 2010 onwards. SSA parameters: $L = 50$ and $r = 7$.

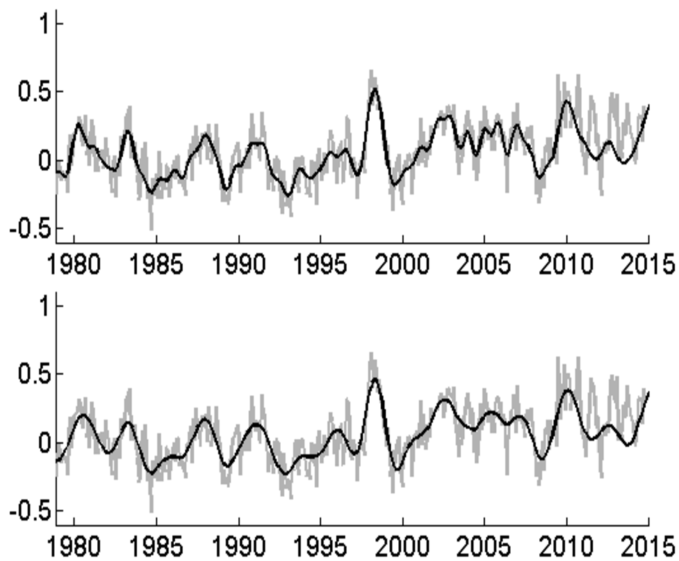

Figure 5. Southern hemisphere temperature. Gray: the original time series for the period from Dec 1978 to Sept 2014. Black: the SSA approximation until Dec 2009 and the forecast from Jan 2010 onwards. SSA parameters: top: $L = 50$ and $r = 7$; bottom: $L = 50$ and $r = 5$.

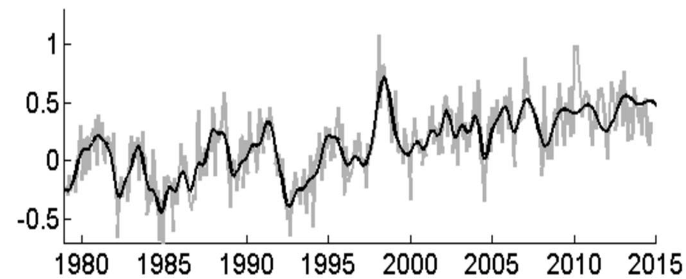

Figure 7. Earth land temperature. Gray: the original time series for the period from Dec 1978 to Sept 2014. Black: the SSA approximation until Dec 2009 and the forecast from Jan 2010 onwards. SSA parameters: $L = 50$ and $r = 7$.

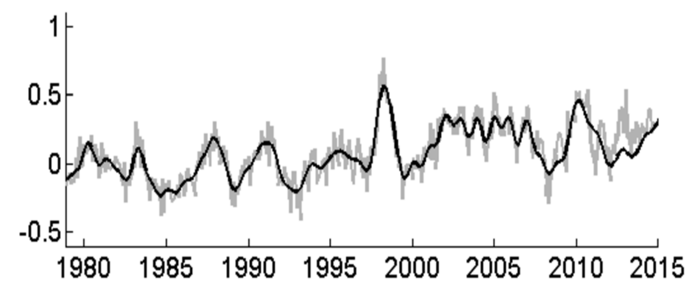

Figure 8. Earth ocean temperature. Gray: the original time series for the period from Dec 1978 to Sept 2014. Black: the SSA approximation until Dec 2009 and the forecast from Jan 2010 onwards. SSA parameters: $L = 50$ and $r = 7$.

$L = 50$ and $r = 5$ is unable to catch the volatility of the actual time series, but this forecast has shown the main trend (more precisely, the absence of any trend) very accurately. However, in Figure 3 the forecast for $L = 24$ and $r = 5$ is able to catch oscillation with right frequency but wrong amplitude. Figures 4–8 are similar to previous ones and self-explanatory.

4. SSA DETECTION OF STRUCTURAL BREAKS

The SSA algorithm of change-point detection was developed by Moskvina and Zhigljavsky in [5] and its idea is the consideration of a SSA-dissimilarity measure $D(x_{u:v}, x_{s:t})$ between the past time series segment $x_{u:v} = (x_u, x_{u+1}, \dots, x_v)$ and the most recently observed time series segment $x_{s:t} = (x_s, x_{s+1}, \dots, x_t)$ for a time series

$$\dots, \underbrace{x_u, x_{u+1}, \dots, x_v}_{\text{past}}, \dots, \underbrace{x_s, x_{s+1} \dots, x_t}_{\text{most recent}}, \dots$$

observed in real time, where $u < v \le s < t$ and t corresponds to the present time. Here we consider online monitoring for structural breaks and we refer to [3, Ch. 3] for the SSA approach to offline change-point detection.

