Romanian Reports in Physics, Vol. 68, No. 2, P. 788–798, 2016

MODELING TEMPERATURE EVOLUTION. CASE STUDY

A. BARBULESCU

Technical University of Civil Engineering, Doctoral School,
122–124 Lacul Tei Bd., Bucharest, Romania
E-mail: alinadumitriu@yahoo.com

Received: March 31, 2014

Abstract. In this article we present the results of the study of daily temperatures collected in 2010, at Techirghiol, Romania. Statistical analyses of the air and water temperature of Techirghiol Lake have been performed and mathematical models have been proposed. Our approach is based on a hybrid technique: the decomposition of data series in a deterministic part and a residual one, followed by the residual modelling using the general regression neural network technique. Also, the correlation between the air and water temperature has been studied, proving the existence of a linear dependence between them.

Key words: Techirghiol Lake, temperature, trend, residual, GRNN.

1. INTRODUCTION

The climate dynamic is a topic of big interest due to the big impact of the meteorological changes on the human and economic life. The lakes are systems whose characteristics and dynamic are also influenced by the climate conditions. Preserving the water quality is important for the impact on the human and environmental health. Therefore, an increasing number of researches was performed on this field, as those done for Romania [1–7].

Techirghiol Lake, located in Dobrudja, Romania, 15 km South from Constanta City (Fig. 1), has a surface of 10.68 km² and is separated from the sea by a sand-belt with a maximum depth of 9 m. It is a hypersaline lake, whose sapropelic mud has curative properties [8]. The main water supply of the lake is formed by the rivers Movilita, Biruinta and Techirghiol and the fresh water springs.

In 2000, after the declaration of Techirghiol Lake as protected area, a campaign for monitoring the environment quality, especially of the physics and chemical parameters of lake water' started. Contributions regarding the ecological and hydro-chemical features of this lake have been brought in [1, 6, 9, 10] and of its morphometrical and hydrological features, in [11], but studies on the relations between the water and air temperature, that influences the evaporation and the salinity of the lake, are not registered. Therefore, to complete the knowledge on the climate in the area of Techirghiol Lake, we present the results of the statistical analysis of water and air temperatures and we model their evolution and dependence.

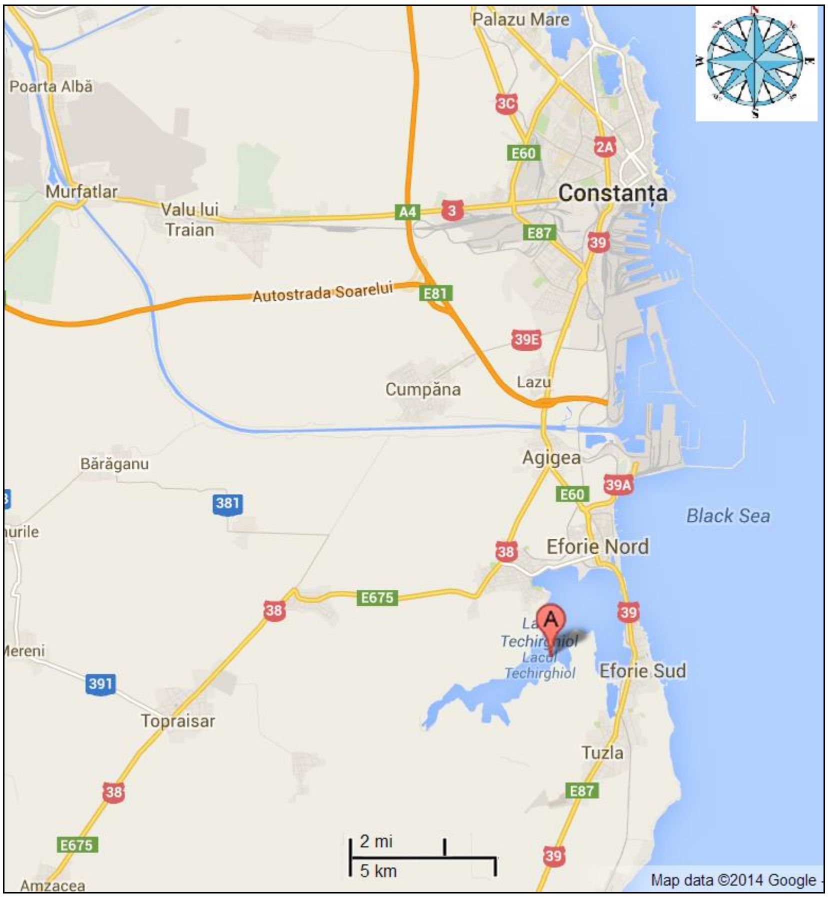

Fig. 1 – Location of Techirghiol Lake on the map of Dobrogea district.

