Journal of
Flood Risk Management

Modelling fluctuations of Caspian Sea levels using a mixed probability distribution

I. Kozhevnikova[1], V. Shveikina[2] and E. Domínguez[3,4]

1 Moscow State University, Moscow, Russia
2 Institute of Water Problems, Russian Academy of Sciences, Moscow, Russia
3 Pontificia Universidad Javeriana, Bogotá, Colombia
4 CeiBA Complejidad, Bogotá, Colombia

Correspondence
E. Domínguez, Departamento de Ecología y Territorio, Pontificia Universidad Javeriana, Bogotá Transversal 4 no. 42-00, Edificio J. Rafael Arboleda, S.J. Piso 8
Tel: +57 1 320 8320 Ext. 4821
Fax: +57 1 320 8320 Ext. 4859
Email: e.dominguez@javeriana.edu.co

DOI: 10.1111/j.1753-318X.2011.01116.x

Key words
Caspian Sea; hydrological modelling; mixed probability distribution; stochastic processes.

Abstract

By applying a nonlinear regression root analysis, this paper evaluates the significance of an emerging third stable steady state for the Caspian Sea level. A theoretical probability density function (PDF) is built and fitted to empirical data by mixing three different normal distributions. From this mixed probability distribution, a drift coefficient is obtained for the corresponding diffusive process. A stochastic model is then proposed, which uses the derived drift coefficient. This model demonstrates better performance than that exhibited by polynomial approximations and open possibilities for hydrological risk assessment of systems with several stable steady states.

Introduction

This work studies the building of a suitable mathematical model to describe the oscillation dynamics of the Caspian Sea level. This task has been of importance since the level of the Caspian Sea decreased 1.8 m in 9 years when usually the increments of sea water levels had not exceeded 0.4–0.5 m for a period of 100 years. A review of the Caspian Sea level (fluctuation problem has been presented by Asarin 1997). Currently, this problem remains relevant because of ongoing global warming and increasing ocean levels around the world (Naidenov and Kozhevnikova, 2001; Elguindi and Giorgi, 2007). From the historical records of the Caspian Sea levels, several shifts in the annual mean water level (AMWL) from low to high values (and vice versa) have been observed. In the middle of the 16th century, the annual mean Caspian Sea water level was −26.6 m above the Baltic Sea Datum (m.a.b.s.d.). For the following century, an increase in the AMWL took place, reaching an average value of −23.9 m.a.b.s.d. Subsequently, at the beginning of the 17th century, the mean water stage decreased to the mark of −26.0 m. After this intense level depletion, a period of high AMWL began, continuing until the beginning of the 19th century (1805) with a mean annual level record of −22.0 m. From the beginning of standard water level measurements (1830) to the beginning of the 20th century, the Caspian Sea level oscillated around −25.8 m. Following that, from 1900 to 1929, no significant changes were observed, and a mean water level of −26.2 m was established. This almost stable water stage then experienced a rapid decrease of 1.8 m during the period of 1930–1939. With a slower rate, the water level decrease continued until 1960 when the mean water level stabilised at −28.4 m.a.b.s.d. Then, for the first half of the 1970s, the AMWL reached an extreme value of −29.30 m, the lowest mark for the last 150 years (see Figure 1). The water level reduction for the 1900–1977 period reached 3 m. Measuring from its initial position, the shoreline recession reached 10 km. In the expanded shore territory, construction took place. Mathematical models implemented at this time predicted a further decrease in the water level, encouraging the redistribution of water affluences from the northern rivers to feed the Caspian Sea in order to stop a catastrophic water level decrease, but starting in 1978, a extremely intense increase in the water levels began (>1 m in 7 years) stopping at −27.97 m in 1985, flooding all shore facilities that occupied the additional shore produced by the Caspian Sea water level recession. Increasing AMWL values resulted in a catastrophe for Caspian Sea shore inhabitants. The increasing trend of the AMWL continued for another ten years, reaching −26.66 m. Nowadays,

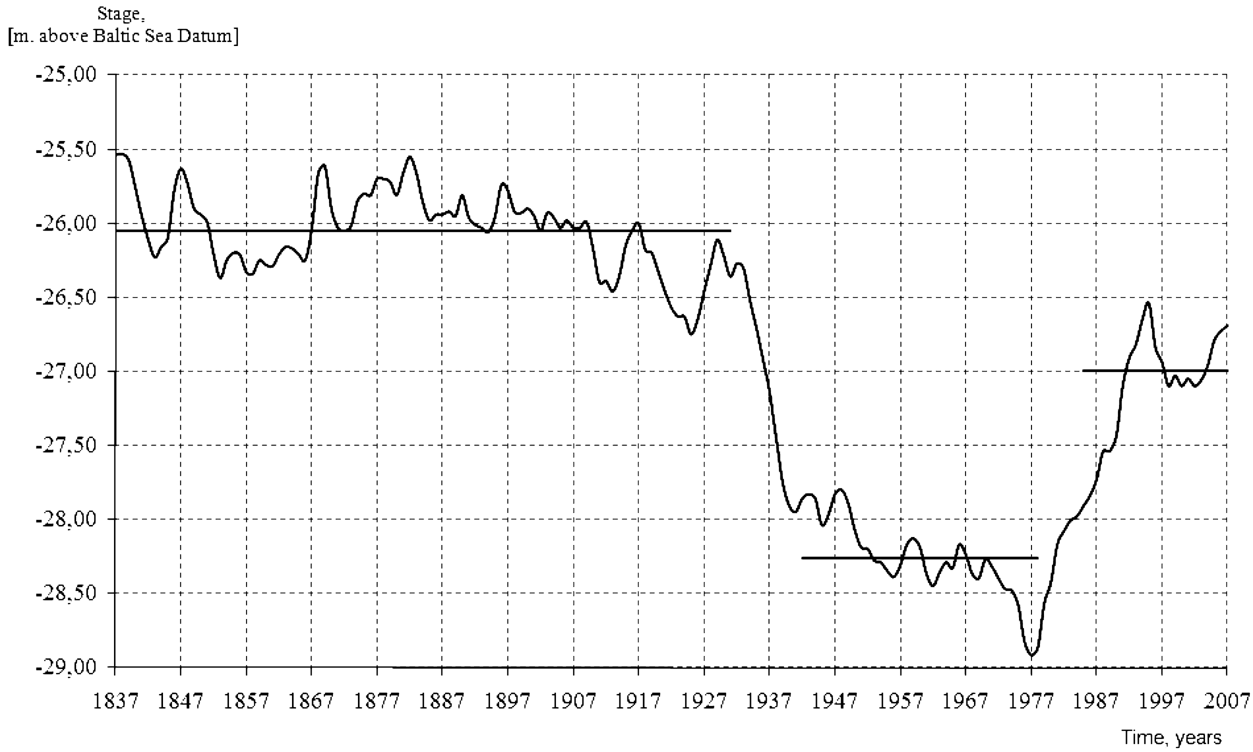

Figure 1 Caspian Sea annual mean water level fluctuations at the Baku City station.

