Ann. Geophysicae 17, 273–279 (1999) © EGS – Springer-Verlag 1999

Annales Geophysicae

Cluster regression model and level fluctuation features of Van Lake, Turkey

Z. Şen, M. Kadıoğlu and E. Batur

Istanbul Technical University, Meteorology Department, Hydrometeorology Research Group, Maslak 80626 Istanbul, Turkey

Received: 2 February 1998 / Revised: 8 June 1998 / Accepted: 25 June 1998

Abstract. Lake water levels change under the influences of natural and/or anthropogenic environmental conditions. Among these influences are the climate change, greenhouse effects and ozone layer depletions which are reflected in the hydrological cycle features over the lake drainage basins. Lake levels are among the most significant hydrological variables that are influenced by different atmospheric and environmental conditions. Consequently, lake level time series in many parts of the world include nonstationarity components such as shifts in the mean value, apparent or hidden periodicities. On the other hand, many lake level modeling techniques have a stationarity assumption. The main purpose of this work is to develop a cluster regression model for dealing with nonstationarity especially in the form of shifting means. The basis of this model is the combination of transition probability and classical regression technique. Both parts of the model are applied to monthly level fluctuations of Lake Van in eastern Turkey. It is observed that the cluster regression procedure does preserve the statistical properties and the transitional probabilities that are indistinguishable from the original data.

Key words. Hydrology (hydrologic budget; stochastic processes) · Meteorology and atmospheric dynamics (ocean-atmosphere interactions)

Introduction

Lakes are natural, inland, free-surface bodies and they respond to atmospheric, meteorologic, geologic, hydrologic and astronomic influences. Hence, the behavior of natural lakes requires knowledge of these driver events recorded within or around the lake catchment. Streamflow data combine various factors influencing the hydraulic balance of a drainage area and, similarly, lake-water level fluctuations represent the end result of the complex interplay of the various water balance components. Among those components are the flow of incoming or outgoing rivers and streams, direct precipitation onto the lake surface and the groundwater exchange. Furthermore, meteorological factors, including precipitation over the lake drainage area, evaporation from the lake surface, wind velocity, humidity and temperature in the adjacent lower atmosphere, all play significant roles in lake water level fluctuations. Some lakes in semi-arid regions are closed with no outlets. Large lakes modify the precipitation over and around the free water surface. Simultaneous measurements of all the effective factors on the lake water level fluctuations are difficult and complete measurements for the application of hydrological water balance equations are missing for many large natural lakes of the world. Perhaps, the simplest lake behavior measurement sequences are the lake water level time series, which include in their structure all the other possible effects combined. It may, therefore, be sufficient to examine and model these fluctuations with the hope of finding simple predictors in the future.

Since, gradual (trend) or abrupt (shifts) climatic change questions have gained particular attention in recent years (ride the preoccupation about impacts of greenhouse gasses on the climate), most of the researches on lake level changes are concerned with meteorological factors of temperature and precipitation data. This point has been assessed in detail by Slivitzky and Mathier (1993). Along this line of research Hubert *et al.* (1989), Vannitsem and Demaree (1991) and Sneyers (1992) used statistical methods to show that temperature, pressure and flow series in Africa and Europe have altered several times during the present century. On the other hand, as stated by Slivitzky and Mathier (1993),

Correspondence to: Z. Şen

most of the modeling of levels and flow series on the Great Lakes have assumed stationarity of the time series using auto regressive-moving average (ARIMA) processes presented by Box and Jenkins (1976). It has been assumed throughout this work that lake level fluctuations do not have a stationarity property and therefore, classical models such as ARIMA processes cannot stimulate lake levels reliably. Multivariate models using monthly lake variabilities failed to adequately reproduce the statistical properties and persistence of basin supplies (Loucks, 1989; Iruine and Eberthardt, 1992).

Spectral analysis of water levels pointed to the possibility of significant trends in lake level hydrological variables (Privalsky, 1990; Kite, 1990). Almost all these scientific studies relied significantly on the presence of an autocorrelation coefficient as an indicator of long term persistence in lake level time series. However, many researchers have shown that shifts in average lake level might introduce unrealistic and spurious autocorrelations. This is the main reason why the classical statistical models often fail to reproduce the statistical properties. However, Mathier *et al.* (1992) were able to reproduce adequately the statistical properties of a shifting-mean model.