2. METHODOLOGY

The series analysed are formed by the daily air and water mean temperatures, registered in the period 1st of January – 31st of December 2010, data taken from the Annual reports of the Agency for Environment, Constanta and the National Agency of Meteorology.

Statistical tests have been performed for checking different hypotheses on the data series: the Kolmogorov-Smirnov and Shapiro-Wilk tests (for normality), the autocorrelation function (for autocorrelation), the box plot (for outliers' detection)

[12], the break point detection (by the segmentation procedure of Hubert [13] and the modified dynamic programming algorithm – mDP [14, 15]).

Most of the univariate methods for outliers' detection rely on the assumption of an underlying known distribution of the data, which is assumed to be identically and independently or normally distributed [16]. When a variable is not normally distributed, a boxplot may be more effective in identifying outliers. Cases with values between 1.5 and 3 box lengths (inter-quartile range) from the upper or lower edge of the box are identified as outliers.

Segmentation of a time series means dividing it into sub-series with statistical characteristics that are similar within each sub-series and different between sub-series [14]. The segmentation procedure of Hubert belongs to the segmentation with regression-by-constant. Its principle is to cut the series in m segments (m > 1) such that the calculated means of the neighbours sub-series significantly differ. To limit the segmentation, the means of two contiguous segments must be different. This constraint is satisfied through the use of the Scheffé's test [13].

mDP [15] is the modification of the DP algorithm by the remaining cost concept of Gedikli [14]. The problem can be formulated as an optimization one. If $(x_t)_{t=\overline{1,T}}$ is a series and $0 < t_1 < t_2 <...< t_K = T$, the segmentation cost *J*(*t*) is defined as $J(t) = \sum_{k=1}^{K} d_{t_{k-1}+1, t_k}$, where $d_{s,t}\,(0 \le s < t \le T)$ is the segment error corresponding to $[s,t]$, given by: $d_{s,t} = \sum_{\tau=s}^{t} \left(x_\tau - \left(\sum_{j=s}^{t} x_j \right) / (t-s+1) \right)^2$. The optimal segmentation is defined as $\hat{t} = \arg\min_{t \in T} J(t)$ [15].

For modeling purposes two techniques were used:

1) To build models for the air and water temperatures, we decomposed the series into a sum of a deterministic trend and a residual that has been modelled using the generalized regression neural network technique (GRNN). This type of neural network has four layers (Fig. 2):

– *Input layer*, containing one neuron, for each predictor variable;

– *Hidden layer*, containing one neuron for each case in the training data set, and storing the predictor variables values and target values. The activation function used was the exponential one and the kernel function was the Gaussian one;

– *Summation layer*, formed by two neurons: the denominator summation unit (that adds up the weight values coming from each of the hidden neurons) and the numerator summation unit (that adds up the weight values multiplied by the actual target value for each hidden neuron);

– *Decision layer*, formed by one neuron, which divides the values from the previous layer.

The conjugate gradient method has been used to refine the parameters estimates.

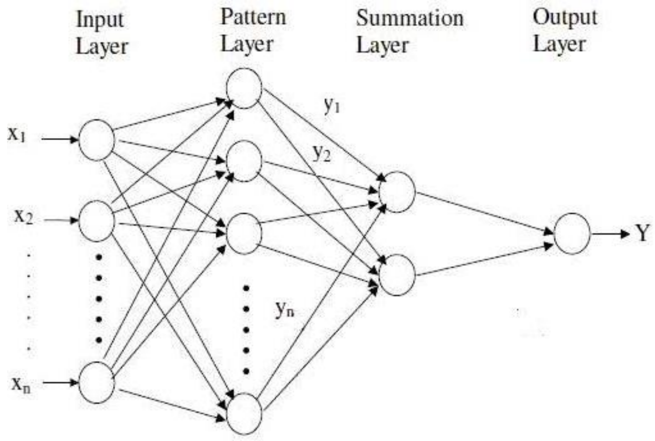

Fig. 2 – Scheme of GRNN.