the Caspian Sea AMWL oscillates around −27.00 m. The observed Caspian Sea level dynamics with long-term oscillations about stable AMWL values and sudden AMWL shifts suggest nonlinear patterns that should be represented by nonlinear models able to reflect several stable steady states in their solutions. In fact, the histogram of yearly averaged water level values shows three stable steady states (see Figure 2). This multimodal pattern of the Caspian Sea level and its divergence with a normal probability distribution pattern were first presented in the works of V.I. Naidenov (Naidenov and Podsechin, 1992) for Russian lakes. Other reports have shown nonstationary and non-Gaussian water level behaviour for the Great Lakes (Walton *et al.*, 1990). This phenomenon of non-Gaussian distributions for water levels in closed lakes and reservoirs is a common feature of several lakes and reservoirs (Chad, Chany, Bolshoie Solionoye, Balhash and Khanka lakes), and it is also inherent in diverse natural phenomena with oscillations in their various steady states. This work shows that the level oscillations of the Caspian Sea are caused by the interactions of several random processes, each of which describes a ‘determined component of process variability’ and is conditioned by the influence of a certain driving factor. According to the Intergovernmental Panel on Climate Change (IPCC) framework, there can be found several forecasts for the next 100 years for the Caspian Sea level oscillation behaviour. Some of these forecasts predict a level increase whereas others predict decreasing or stabilisation near the −27.00-m mark. Such forecast diversity with one stable steady state as the outcome reflects neither the observed Caspian Sea patterns nor their future behaviour. This paper developed a model of inland lake water level oscillations for which randomness, nonlinear dynamics and minimal discrepancy between the observed and modelled AMWL probability density distributions are required. In detail, the model fitness for this probabilistic forecast is demonstrated thorough an assessment of the independence of the sequence of residuals, its probability density distribution determination and the absence of hidden periodicities. With this validation, the model can be used to determine the hydrological risk associated with different exceedance probabilities for any particular mean annual water level value. The developed method was applied to the Caspian Sea level case but can be useful for any inland reservoir, river dam, river-related water treatment pool, etc.

Approximation of observed data using a polynomial regression

There are different methods for building a model that describes water level oscillations. The first method employed here, developed by the authors, is based on a polynomial approximation to describe the residuals of the stages. This polynomial approximation is then used as a drift coefficient in a stochastic differential equation for which the corre-

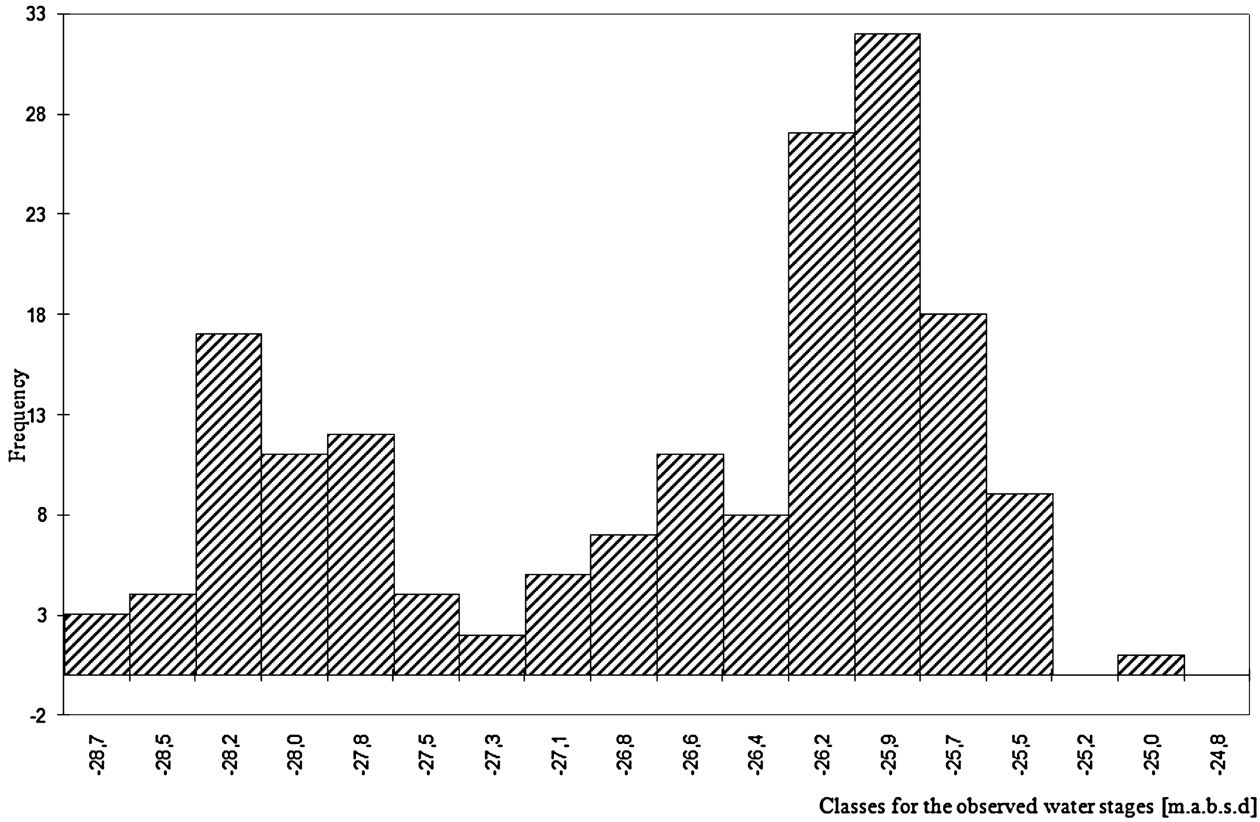

Figure 2 Histogram for observed annual mean water levels of the Caspian Sea.

sponding Fokker–Planck–Kolmogorov (FPK) equation is derived (Siegert *et al*., 1998). Then, from the solution of this FPK equation, we obtain a stationary probability density distribution. This method, which uses a fifth-order nonlinear regression model, was applied by Kozhevnikova and Shveikina to describe the Caspian Sea water level oscillations (Kozhevnikova and Shveikina, 2008). In their study, an observed series from 1900 until 2000 was used, and only two stable steady states were fixed by the regression analysis and were within the histogram for the observed levels. Later, a model of the same type was used with an expanded, up-to-date data set (1900–2006). This data set, with monthly resolution, is presented in Figure 3. In the phase space, with the coordinate system water level $X(t)$ versus water level increments (the most recent difference $\Delta X = X(t) - X(t-1)$), the water level oscillations represent a trajectory that is known as a phase portrait. This trajectory, with a functional dependency, links the water level $X(t)$ with the rate of change at each time point. This trajectory demonstrates a complex pattern, shown in Figure 4. The behaviour of this system is described using a fifth-order regression with coordinates OX representing the water level and OY representing its increment. In order to avoid large regression errors, transformation of the time series $X(t)$ yields the following (Seber, 1977):

$$Z(t) = \frac{2X(t) - X_{\max} - X_{\min}}{X_{\max} - X_{\min}} \tag{1}$$

where $X_{\max} = \max_{1 \le t \le N} X(t)$, $X_{\min} = \min_{1 \le t \le N} X(t)$ and $|Z(t)| \le 1$. The function $Z(t)$ is described as a normalised time series. For the time series presented in Figure 3, $X_{\max} = -29.15$ and $X_{\min} = -25.31$ m above the Baltic Sea datum.