In this study, a cluster, linear-regression model has been developed and then used to simulate monthly lake level fluctuations, in a way that preserves the statistical properties and the correlation coefficient. The method is applied to water level fluctuations in Lake Van, eastern Turkey.

Study area features

The world's largest soda lake, Lake Van, is located on the Anatolian high plateau in eastern Turkey (38.5°N and 43°E), (Fig. 1). Lake Van area has very severe winters with frequent temperatures below 0 °C. Most of the precipitation falls during winter season in the form of snow and towards the end of spring heavy rainfalls occur. High runoff rates occur in spring during snowmelt and more than 80% of annual discharge reaches the lake during this period. The summer period, (July to September), is warm and dry with average temperatures of 20 °C. Diurnal temperature variations are about 20 °C. Lake Van has a large drainage basin of 12 500 km^2, (Fig. 2). The lake surface presently averages about 3600 km^2 in extent (Kempe *et al.*, 1978). The surface is approximately 1650 m above sea level. The lake is surrounded by hills and mountains reaching 4000 m. The volcanic mass Süphan mountain rises to 4434 m.

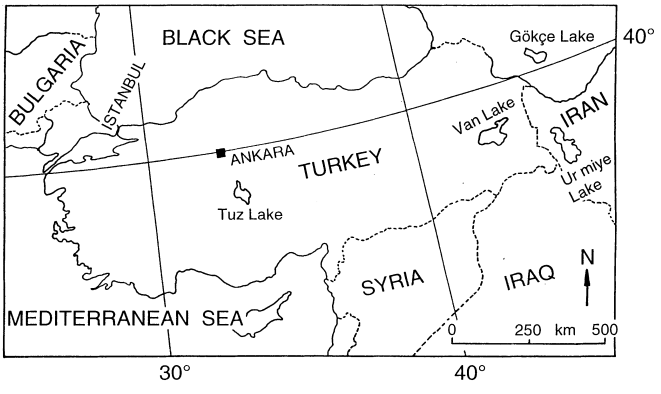

Fig. 1. Location map

Lake Van has no natural outlets. It is the world's fourth closed basin lake, with a volume of about 600 km^3. It is calculated that, on average, annually 4.2 km^3 of water is lost into the atmosphere by evaporation. This is balanced by the long-term averages of annual surface runoff and precipitation amounts as 2.5 km^3 and 1.7 km^3, respectively. Kadıoğlu *et al.* (1997) have shown that the water level fluctuations are entirely dependent on the natural variability of the hydrological cycle and climatic change effects the drainage basin.

Lake Van has been subject to a net water level rise of about 2 m during the last decade and consequently the low-lying, inundated areas along the shore are now giving problems to local administrators, governmental officials, irrigation activities and to people's property. Figure 3 shows monthly lake level fluctuations during the 50 y 1944–1994. Each year, water level rises from January to June and falls in the second part of the year. In this figure, the vertical axis indicates the readings from a staff gauge located at western corner of the lake. These are superimposed on a larger scale of fluctuation. Although the long-range lake level average is at 1648 m, it is now 2 m higher. Under prevailing climatic conditions, the lake water level fluctuations have yearly amplitudes of 40–60 cm. Degens and Kurtman (1978) calculated average amplitude as 49.7 ± 18 cm for period from 1944–1974. However, the complete record from 1944 to 1994 carries average and standard deviation values of 136.8 cm and 70.8 cm, respectively. Comparison of these two periods shows that there are 275% and 383% increases in the mean lake level and amplitude, respectively. Of course, these increases are as a result of lake water level rise after 1974 caused by the climate change in the area. As a result, the region has become more humid with more precipitation but less evaporation.