The data was divided in two parts, one for training and one for validation. The estimation of models' quality was done by analysing the correlation between the actual and the predicted values, the maximum and the mean square error. For details concerning GRNN [17].

2) For modeling the dependence between the temperature of water of Techirghiol Lake and the air temperature, a linear regression model was proposed. We remember that this type of model has the equation:

$$y_t = \beta_0 + \beta_1 x_t + z_t\,,\ t = \overline{1,T}\,,$$

where: y_t is the dependent variable, x_t – the independent one and z_t – the residual.

The conditions that must be satisfied by the variables z_t are: the expectance of z_t is null for all *t*; the variance of z_t is constant, for all *t*; (z_t) are not correlated, are Gaussian and identically distributed.

The work has been carried on by using Khronostat, Minitab, segmenter.21 and DTREG softwares.

3. RESULTS AND DISCUSSION

Firstly, we present the results of statistical analysis. The Kolmogorov-Smirnov and Shapiro-Wilk test have been performed at the significance level of

0.05. The probability plots for both series are presented in presented Fig. 3, where: mean is the average of the registered values, StDev – the standard deviation, N – the data number, KS – the values of the Kolmogorov-Smirnov statistics, and
P-value the minimum probability value at which the null hypothesis can not be rejected. If the P-value is less than the significance level, the normality hypothesis is rejected, as in our cases.

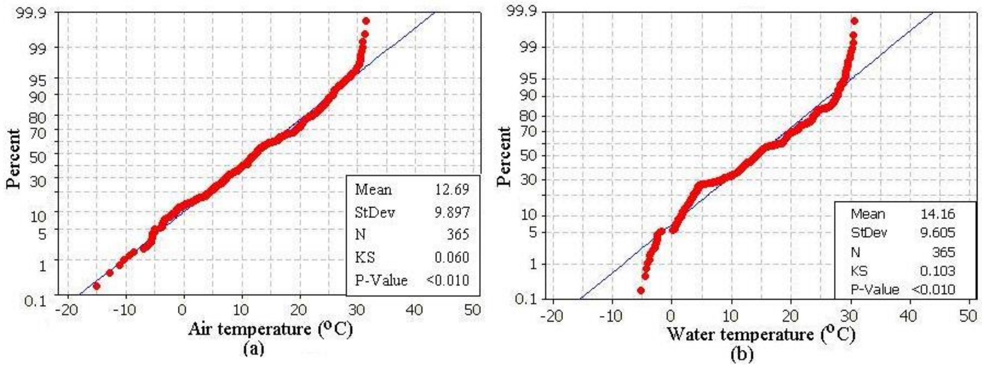

Fig. 3 – Probability plot of: (a) air temperature; (b) water temperature.

After the study of autocorrelation function, the hypothesis that the series is not correlated has been rejected. The boxplot analysis (Fig. 4) leads us to reject the hypothesis of the outliers' existence for both series.

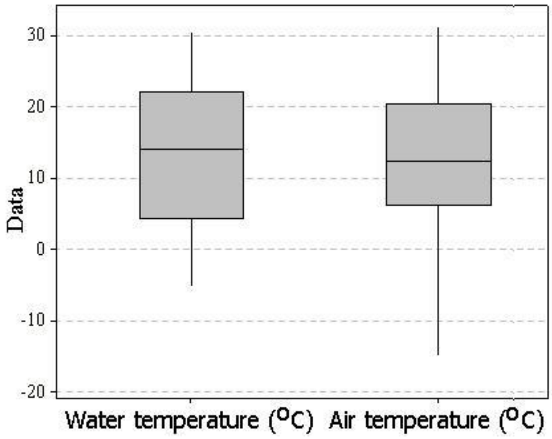

Fig. 4 – Boxplot of air and water temperature series.