The regression equation that describes the time series $X(t)$ was obtained by applying the minimised squared error method using the observed data from the Makhachkala station, and for the normalised time series, $Z(t)$ is:

$$Z(t+1) = Z(t) + \Phi(Z(t)) + \sigma\varepsilon(t) \tag{2}$$

where $\Phi(Z(t)) = -0.0021 - 0.0315Z(t) + 0.01666Z^2(t) + 0.1741Z^3(t) - 0.1204Z^4(t) - 0.2092Z^5(t)$ and $\varepsilon(t)$ is the model residual, which should be a Gaussian-distributed value with an expected value of zero and a standard deviation of $\sigma = 1$. This regression equation is presented in Figure 5. In order to obtain the required accuracy, the regression coefficients were calculated with a precision of nine decimal places. Establishing the equilibrium position of this curve, the equation $\Phi(Z) = 0$ was solved. The roots for the

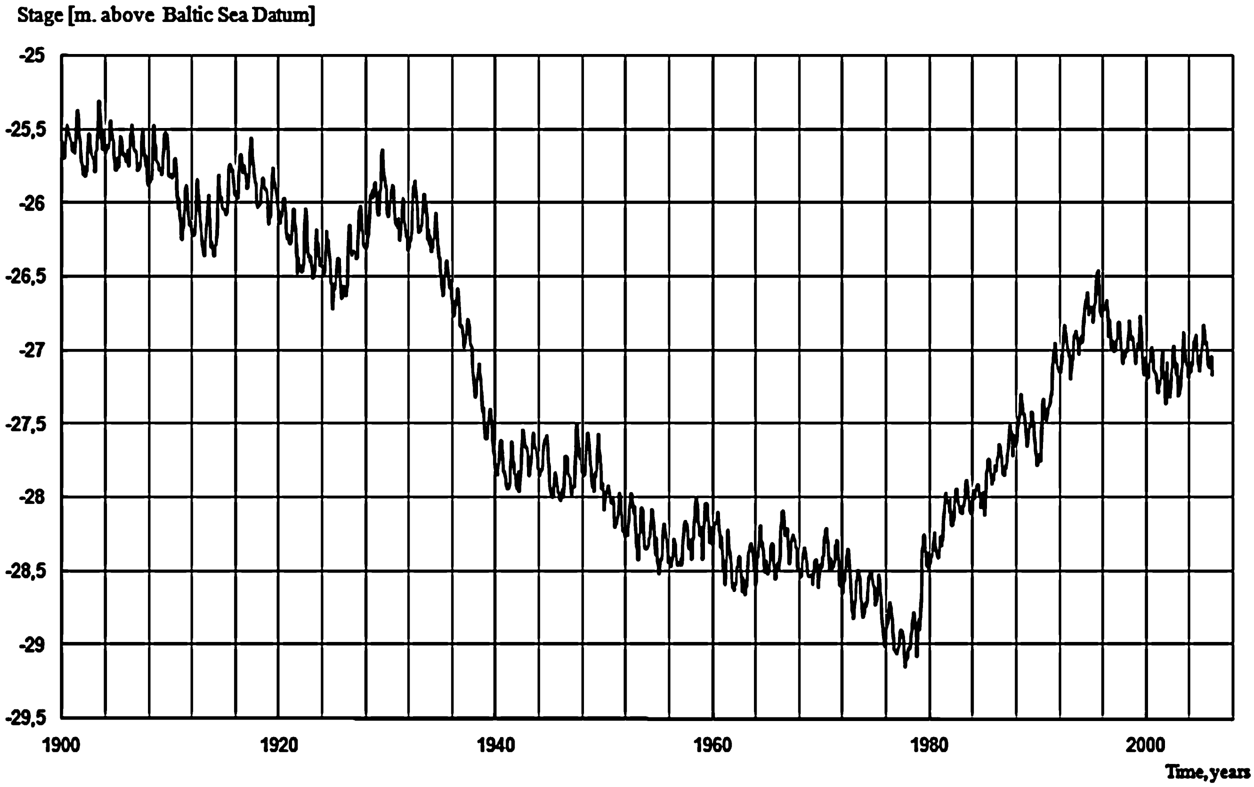

Figure 3 Caspian Sea monthly mean level fluctuations at the Makhachkala City station.

polynomial $\Phi(Z)$ are: $Z_1 = -0.6601$; $Z_2 = -0.3645$; $Z_3 = -0.0226$; $Z_4 = -0.7014$; and $Z_5 = -0.9565$.

A characteristic property of the function $\Phi(Z)$ is the presence of five real roots that give the values for the equilibrium water levels ($X_1 = -28.44$, $X_2 = -27.91$, $X_3 = -27.32$, $X_4 = -26.07$ and $X_5 = -25.63$). However, here, X_1, X_3 and X_5 are stable, and X_2, X_4 are unstable steady states. It is known that the minima of the potential $U(Z) = -\int\Phi(Z)dZ$ correspond to stable steady state conditions, and the maxima correspond to unstable steady states. Additionally, the extreme values for $U(Z)$ agree with the roots of $\Phi(Z)$. The expression $U(Z)$ has three minima at the points Z_1, Z_3 and Z_5, which are stable steady state levels. Using the regression between the increment of the water level and the Caspian Sea level [$\Phi(Z) = -\partial U/\partial Z$], it is possible to use a continuous model for the oscillation of the Caspian Sea level in the form of a diffusion process:

$$dZ_t = -\frac{dU}{dZ}dt + \bar{\sigma}dW_t \tag{3}$$

where U_t is the Caspian Sea level potential, $-\partial U/\partial Z$ is the drift coefficient, $\bar{\sigma}$ is the diffusion coefficient and W_t is a standard Wiener process. The quality of this model depends on the accuracy of the drift coefficient determination. Use of the above-mentioned polynomial regression to establish the drift coefficient does not provide the required accuracy and a better definition of the drift coefficient would provide better performance of the model given by Eqn (3) if the drift coefficient were defined through the observed density distribution function.