The formal description of SSA change-point detection algorithm is as follows. For a time series $x_1, x_2, \ldots$, choose a value of the parameter L and construct vectors $X_i = (x_{i-L+1}, \ldots, x_i)^T$. Then consider the trajectory matrix $\mathbf{X} = (X_{n-m}, \ldots, X_n)$ corresponding to past time series segment and compute the leading eigenvectors $U_1, U_2, \ldots, U_r$ of $\mathbf{X}$. Note that the matrix $U = (U_1, \ldots, U_r)$ describes a 'structure' of the past time series segment $x_{u:v}$ with $u = n - m - L + 1$ and $v = n$.

For the same value of the parameter L, compute vectors $X_i = (x_{i-L+1}, \ldots, x_i)^T$, $i \in \{n+p, \ldots, n+q-1\}$ corresponding the most recently observed time series segment $x_{s:t}$ with $s = n + p - L + 1$ and $t = n + q - 1$.

In [5] the SSA-dissimilarity measure is defined by the statistic

$$D_{n,L,p,q} = \frac{1}{L(q-p)} \sum_{i=n+p}^{n+q-1} [X_i^T X_i - X_i^T U U^T X_i]$$

which can be viewed as a distance between the recent vectors $X_{n+p}, \ldots, X_{n+q-1}$ and the subspace $\mathcal{L}(U_1, U_2, \ldots, U_r)$ for the past time series segment, where r is the dimension of the signal subspace. The statistic $D_{n,L,p,q}$ means the presence of a change-point in a time series if D_n goes above some threshold. Note that the typical choice of parameters is $p = L$, $q = L + 1$ that leads to the statistic

$$(1) \qquad D_n = \frac{1}{L}[X_n^T X_n - X_n^T U U^T X_n].$$

For convenience of practical application, the SSA algorithm of change-point detection can be written in the form resembling the popular CUSUM procedure of detection of jumps in the mean. First, we compute the normalized statistic

$$d_n = D_{n,L,p,q} / D_{n,L,L-N,0}$$

and then we define the process W_1, W_2, W_3,... which is calculated sequentially by

$$W_n = \max\Big\{W_{n-1} + d_n - d_{n-1} - 1/(3LQ),\ 0\Big\},$$

$n = 2, 3, \ldots$, and starting with $W_1 = 0$, where $Q = q - p$. If the SSA detection statistic W_n becomes larger than a threshold, then this indicates the presence of a change in the structure of time series. Under the assumption of normality, the value of the threshold was found in [5] to be approximately equal to

$$h = \frac{2t_\alpha}{LQ}\sqrt{Q(3LQ - Q^2 + 1)/3}$$

where t_α is the $(1-\alpha)$-quantile of the standard normal distribution.

To demonstrate the performance of the SSA algorithm of change-point detection, we consider an artificial time series

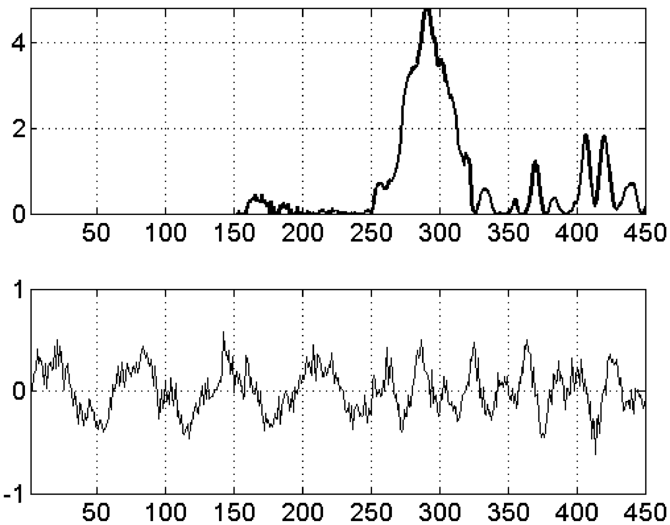

Figure 9. The SSA detection statistic W_{n+L} (top) for the simulated time series (bottom).

$x_1, \ldots, x_{450}$ defined by

$$x_k = \begin{cases} 0.3 \sin \frac{2\pi k}{65} + 0.1 \sin \frac{\pi k}{10} + 0.1\epsilon_k & k = 1, \ldots, 250 \\ 0.2 \sin \frac{2\pi k}{35} + 0.2 \sin \frac{\pi k}{10} + 0.1\epsilon_k & k = 251, \ldots, 450 \end{cases}$$

where ϵ_k are i.i.d. random variables with normal distribution $N(0, 1)$. In Figure 9 (bottom) we show the simulated time series. We run the SSA algorithm of change-point detection with parameters $L = 50$, $r = 5$, $Q = 1$ splitting the time series $x_1, \ldots, x_{n+L}$ into 2 pieces: the past time series segment $x_{1:n}$ and the most recently observed time series segment $x_{n+1:n+L}$. Note that the corresponding threshold is $h = 0.73$ for selected values of parameters. We depict the SSA detection statistic in Figure 9 (top) and we can see that the statistic W_n goes well above the threshold announcing the presence of the structural break in the time series.