For comparison reasons both algorithms for change points detections have been used setting the confidence level of 99% in the Scheffé test. The change points provided by Hubert’s procedure are presented in Table 1 (the first four columns for air temperatures series and the last four, for water temperatures series). Analyzing the results we remark the existence of the same number of the change points, those for the water temperature appearing after those of air temperature, as expected, since the last ones are influenced by the first ones.

Table 1

Change points in the series of air and water temperatures (Hubert’s procedure)

Segments of air temperatures series				Segments of water temperatures series			
Beginning	End	Mean	StDev	Beginning	End	Mean	StDev
1 Jan	13 Jan	19.123	4.382	1 Jan	17 Feb	6.246	2.828
14 Jan	10 Feb	11.439	4.132	18 Feb	29 Mar	10.72	1.821
11 Feb	19 Mar	20.305	2.718	30 Mar	02 May	19.747	1.956
20 Mar	23 May	28.512	3.136	3 May	06 Jun	25.834	1.466
24 May	13 Jul	37.986	2.487	7 Jun	19 Jul	30.981	1.615
14 Jul	28 Aug	43.298	2.445	20 Jul	01 Sep	34.686	1.235
29 Aug	30 Sep	35.164	1.756	2 Sep	02 Oct	26.165	1.716
1 Oct	30 Nov	27.177	2.465	3 Oct	09 Dec	18.447	2.051
1 Dec	31 Dec	16.635	4.676	10 Dec	31 Dec	9.195	1.884

The data series (the grey lines) and the segments (the black lines) obtained by running mDP are drawn in Fig. 5.

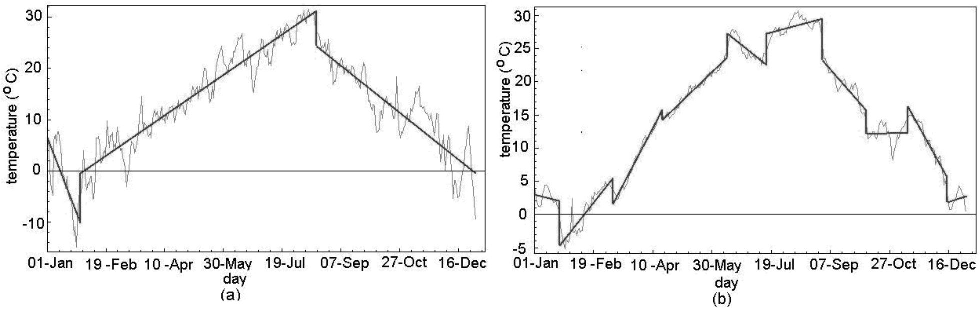

Fig. 5 – Results of mDP algorithm for change points’ detection for the series of:
a) air temperature; b) water temperature.

The segments are: 1 Jan – 28 Jan, 29 Jan – 17 Aug, 18 Aug – 31 Dec, for air temperature, respectively: 1 Jan – 21 Jan, 22 Jan – 7 Mar, 8 Mar – 18 Apr, 19 Apr –12 Jun, 13 Jun –15 Jul, 16 Jul – 31 Aug, 1 Sep – 7 Oct, 8 oct – 11 Nov, 12 Nov – 14 Dec, 15 Dec – 31 Dec. mDP algorithm provided a different number of segments by comparison to the Hubert procedure, due to the technique of optimization used. Since the break points are not the same, we decided to use the entire series for modeling. Since the data has high variability, due to the seasonal component, we choose a hybrid approach for modeling the series.

To present the models, let us denote by: y_t – the values of water temperature, x_t – those of the air temperature, $\hat{x}_t$ – the values of the deterministic trend of air temperature, $\hat{y}_t$ – the values of the deterministic trend of water temperature, $t = \overline{1,365}$ being the time (1 corresponds to 01.01.2010 and 365 to 31.12.2010).

3.1. MODEL FOR WATER TEMPERATURE

Firstly, a cyclic trend has been determined. It has the equation:

$$\hat{y}_t = 13.6654 + 13.2957\cos(0.0165t - 3.41393). \tag{1}$$

Let us denote by $e_t = y_t - \hat{y}_t$, $t = \overline{1,365}$, the error, *i.e.* the difference between the registered data and the trend's estimated value. The data, the trend and the error are represented in Fig. 6a.