Component separation of mixed probability distributions

Component separation of mixed probability distributions is a common issue for modelling several natural processes. In hydrological research, a distribution of one kind is usually used to describe an entire recorded time series. Selection of the distribution depends on the type of problem that is being solved. For instance, for the description of infrequent events, the use of an exponential distribution is more suitable (Naidenov and Shveikina, 2005; Dolgonosov and Korchagin, 2007). For river runoff, a very good description can be reached using the gamma distribution, and in some situations, the three-parameter Weibull distribution (known as the Kritsky–Menkel distribution in Eastern hydrology) can be used. In order to build a mixed distribution, the above-mentioned distributions can be used in addition to the Poisson and binomial distributions, among others (Isaenko

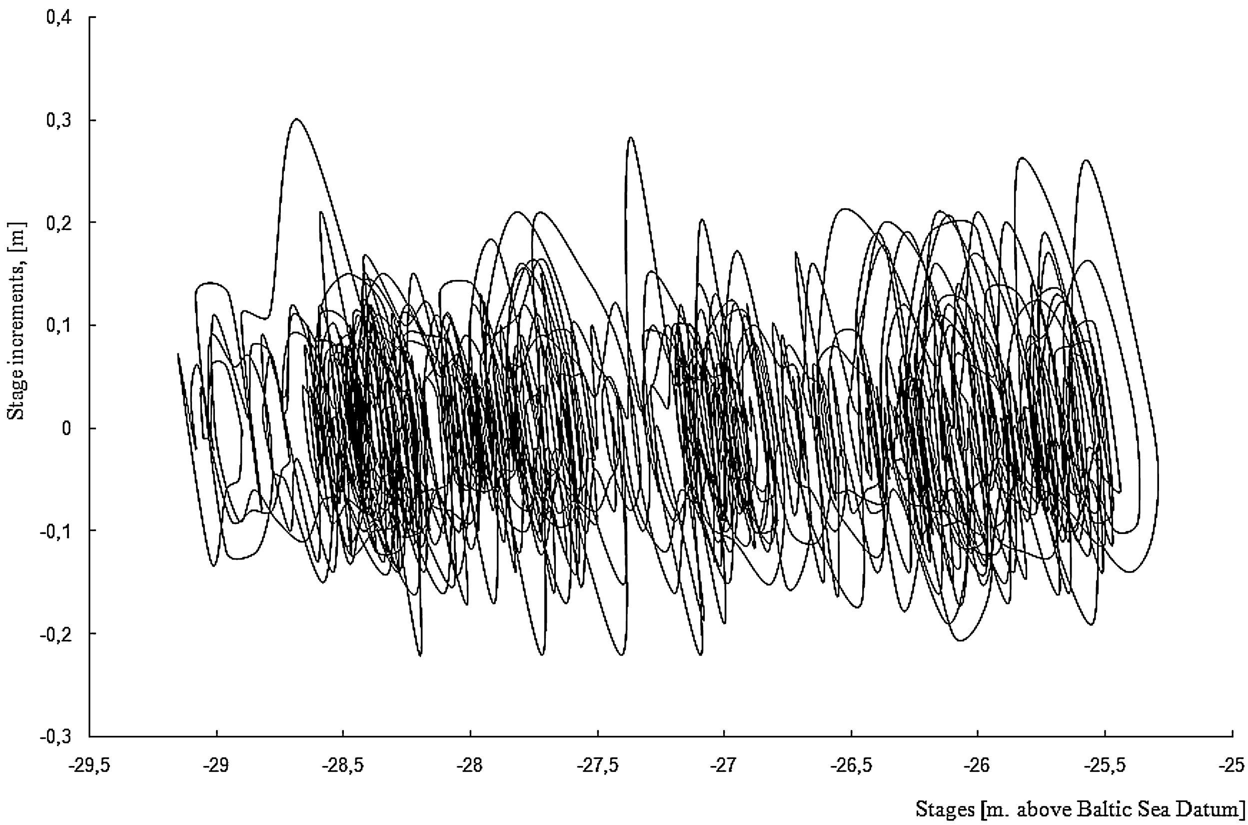

Figure 4 A water level oscillation phase portrait.

and Urbakh, 1976). The estimation of theoretical parameters for a mixed distribution can be done with different methods including the method of moments, the maximum likelihood method, the least squared error method, some types of min–max procedures and graphical methods. In this work, Gaussian distributions were used to build the mixed density distributions, and sample mean and variance values from the stable steady state subseries were applied as initial values to estimate the distribution parameters and to define the first iterative approximation.

The time series analysis allowed for a separation of three stable steady state condition sets. Therefore, the studied process can be approximated using the merger of three normal distributions. The behaviour of the time series allowed for a separation of three periods. The first period includes 37 years of observations and begins at the time when the sea levels were around −26 m above Baltic Sea datum. The second period includes the sudden decrease in water level and its stabilisation near −28 m. The third period contains observations from 1986 until today, during which time the water level oscillates around −27 m. For each mentioned period, using the D'Agostino criterion, we tested and verified the hypothesis regarding the normal fit of the water level oscillations. A chi-squared test is usually applied for such verification, but it is well known that it is better to use specialised distribution-oriented tests, tailored to the normal distribution in this case. In addition, the D'Agostino test is well suited for short length samples ($n = 10 \ldots 12$). The nonsuitability of the chi-squared test for samples with lengths of $n = 10 \ldots 52$ is discussed in (D'Agostino, 1972), where 5000 to 23 500 normally distributed realisations with lengths from 10 to 52 independent elements were tested using the chi-squared test, resulting in a 68% rejection of the null hypothesis of the normality of the analysed, previously generated as normally distributed, data sets. Such a situation has encouraged the application of the D'Agostino test.

The D'Agostino test (Hald, 1952; D'Agostino, 1972) is not commonly used in hydrological assessment tests. For the D'Agostino test, given the observed data $X_1 \ldots X_n$, the variational series is set as $X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}$. One may now introduce the function:

$$D = \frac{T}{n^2 S}, \tag{4}$$

where

$$T = \sum_{i=1}^{n} \left[i - \frac{1}{2}(n+1) \right] X_{(i)} \tag{5}$$

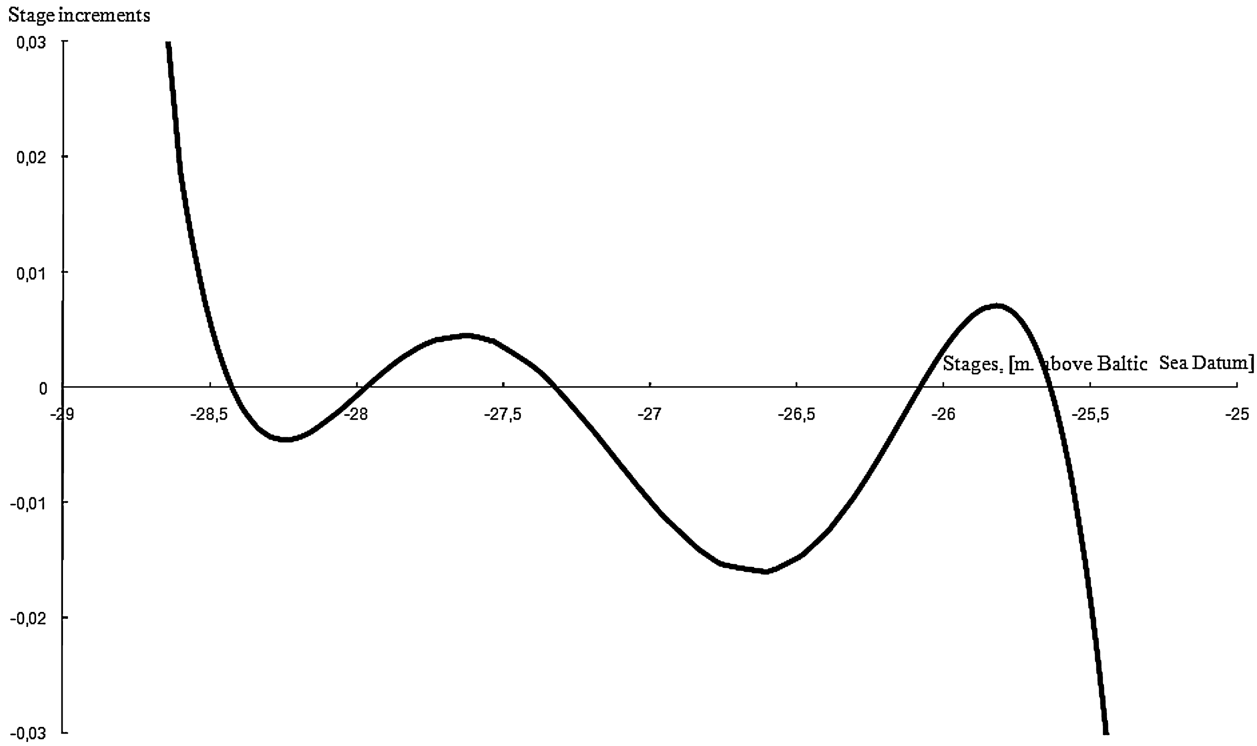

Figure 5 Polynomial regression with five real roots.

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2 \quad (6)$$

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n}(X_i) \quad (7)$$

Previous work (Aivazyan, 1968) has shown that if a subset $X_1 \ldots X_n$ is taken from a normally distributed set, then the asymptotic values for the mathematical expectation and standard deviation for the function D read:

$$E(D) \approx \frac{2}{2\sqrt{\pi}} = 0.28209479 \quad (8)$$

$$\sqrt{Var(D)} \approx \frac{0.02998398}{\sqrt{n}} \quad (9)$$

and then the function:

$$Y = \frac{D - E(D)}{\sqrt{Var(D)}} \quad (10)$$

has a zero mathematical expectation and a standard deviation of one when $n \to \infty$. If the subset does not belong to a normal distribution, then the value for Y will differ from zero by being positive or negative. It has been shown that if the alternative distribution has excess, then the Y-value will be greater than zero as well. Tables with critical values for different significance levels and for short and long data sets are available (Hald, 1952). For instance, for the first period data set (with a mean stage of −26 m), the D'Agostino criterion with a level of significance of 0.2 has a value of −0.137 and a confidence interval of −1.661 to 0.759. With a D'Agostino empirical value of 0.281, this falls within the confidence interval, indicating that, with a significance level of 0.2, the analysed data set follows a normal distribution. Analysis of the second and third period data sets (with mean stages of −27 and −28 m, respectively) with D'Agostino's criterion indicates that these data sets also follow normal distributions. Thus, the problem of a theoretical distribution fit to the empirical data leads to a statistical determination of the unknown mixed distribution parameters.

Theoretical moments of a mixed distribution can be expressed through the first and second moments of each individual distribution according to the expression:

$$f(x) = p_1 f_1\left(x, \mu_1, \sigma_1^2\right) + p_2 f_2\left(x, \mu_2, \sigma_2^2\right) + \left(1 - p_1 - p_2\right) f_3\left(x, \mu_3, \sigma_3^2\right), \quad -\infty < x < \infty \quad (11)$$

where $f_i\left(x, \mu_i, \sigma_i^2\right)$ is the probability density distribution of the i-th component of the mixed distribution and p_i is the weight of the i-th component. To define these parameters, the method of moments was applied. To do this, it is necessary to calculate the moments up to the eighth-order moment because this equation has eight unknown parameters. The equations for the statistical moments of a Gaussian distribution N(μ,σ²) are (Shiriaev, 2007):

$$m_1 = \mu \tag{12}$$

$$m_j = \mu^j + \sum_{k=1}^{\left[\frac{j}{2}\right]} C_j^{2k}\sigma^{2k}\mu^{j-2k}(2k-1)!!, \quad j \geq 2 \tag{13}$$

where j is the order of the normal distribution moment and (2k − 1)!! is the product of odd numbers through (2k − 1) inclusively. The equations for the moments of a mixed distribution with q components have the form:

$$m_{q,1} = \sum_{i=1}^{q} p_i\mu_i, \quad q = 3, \tag{14}$$

$$m_{q,j} = \sum_{i=1}^{q} p_i\left(\mu_1^j + \sum_{k=1}^{\left[\frac{j}{2}\right]} C_j^{2k}\sigma_i^{2k}\mu_i^{j-2k}(2k-1)!!\right), \quad j \geq 2, \tag{15}$$

$$\sum_{i=1}^{q} p_i = 1 \text{ and } p_3 = 1 - p_1 - p_2 \tag{16}$$

Taking in account the above equations, one can set the following expressions:

$$m_1 = p_1\mu_1 + p_2\mu_2 + (1 - p_1 - p_2)\mu_3 \tag{17}$$

$$m_2 = p_1(\mu_1^2 + C_2^2\sigma_1^2) + p_2(\mu_2^2 + C_2^2\sigma_2^2) + (1 - p_1 - p_2(\mu_3^2 + C_2^2\sigma_3^2)) \tag{18}$$

$$\begin{aligned} m_3 = {} & p_1(\mu_1^3 + C_3^2\sigma_1^2\mu_1) + p_2(\mu_2^3 + C_3^2\sigma_2^2) \\ & + (1 - p_1 - p_2)(\mu_3^3 + C_3^2\sigma_3^2) \end{aligned} \tag{19}$$

$$\begin{aligned} m_4 = {} & p_1(\mu_1^4 + C_4^2\sigma_1^2\mu_1^2 + C_4^4\sigma_1^4 \cdot 1 \cdot 3) \\ & + p_2(\mu_2^4 + C_4^2\sigma_2^4\mu_2^2 + C_4^4\sigma_2^4 \cdot 1 \cdot 3) \\ & + (1 - p_1 - p_2)[A_4(\mu_3, \sigma_3)] \end{aligned} \tag{20}$$

$$\begin{aligned} m_5 = {} & p_1(\mu_1^5 + C_5^2\sigma_1^2\mu_1^3 + C_5^4\sigma_1^4\mu_1^1 \cdot 1 \cdot 3) + p_2[A_5(\mu_2, \sigma_2)] \\ & + (1 - p_1 - p_2)[A_5(\mu_3, \sigma_3)] \end{aligned} \tag{21}$$

$$\begin{aligned} m_6 = {} & p_1(\mu_1^6 + C_6^2\sigma_1^2\mu_1^4 + C_6^6\sigma_1^4\mu_1^2 \cdot 1 \cdot 3 + C_6^6\sigma_1^6 \cdot 1 \cdot 3 \cdot 5) \\ & + p_2[A_6(\mu_2, \sigma_2)] + (1 - p_1 - p_2)[A_6(\mu_3, \sigma_3)] \end{aligned} \tag{22}$$