SSA change-point detection for temperature records

In this section, we apply the SSA-based algorithm of change-point detection with the following parameters. We take the window length $L = 50$, the number of the leading eigenvectors $r = 5$, $Q = 1$ and divide the time series $x_1, x_2, \ldots, x_t$ for each t into 2 parts: (i) the past time series segment $x_{1:n}$ and (ii) the most recently observed time series segment $x_{n+1:n+L}$ with $n + L = t$. We also performed the SSA change-point detection with $r = 7$ for temperature records and obtained results which are similar to the case $r = 5$ and not given here for sake of brevity. Under the assumption of normality, the threshold is given by $h = 0.73$. Since the assumption of normality is moderately violated, we should consider a larger value for the threshold, see [5].

In Figures 10–17 we depict the change-point detection statistic W_{n+L} which is the CUSUM modification of the

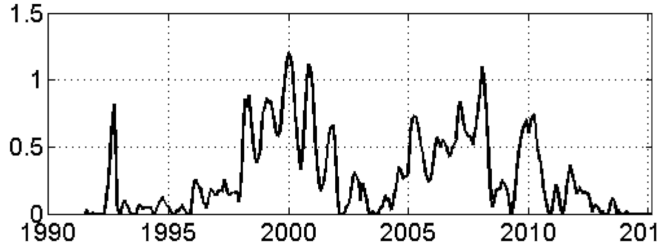

Figure 10. The SSA detection statistic W_{n+L} for global Earth temperatures.

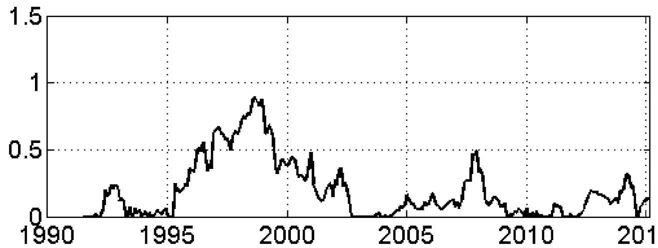

Figure 11. The SSA detection statistic W_{n+L} for North Pole temperatures.

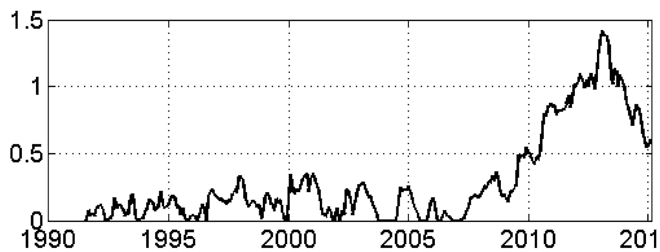

Figure 12. The SSA detection statistic W_{n+L} for South Pole temperatures.

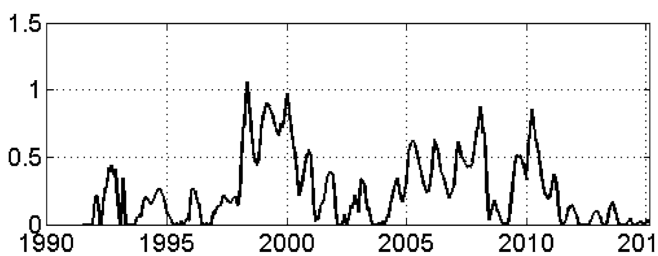

Figure 13. The SSA detection statistic W_{n+L} for Northern hemisphere temperatures.

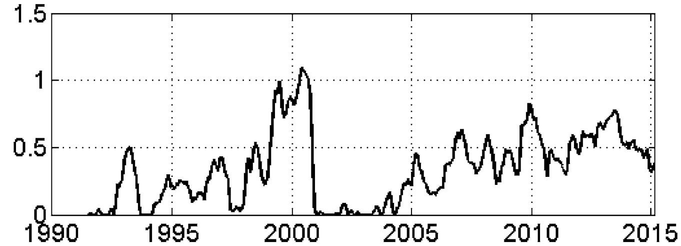

Figure 14. The SSA detection statistic W_{n+L} for Southern hemisphere temperatures.