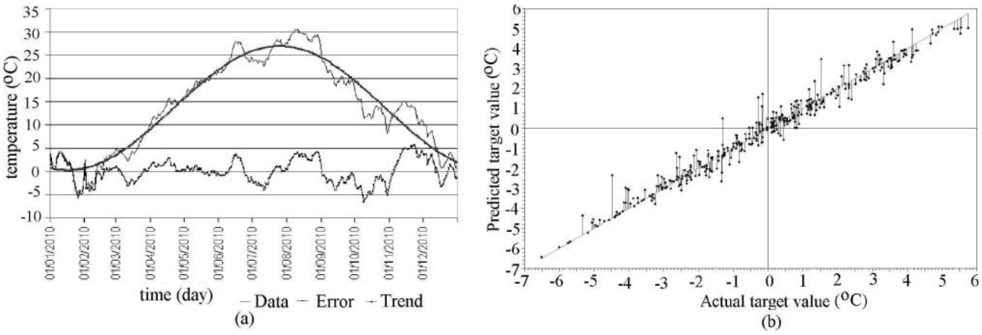

Fig. 6 – a) The series of water temperature, the trend and the error;
b) predicted values of error *versus* actual values of error in GRNN model.

The correlation coefficient between y_t and $\hat{y}_t$ in (1) is equal to 0.9645 and the residual standard deviation of 2.53.

Even if the trend describes well the temperature evolution, we tried to improve the adjustment quality, determining a model for e_t by using the GRNN. After a number of experiments, the number of neurons in the input layer has been set to 6, since the predictors for e_t were the lagged data, $e_{t-1},....,e_{t-6}$. The modeling results are the following: the unexplained variance after the model fit = = 0.24633, the proportion of variance explained by model = 0.96074, the correlation between actual and predicted = 0.9823, the maximum error = 2.88, the mean squared error = 0.2502. Also, testing the hypothesis on the new residual (that in the GRNN model), ε_t, we accepted the hypothesis that it is Gaussian, with the same variance and is not correlated. The chart of the predicted values of e_t *versus* the actual values of e_t in GRNN model is given in Fig. 6b.

Concluding, the model for the water temperature can be written as:

$$y_t = 13.6654 + 13.2957\cos(0.0165t - 3.41393) + g_t + \varepsilon_t, \qquad (2)$$

where g_t is the residual fit and ε_t, the new residual (that in the GRNN model).

3.2. MODEL FOR AIR TEMPERATURE

As in the previous section, the trend of the air temperature has been determined. It has the equation:

$$\hat{x}_t = 12.1946 + 13.1506\cos(0.0165t - 3.3654). \qquad (3)$$

Let us denote by $s_t = x_t - \hat{x}_t$, $t = \overline{1,365}$, the error. The correlation between x_t and $\hat{x}_t$ in (3) is equal to 0.9246 and the residual' standard deviation is 3.77.

The same procedure as for the water temperature modeling has been used. After determining the GRNN model for s_t, the calculated unexplained variance was of 0.24633, the proportion of variance explained by model = 0.96136, the correlation between actual and predicted = 0.9805, the maximum error = 3.78, the mean squared error = 0.246. The data, the trend and the residual are represented in Fig. 7a and the model for the error, in Fig. 7b.

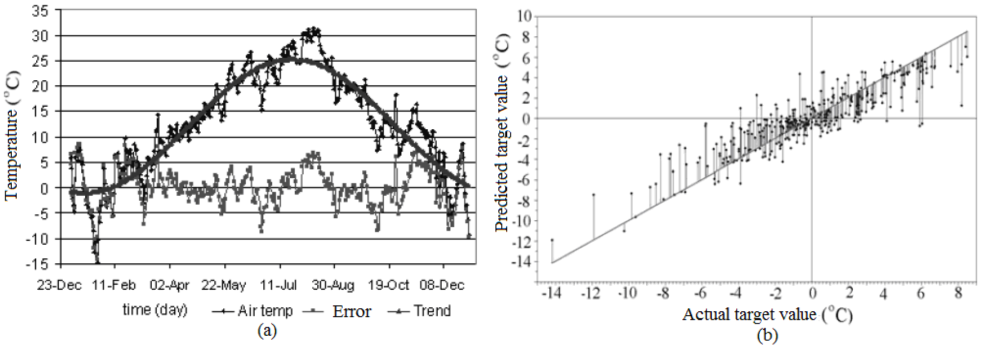

Fig. 7 – a) The series of air temperature, the trend and the error;
b) predicted values of error *versus* actual values of error in GRNN model.