$$\begin{aligned} m_7 = {} & p_1(\mu_1^7 + C_7^2\sigma_1^2\mu_1^5 + C_7^4\sigma_1^4\mu_1^3 \cdot 1 \cdot 3 + C_7^6\sigma_1^6\mu_1 \cdot 1 \cdot 3 \cdot 5) \\ & + p_2\mathrm{A}_7(\mu_2, \sigma_2) + \mathrm{B}_7 \end{aligned} \tag{23}$$

$$\begin{aligned} m_8 = {} & p_1(\mu_1^8 + C_8^2\sigma_1^2\mu_1^6 + C_8^4\sigma_1^4\mu_1^4 \cdot 1 \cdot 3 + C_8^6\sigma_1^6\mu_1^2 \cdot 1 \cdot 3 \cdot 5 \\ & + C_8^8\sigma_1^8 \cdot 1 \cdot 3 \cdot 5 \cdot 7) + p_2\mathrm{A}_8(\mu_3, \sigma_3) + \mathrm{B}_8 \end{aligned} \tag{24}$$

where $B_k(1 - p_1 - p_2)A_k$, k = 7.8 and

$$A_j = \mu_i^j + \sum_{k=1}^{\left[\frac{j}{2}\right]} C_j^{2k}\sigma_i^{2k}\mu_i^{j-2k}(2k-1)!!, \quad i = 4, \ldots, 8.$$

Table 1 Parameters of the merged distribution with q components

	q = 1	q = 2	q = 3
μ_q	$\mu_1 = -25.86$	$\mu_2 = -28.27$	$\mu_3 = -26.97$
σ_q	$\sigma_1 = 0.548$	$\sigma_2 = 0.439$	$\sigma_3 = 0.192$
p_q	$p_1 = 0.50$	$p_2 = 0.61$	$p_3 = 0.16$

Noncentred statistical moments in the left side of the above equations can be evaluated directly from the observed data as:

$$\hat{m}_i = \frac{1}{n}\sum_{k=1}^{n} x_k^i \tag{25}$$

where i is the order of the given moment.

To find the mixed distribution parameters, the minimisation of the following goal function is applied:

$$Q = \sum_{j=1}^{8}\left(\frac{m_j - \hat{m}_j}{\hat{m}_j}\right)^2 \tag{26}$$

Using an iterative method to solve the above algebraic systems, the parameters for the merged probability distribution were found (Table 1).

The values in Table 1 allow us to build a merged distribution that is well fitted to the empirical distribution (Figure 6).

Solution of the inverse problem for a model of the Caspian Sea oscillations

Traditionally, if the selected model is a diffusion process model, then the corresponding FPK equation can be obtained and the stationary probability density distribution can be found (Sveshnikov, 1968; Gardiner, 1985; Siegert *et al.*, 1998). The obtained distribution differs from the empirical distribution. Therefore, a special approach is used, which is based on the use of a theoretical probability density function (PDF) fitted to the empirical PDF. Thus, it is supposed that the fitted PDF satisfies the FPK equation. Using the found theoretical PDF, a stochastic model is then built. Following this approach, the theoretical PDF is made by mixing three Gaussian distributions, and this mixed PDF is defined by eight parameters. Recalling the model Eqn (3) for the Caspian Sea oscillations, the corresponding FPK equation reads:

$$\frac{\partial p(y,t|x)}{\partial t} = -\frac{\partial[\Phi(y)p(y,t|x)]}{\partial y} + \frac{\bar{\sigma}^2}{2}\frac{\partial^2[p(y,t|x)]}{\partial y^2} \tag{27}$$

The stationary solution for this FPK equation, assuming ∂p/∂t = 0, is expressed as:

$$p_s(x) = \frac{C}{\bar{\sigma}^2}\exp\left[\frac{2}{\bar{\sigma}^2}\int_{-\infty}^{x}\Phi(u)du\right], \quad X_{\min} \leq x \leq X_{\max} \tag{28}$$

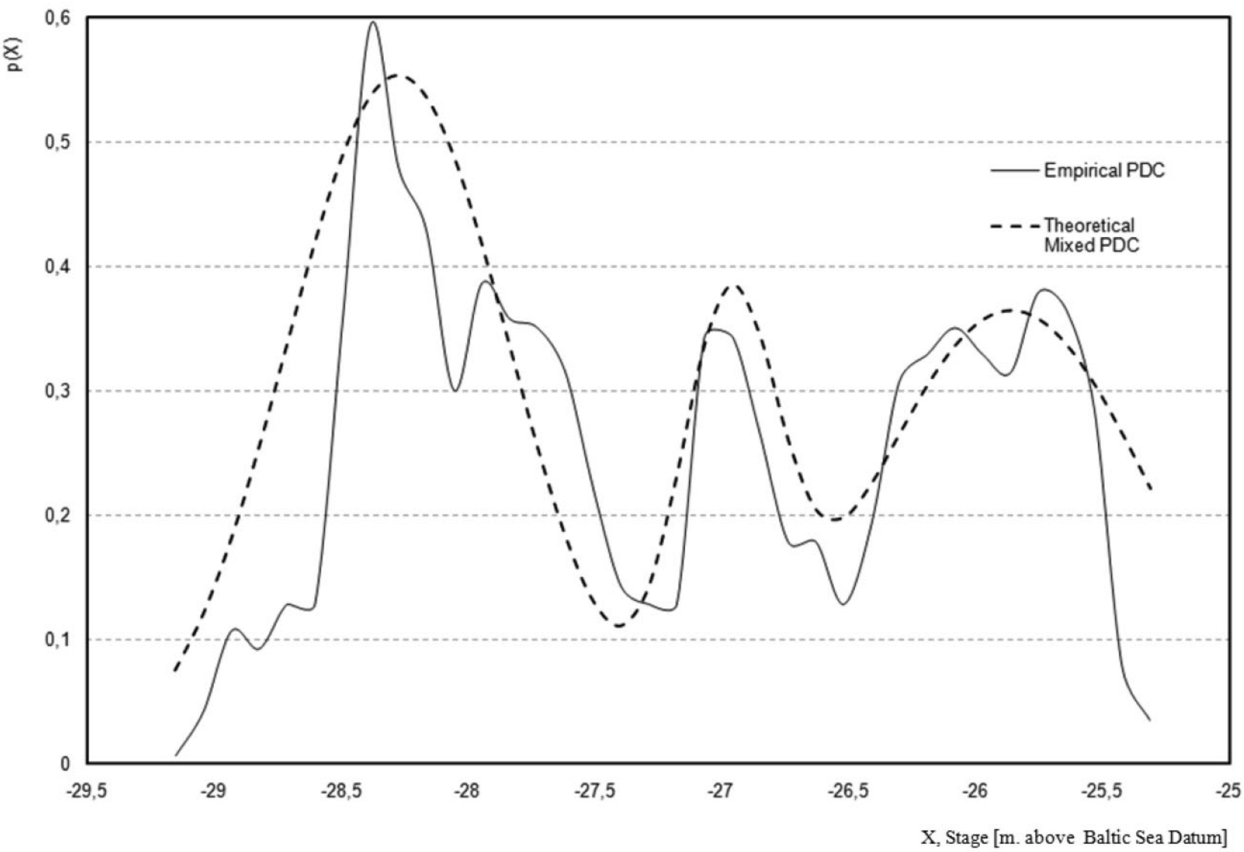

Figure 6 Empirical and theoretical mixed probability density functions. PDC: Probability Density Curve.