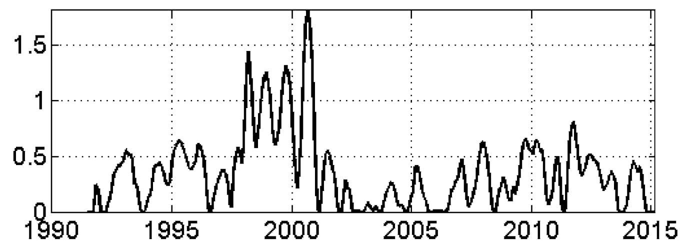

Figure 15. The SSA detection statistic W_{n+L} for Tropics temperatures.

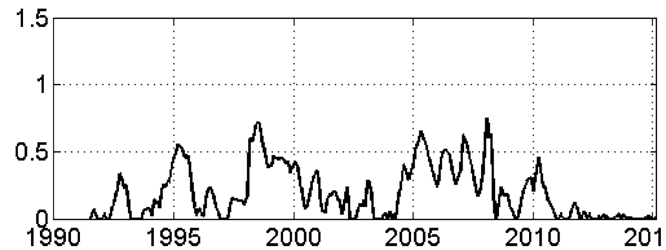

Figure 16. The SSA detection statistic W_{n+L} for Earth land temperatures.

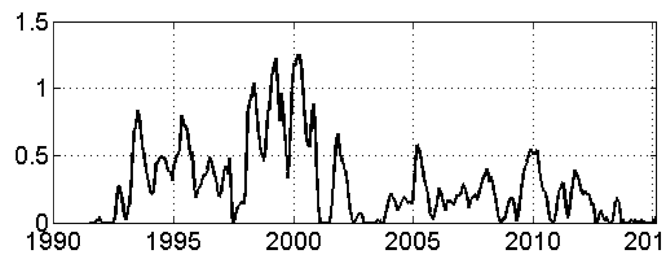

Figure 17. The SSA detection statistic W_{n+L} for Earth ocean temperatures.

SSA-dissimilarity measure between the segment $x_1, \dots, x_n$ and the segment $x_{n+1}, \dots, x_{n+L}$ for $n = 150, \dots, 436$, where 436 is the length of original time series.

In all these figures we can see that the statistic W_n does not go much above the threshold which indicates that the structural change, which happened around 1998, is not very significant. We should however note that in Figure 15 the large values of W_n occur for Tropics temperatures for the period 1997–2000. Indeed, as shown in Figure 6, Tropics temperatures for this period have rather unusual behavior.

Also, in Figure 12 the large values of W_n occur for South Pole temperatures for the period 2013–2014. As shown in Figure 3, South Pole temperatures for this period have oscillation with nontypical frequency.

Overall, we can conclude that the temperature records do not have significant and statistically justifiable structural breaks except perhaps the change, which happened around 1998.

5. CONCLUSION

In this paper, we have compared the forecasts of Earth temperature records made by A. Pepelyshev and A. Zhigljavsky in 2009 and published in [2] with the data actually observed during 2010–2014. We have demonstrated that the forecasts made by A. Pepelyshev and A. Zhigljavsky were quite accurate and have captured the main patterns of the temperature series very well. We have also shown that mostly the temperatures have neither increased or decreased but continued to be quite volatile. Also the SSA algorithm of change-point detection does not announce the presence of essential structural breaks in temperature records. Except perhaps a small change happened around 1998.

Received 20 January 2015

REFERENCES

[1] Ghil, M. and Vautard, R. (1991). Interdecadal oscillations and the warming trend in global temperature time series. Nature, 350, 324–327.

[2] Pepelyshev, A. and Zhigljavsky, A. (2010). Assessing the stability of long-horizon SSA forecasting. Statistics and Its Interface, 3, 321–327. MR2720135

[3] Golyandina, N., Nekrutkin, V. and Zhigljavsky, A. (2001). Analysis of Time Series Structure: SSA and Related Techniques. Chapman & Hall/CRC. MR1823012

[4] Golyandina, N. and Zhigljavsky, A. (2013). Singular Spectrum Analysis for Time Series. Springer Briefs in Statistics, Springer-Verlag. MR3024734

[5] Moskvina, V. and Zhigljavsky, A. (2003). An algorithm based on singular spectrum analysis for change-point detection. Communications in Statistics – Simulation and Computation, 32(2), 319–352. MR1983338

[6] Schlesinger, M. E. and Ramankutty, N. (1994). An oscillation in the global climate system of period 65–70 years. Nature, 367, 723–726.

[7] Vautard, R., Yiou, P. and Ghil, M. (1992). Singular-spectrum analysis: A toolkit for short, noisy chaotic signal. Physica D, 58, 95–126.

[8] Zhigljavsky, A. (2010). Singular spectrum analysis for time series: Introduction to this special issue. Statistics and Its Interface, 3(3), 255–258. MR2720131

Vladimir Kornikov
St. Petersburg State University
Universitetskiy prospekt 28
St. Petersburg, 198504
Russia
E-mail address: vkornikov@mail.ru