Concluding, the model for the air temperature can be written as (4):

$$y_t = 12.1946 + 13.1506\cos(0.0165t - 3.3754) + h_t + \varepsilon_t', \qquad (4)$$

where h_t is the residual fit and $\varepsilon_t^{'}$, the new residual (that in the GRNN model), that is normally distributed, with the same variance and is not correlated.

3.3. MODEL OF CORRELATION BETWEEN THE WATER AND AIR TEMPERATURE

Starting from the physical fact that between the water and air temperature there is a correlation, the correlation coefficient has been calculated. Since its value is 0.9501, it results that there is a strong linear dependence between the two variables. Therefore, using the least squares method, the linear model given by (5) has been determined:

$$y_t = 2.445x_t + 0.923 + z_t,\ t = \overline{1, 365}, \tag{5}$$

where z_t is the residual (Fig. 8).

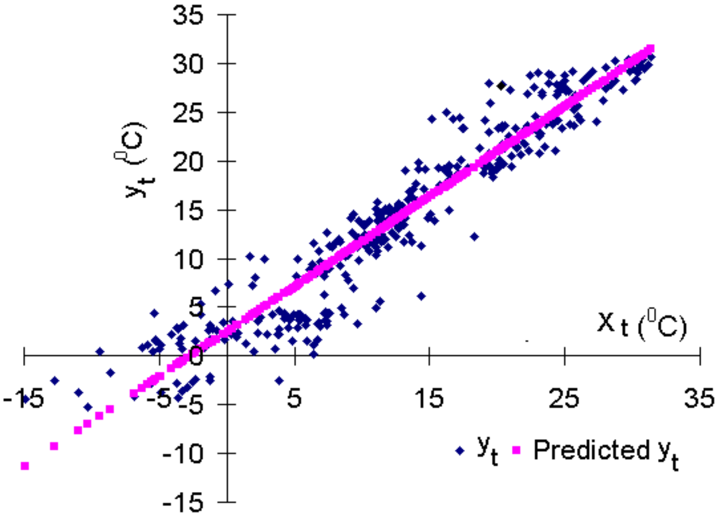

Fig. 8 – Linear model of dependence between the water temperature and air temperature.

In order to validate the model, the “t” test and the “F” test for the coefficients’ significance have been performed, at the significance level of 5%. The p-values associated to these tests were 0.00, so the coefficients and the models are significant. The residual mean was 0, the standard deviation 2.976, and the Kolmogorov-Smirnov test failed to reject the normality hypothesis at the significance level of 1% (the value of the associated statistic being 0.052 and the *p*-value, 0.024). The histogram from Fig. 9 sustains the normality hypothesis.

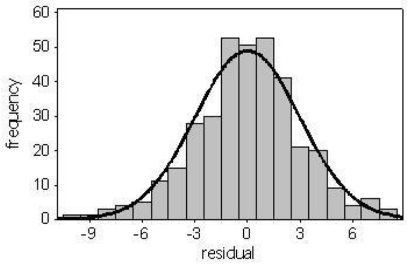

Fig. 9 – Histogram of residual in the linear model.

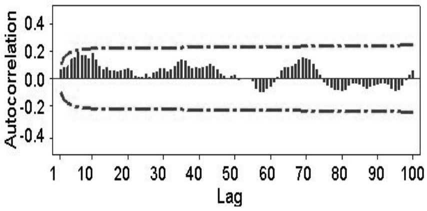

Fig. 10 – The chart of autocorrelation function of residual in the linear model.

The residuals are not correlated, as resulted from the study of the autocorrelation function chart (Fig. 10), where the values of the function (represented by bars) are inside the confidence interval (whose limits are represented by dotted lines), at the confidence level of 95%. Since the residual is not correlated and is normally distributed, it results that it is independent. Finally, the hypothesis of the residual homoscedasticity (the same variance) has been accepted by applying the Levene test, so the linear model proposed is statistically correct.

4. CONCLUSIONS

The analysis of air and water temperatures at Techirghiol Lake in 2011 proved similar statistic properties, in term of normality, correlation and change points and trend existence.