where C is a normalisation coefficient and $\bar{\sigma}^2 = 0.011$. Substituting in Eqn (28), the drift coefficient, as:

$$\Phi(x) = -\frac{dU(x)}{dx} \tag{29}$$

we obtain the following expression for the stationary probability density:

$$p_s(x) = \frac{C}{\bar{\sigma}^2}\exp\left[-\frac{2}{\bar{\sigma}^2}U(x)\right] \tag{30}$$

Supposing that the stationary probability density that satisfies the FPK equation is expressed through a mixed distribution, $f(x)$, with eight parameters, then we have:

$$f(x) = \frac{C}{\bar{\sigma}^2}\exp\left[-\frac{2}{\bar{\sigma}^2}U(x)\right] \tag{31}$$

Taking the logarithm of Eqn (30), one can derive the expression for $U(x)$ as:

$$\frac{2}{\bar{\sigma}^2}U(x) = \ln\frac{C}{\bar{\sigma}^2} - \ln f(x) \tag{32}$$

Then, the drift coefficient reads:

$$\Phi(x) = -\frac{dU}{dx} = \frac{\bar{\sigma}^2}{2}\frac{f'(x)}{f(x)} \tag{33}$$

and the Caspian Sea oscillation model becomes:

$$dX_t = \frac{\bar{\sigma}^2}{2}\frac{f'(X)}{f(X)}dt + \bar{\sigma}dW_t \tag{34}$$

where

$$f(X) = \frac{1}{\sqrt{2\pi}}\sum_{k=1}^{3}\frac{p_k}{\bar{\sigma}_k}\exp\left[-\frac{(X-\mu_k)^2}{2\bar{\sigma}_k^2}\right] \tag{35}$$

and

$$f'(X) = -\frac{1}{\sqrt{2\pi}}\sum_{k=1}^{3}\frac{p_k(X-\mu_k)}{\bar{\sigma}_k^3}\exp\left[-\frac{(X-\mu_k)^2}{2\bar{\sigma}_k^2}\right] \tag{36}$$

In a previous model, some of the authors of this paper expressed the drift coefficient with a fifth-order polynomial. Here, it is expressed in a more complex manner (Figure 7), but the scatter plot dispersion is better represented. It is expected that modelling with this drift coefficient will show improved performance.

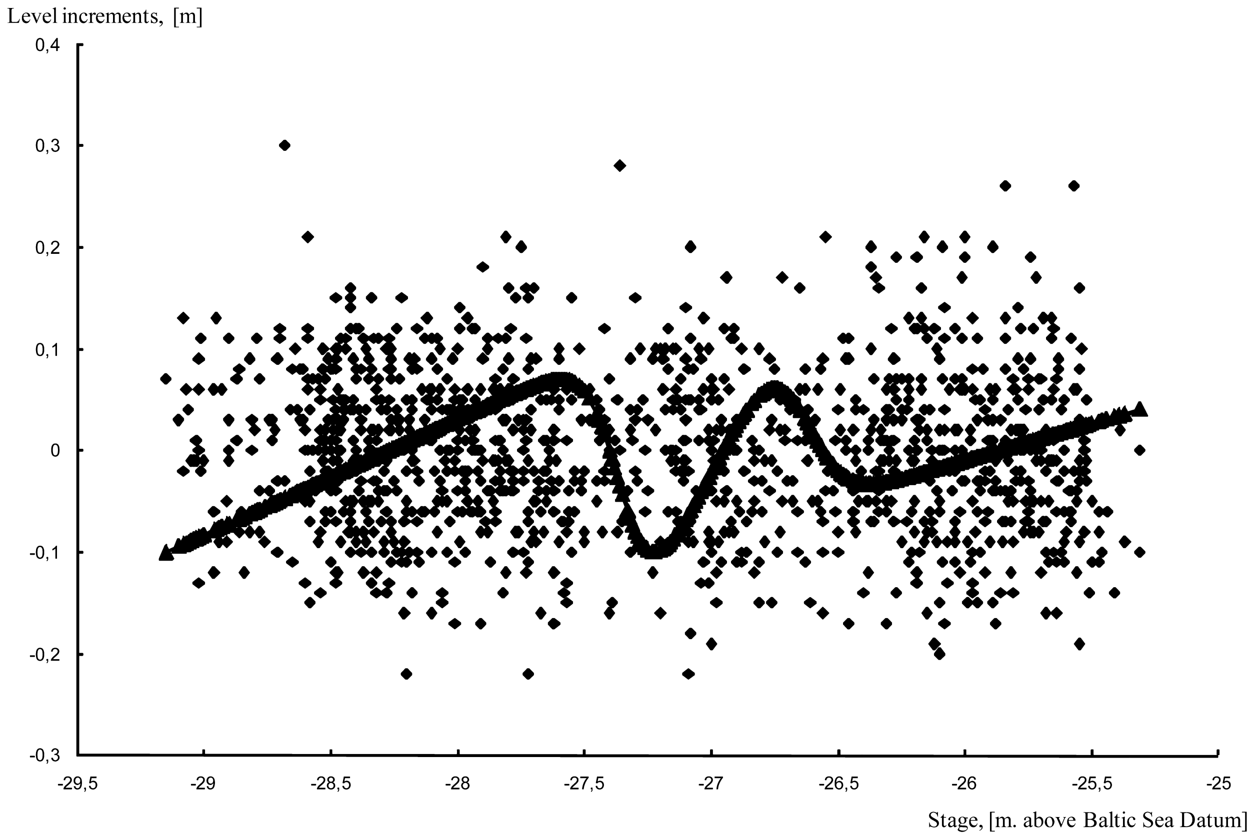

Figure 7 Diffusion process drift coefficient.

The performance of the proposed model is characterised by the sum of the squared errors for the increments of observed data. In this case, the sum is 2.057, and therefore, the residual dispersion will be $2.057/1272 \cong 1.62 \times 10^{-3}$. For reference, the well-known human population models of S.P. Kapitza (Kapitza, 1992) have a residual dispersion on the order of 7.6×10^{-2}, 2.9×10^{-2} and 1.7×10^{-2}. The model presented here has a lower residual dispersion than the Kapitsa model, so the model used to describe the observed water level data is acceptable.

The residual dispersion for the model that uses a drift coefficient defined by the fifth-order nonlinear regression is equal to 11.81, indicating that the approach using the composite PDF allows for a more precise stochastic model of the Caspian Sea fluctuations. Data generated by the stochastic model proposed here is presented in Figure 5.