Whatever the causes might be, there has been a systematic increase in the water level of Lake Van. Level changes will be modeled by means of a simple approach based on the combination of regression line and transition probability methods. The regression analysis is adopted because it furnishes the basis of the short-term persistence through an autocorrelation coefficient and probability, due to its suitability for clustering of points as a result of possible abrupt shifts. These two simple methods are combined together under the name of cluster regression.

Cluster regression model

Classical regression analysis has several assumptions about the normality and independence of the residuals. Furthermore, an implied assumption that skips from the

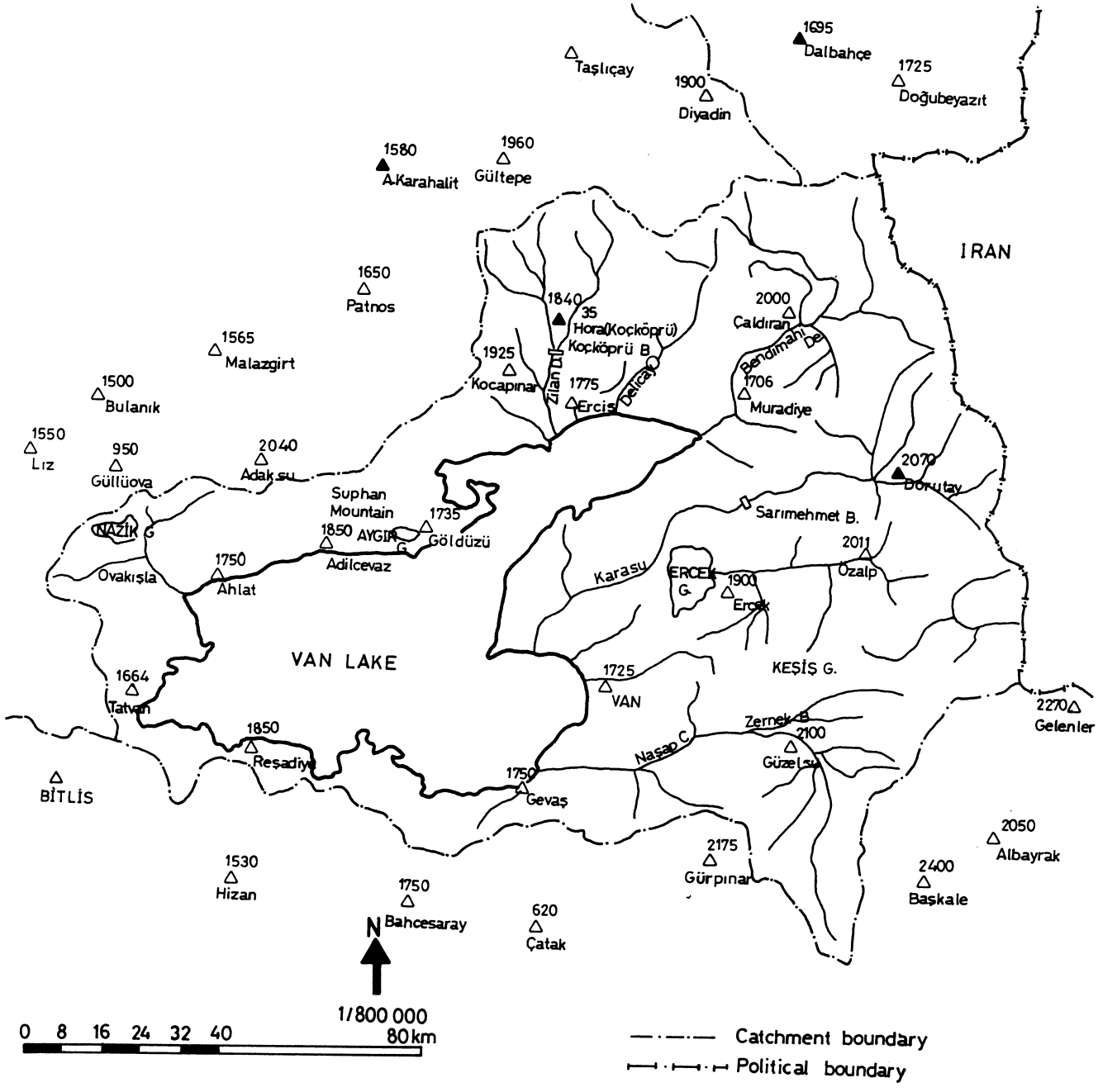

Fig. 2. Lake Van catchment

considerations in most regression line applications is that the scatter diagram should have the points distributed uniformly around a line. Unfortunately, this assumption is often overlooked, especially if the scatter diagram is not plotted. Uniform scatter of the points along the line is possible if the original records are homogeneous and steady with no shifts, trends or seasonalities. If level shifts exist through time then the scatter diagram will include clusters of points along the regression line. Confirmation