The hybrid models proposed by us for the data series fit well the data, since in both case the correlation coefficients between actual values and predicted ones are close to 1 (0.9823 respectively 0.9805), the proportion of variance explained by the model is also close to 1 (0.96074 respectively 0.96136) and the mean square error is small (0.2502 respectively 0.246). Also, the maximum error was less than 4^{o}C.

It has been proved that there is a linear dependence between the water and air temperature, the determination coefficient in the model being of 0.9501.

Acknowledgements. Thanks to my colleagues C. Maftei and L. Barbes for providing me some useful references concerning Techirghiol Lake.

REFERENCES

1. A. Bărbulescu, L. Barbeş, *Assessment of surface water quality Techirghiol Lake using statistical analysis*, Rev Chim (Bucharest) **64**, *8*, 868–874 (2013).

2. A. Bărbulescu, L. Orac, *Corrosion analysis and models for some composites behavior in saline media*, Int J Energ Environ **1**, *2*, 35–44 (2008).

3. E. M. Carstea, C. I. Ioja, R. Savastru and A. Gavrilidis, *Lake water offers an indication of the health of a city, however little research has been done to evaluate the spatial distribution of dissolved organic matter (DOM), within an urban lake*, Rom Rep Phys **65**, *3*, 1092–1104 (2013).

4. I. D. Dulama, C. Radulescu, C. Stihi, I. V. Popescu, I. Ionita, I. A. Bucurica, E. D. Chelarescu, V.O. Nitescu and R. Stirbescu, *Characterization of Olt river water quality using analytical methods*, Rom Rep Phys **65**, *4*, 1519–1527 (2013).

5. M. Faier, Crivineanu, D. Perju, G. A. Dumitrel and D. Silaghi Perju, *Mathematical models describing the emission and distribution of heavy metals in surface waters*, Rev Chim (Bucharest) **63**, *4*, 435–439 (2012).

6. S. Godeanu, L. D. Galatchi, *The determination of the degree of eutrophication of the Lakes on the Romanian Seaside of the Black Sea*, Ann. Limnol. – Int. J. Lim. **43**, *4*, 245–251 (2007).

7. I. Pincovschi, D. S. Stefan, *Effect of temperature and sulphur dioxide pressure on natural water pollution*, Rev Chim (Bucharest) **63**, *9*, 1021–1023 (2013).

8. M. Constantin, *Therapeutic mud* (in Romanian), Edit. Balneară, Bucharest, 2012.

9. P. P. Gastescu, *The aspects regarding the present status of the Siutghiol and Techirghiol lakes* (in Romanian), Annals of the Bucharest University – Geography Series **3**, 134–138 (2003).

10. Romanescu, Gh., Dinu, C., Radu, A., & Török, L., *Ecologic characterization of the fluviatile limans in the south-west Dobrudja and their economic implications (Romania),* Carpathian Journal of Earth and Environmental Sciences **5**, *2*, 25–38 (2010).

11. C.-E. Telteu, L. Zaharia, *Morphometrical and dynamical features of the South Dobrogea lakes, Romania*, Procedia Env Sci **14**, 164–176 (2012).

12. V. Barnett, T. Lewis, *Outliers in Statistical Data*, John Wiley & Sons, New York, 1994.

13. P. J. Hubert, *The segmentation procedure as a tool for discrete modeling of hydrometeorogical regimes*, Stoch Env Res Risk Ass **14**, 297–304 (2000).

14. A. Gedikli, H. Aksoy, N. E. Unal, *Segmentation algorithm for long time series analysis*, Stoch Environ Res Risk Assess **22**, 291–302 (2008).

15. A. Gedikli, H. Aksoy, N.E. Unal, A. Kehagias, *Modified dynamic programming approach for offline segmentation of long hydro-meteorological time series*, Stoch Environ Res Risk Assess **24**, 547–557 (2010).

16. I. Ben-Gal, Outlier detection, In: O. Maimon, L. Rockach, (Eds.) *Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers*, Kluwer Academic Publishers, 2005, pp. 1–16.

17. D. F. Sprecht, *A General Regression Neural Network*, IEEE Transactions on Neural Networks **2**, *6*, 568–576 (1991).