Conclusions

In the work presented here, a number of problems are solved. First, an improved mathematical model of the Caspian Sea water level oscillation, in comparison with the model based on the polynomial approach for the drift coefficient, was built (Figure 8). The developed model is nonlinear and leads to a stationary probability density distribution that is quite similar to the observed density distribution. The result obtained with the polynomial approach, a stationary density distribution, usually differs from the observed distribution. The stationary distribution arises in the system as $t \to \infty$, which means that the probability distribution becomes independent of time. Perhaps the system fluctuations do not tend to any particular realisation. The system itself continues fluctuating, but with increasing time, the distribution will tend to a particular probability distribution that finally will not change with time, thus being stationary. The proposed model is required to tend to the same stationary distribution when $t \to \infty$. Further development of the proposed model will allow for not only a description of the actual dynamics, but a probabilistic forecast for risk assessment as well. Secondly, the statistical analysis of the Caspian Sea level behaviour established the emergence of a third stable steady state in addition to two previously known states. This highlights the climatic transformations that influence the Caspian Sea stages. The search for these climatic driving factors is not a goal of the present statistical analysis, which, in this case, only indicates the existence of such phenomena.

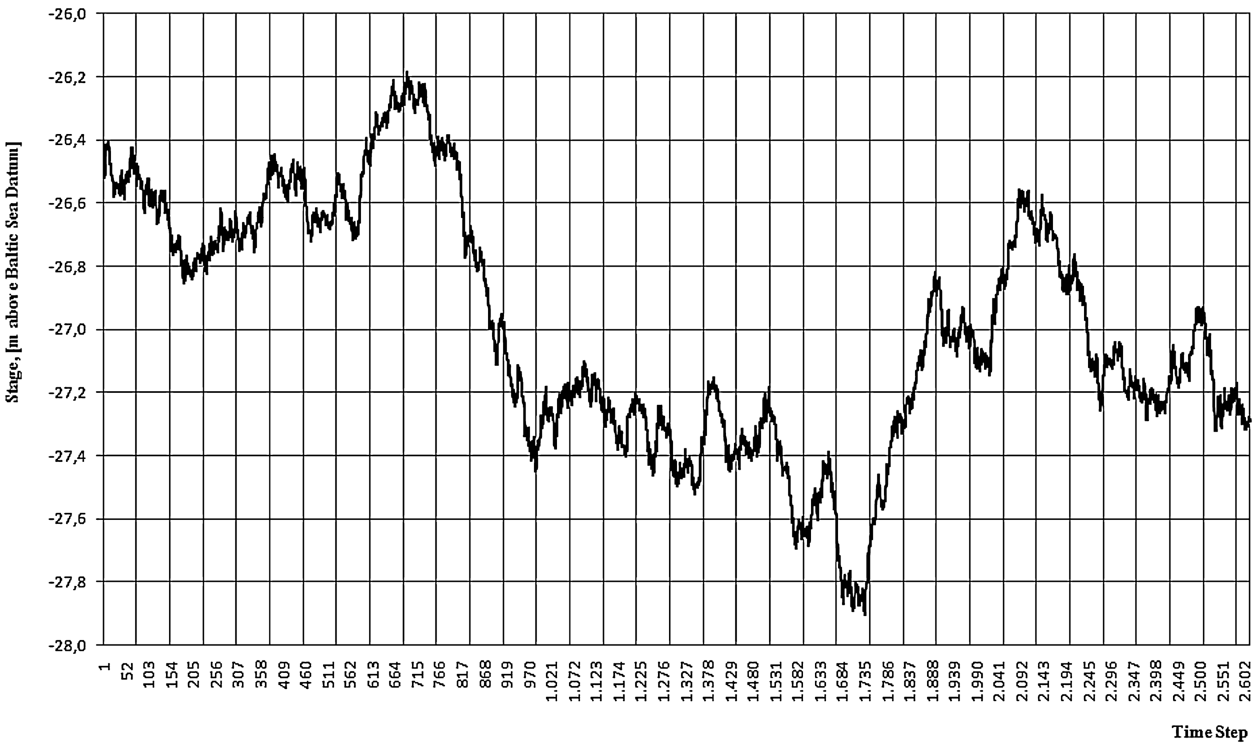

Figure 8 The characteristic realisation of the Caspian Sea level fluctuation model.

The emergence of a third stable steady state for the stage oscillations of the Caspian Sea leads to a novel, not yet solved problem: what is the transition time from one stable steady state to another? In the potential function that describes Caspian Sea level dynamics, three steady states emerge. An assessment of the probability that one of these stages will settle into one of these observables requires more research.

Acknowledgements

The authors wish to acknowledge the reviewers for their helpful comments, which led to the improvement of the final version of this paper.

References

Aivazyan S.A. *Statistical study of the relationships*. Moscow: Mir, Metallurgia, 1968.

Asarin A.E. Problem of Caspian Sea level fluctuations. *Hydrotecnical Construction* 1997, **31**, 645–654.

D'agostino R.B. Small sample probability points for the D test of normality. *Biometrika* 1972, **59**, 219–221.

Dolgonosov B.M. & Korchagin K.A. A non-linear stochastic model describing the formation of daily and mean monthly water flow in river basins. *Water Resour* 2007, **34**, 624–634.

Elguindi N. & Giorgi F. Simulating future Caspian Sea level changes using regional climate model outputs. *Clim Dyn* 2007, **28**, 365–379.

Gardiner C. *Handbook of stochastic methods*. Berlin: Springer-Verlag, 1985.

Hald A. *Statistical theory with engineering applications*. New York: John Wiley & Sons Inc, 1952.

Isaenko O.K. & Urbakh V.Y. *Separation of mixed probability distribution into components. Achievements of science and technique. Theory of probabilities, mathematical statistics and theorethical cybernetics*. Moscow: VINITI, 1976.

Kapitza S.P. A mathematical model for global population growth. *Math Model* 1992, **4**, 65–79.

Kozhevnikova I.A. & Shveikina V.I. Nonlinear dynamics of level variations in the Caspian Sea. *Water Resour* 2008, **35**, 297–304.

Naidenov V.I. & Kozhevnikova I.A. Nonlinear variations of the level of the Caspian sea and the global climate. *Dokl Phys* 2001, **46**, 340–345.

Naidenov V.I. & Podsechin V.P. A nonlinear mechanism of water level fluctuations of inland reservoirs. *Water Resour* 1992, **6**, 5–11.

Naidenov V.I. & Shveikina V.I. Hydrological theory of the hearth global warming. *Russ Meteorol Hydrol* 2005, **12**, 31–38.

Seber G.A.F. *Linear regression analysis*. New York: John Wiley & Sons, 1977.

Shiriaev A.N. *Probability*. Moscow: Nauka, 2007.

Siegert S., Friedrich R. & Peinke J. Analysis of data sets of stochastic systems. *Physics Lett A* 1998, **243**, 275–280.

Sveshnikov A.A. *Problems in probability theory, mathematical statistics and theory of random functions*. Philadelphia: Saunders, 1968.

Walton T.L. & Borgman L.E. Simulation of nonstationary, non-Gaussian water levels on Great Lakes. *J Waterway Port Coast Ocean Eng* 1990, **116**, 664–685.