of such clusters is obvious in Fig. 4, which shows the lake level lag-one scatter diagram for monthly records from Lake Van. The following conclusions are possible from interpretation of the scatter diagram.

a. The lag-one scatter diagram indicates an overall straight-line relationship between the successive lake level occurrences. Existence of such a straight line corresponds to the first order autocorelation coefficient in the monthly lake level time series. Hence, lake level persistence is preserved by this straight line.

b. The scatter of points around the straight line is confined within a narrow band, which implies that the prediction of immediate future levels cannot be very different from the current level provided that there are no shifts in the data.

c. There are different cluster regions along the straight line. Such clusters are not expected in the classical regression approach but the existence of these clusters

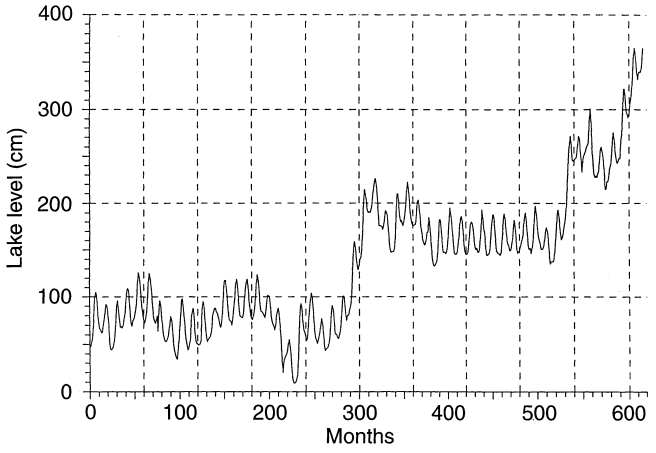

Fig. 3. Lake Van water level fluctuations (1944–1994)

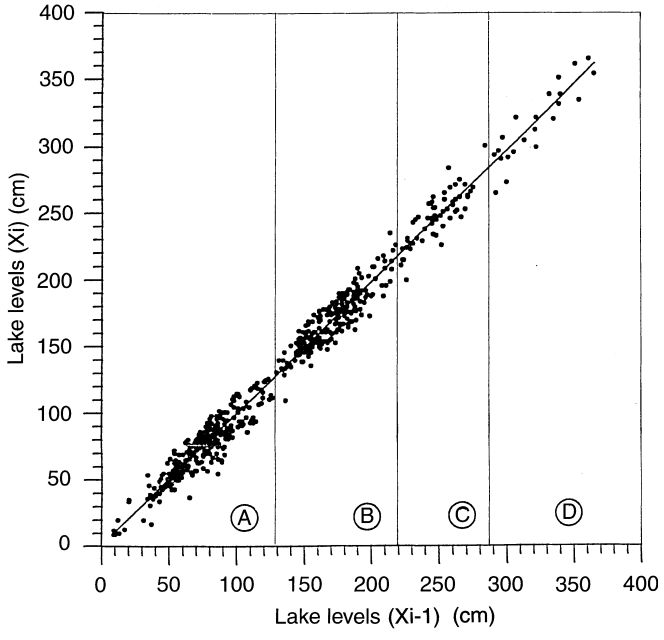

Fig. 4. Lag-one water level fluctuations and cluster boundaries

renders the classical regression analysis into a cluster regression analysis, the basis of which will be presented later in this study. Separate clusters correspond to periods of shifted lake level.

d. Classical regression analysis provides a basis for predicting water levels relative to current, but in the cluster regression line approach, reliable predictions are only possible provided that the probability of cluster occurrences are taken into consideration. Herein, the questions arise as to which cluster is to be taken in future predictions? Should the future prediction remain within the same cluster or not? Any transition from one cluster to another means a shift in the water level. We need, therefore, to know the transitional probabilities among various clusters. The cluster regression depicts not only the autocorrelation coefficient but also the influence domain of each cluster as shown in Fig. 4 along the horizontal axis as A, B, C and D. The influence domains help to calculate the transitional probabilities between the clusters from the original water level records.

e. For any current cluster, it is possible to estimate future normal lake levels by using the regression line equation. For reliable estimations through cluster regression, the following steps are necessary:

1. In order to decide initially which domain of influence (A, B, C or D) should be taken into consideration, a uniform distribution function is considered that assumes random values between 0 and 400 cm.
2. Generate a uniformly distributed random number and, accordingly, decide about the next cluster by considering influence domains. For instance, if the uniformly distributed random number is 272 then from Figure 4 influence domain, C will be the current cluster.
3. Generate another uniformly distributed random number and if the level remains within the same cluster then use the regression equation for estimation. Otherwise, take the average water level value in the new cluster. The new level will be adopted as the midpoint of the cluster domains in Fig. 4. A better estimation might be based on the random variable generation again from a uniform distribution confined within the variation domain of each cluster. Furthermore, the value found in this manner will be added to a random residual value. This will then give the basis of the future water level estimations within the same cluster domain.

Application and discussion

The cluster regression approach has been applied herein to the recorded water level fluctuations of the Lake Van. For this purpose, various lag scatter points of the successive levels are first plotted in Figs. 4–6. The

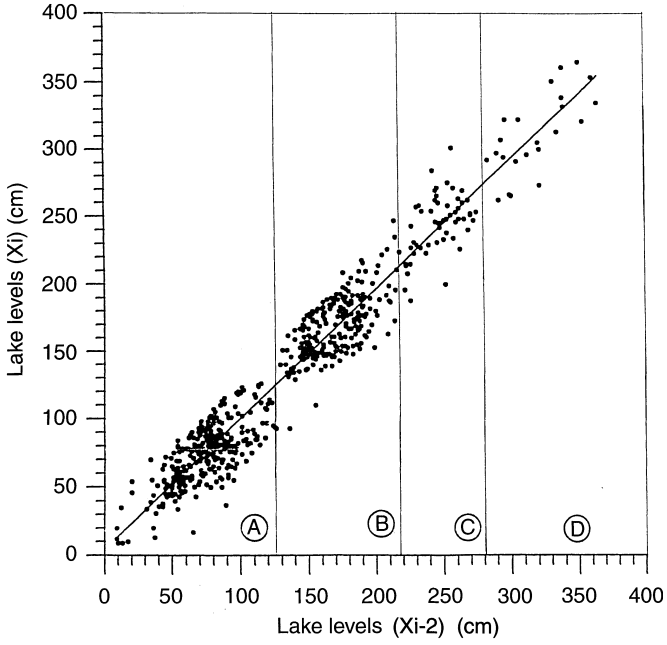

Fig. 5. Lag-two water level fluctuations and cluster boundaries

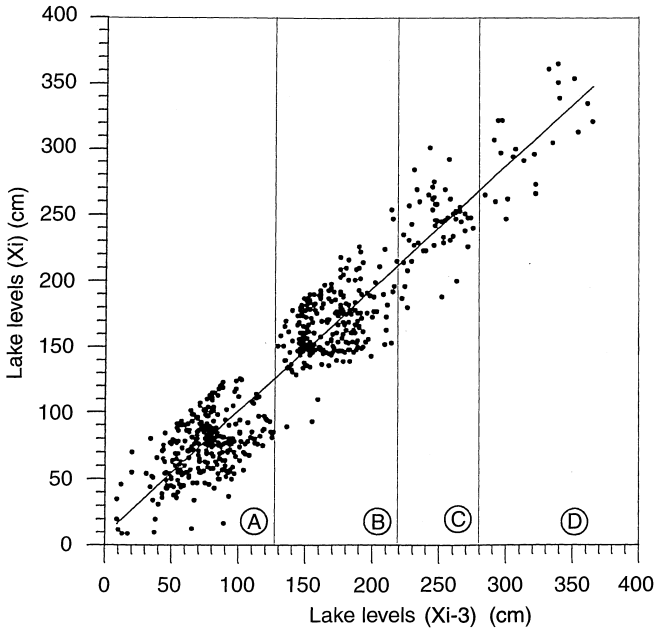

Fig. 6. Lag-three water level fluctuations and cluster boundaries

general appearance of these figures implies the applicability of the cluster regression equation steps as mentioned in the previous section. In all the figures there are straight lines and the transition boundaries between clusters of A, B, C and D are given in Table 1 in addition to the boundaries of each cluster at different lags up to 9.

Here, A is considered as a low lake-level cluster, where only transitions from low level to low level are allowed. B and C refer to lower-medium and upper-medium level clusters and, finally, D is the cluster that includes highest levels only.

It is obvious from Table 1 that the transition limits between A and B, and B and C are practically constant on average for all lags and equal to 129 and 219, respectively. However, the upper limit transition between C-D increases with the increase in the lag value. The difference between the first and ninth lags has a relative error percentage of $100 \times (308-285)/308 = 7.4$ which may be regarded as small for practical purposes.

Table 1. Cluster regression boundaries and coefficients

Lag	Transboundary values			Regression coefficients	
	A-B	B-C	C-D	a	b
1	130	220	> 285	0.985	1.459
2	125	218	> 280	0.960	4.564
3	129	215	> 280	0.930	8.239
4	130	219	> 281	0.901	11.725
5	128	222	> 280	0.878	14.486
6	128	212	> 287	0.862	16.293
7	130	221	> 296	0.853	17.114
8	132	225	> 302	0.852	16.835
9	131	222	> 308	0.858	15.532
Average	129	219			

The scatter diagrams in Figs. 4–6 yield the following specific interpretations for Lake Van level fluctuations.

a. The scatter diagrams have four clusters with the most dense point concentration in cluster A that represents low water level following low water levels. Extreme values of water level fluctuations have the least frequency of occurrences in cluster D.

b. Irrespective of the lag value, points in the scatter diagram deviate from the regression line within a narrow band. This indicates that once the water level is within a certain cluster it will remain within this cluster with comparatively very high probability as will be argued later in this work. Furthermore, the transitions between the clusters are expected to take place rather rarely and in fact between the adjacent clusters only.

c. In none of the scatter diagrams is transition of water level possible from one cluster to another non-adjacent one. This may be confirmed from the calculated transition matrix elements because there are no elements except along the main and the two off diagonals.

Herein, only lag-one regression line will be considered to model the lake levels by considering transitional probabilities between adjacent clusters. The monthly level time series data for lake Van from 1944 to 1994 yield lag-one transition probability matrix, $[M]$ as follows

$$[M] = \begin{array}{c} \\ \text{A} \\ \text{B} \\ \text{C} \\ \text{D} \end{array} \begin{array}{c} \begin{array}{cccc} \text{A} & \text{B} & \text{C} & \text{D} \end{array} \\ \begin{bmatrix} 291 & 2 & 0 & 0 \\ 1 & 233 & 4 & 0 \\ 0 & 3 & 54 & 2 \\ 0 & 0 & 1 & 20 \end{bmatrix} \end{array} \quad (1)$$

The diagonal values in this matrix are the numbers of transitions within each cluster. For instance, there are 291 transitions from low levels to low levels within cluster A. In the same matrix, inter-cluster transitions occur rather rarely along the off diagonals, such as 4 transitions from cluster B to C. In classical stochastic processes the calculation of transition matrix elements are based on the fundamental assumption that the process is time reversible. This is equivalent to saying that transitions as $A \rightarrow B$ is the same as $B \rightarrow A$. Consequently, the resulting matrix must be symmetrical. However, in the proposed method of cluster regression technique only one way transitions along the time axis toward future is allowed. This means that the transition along the time axis is irreversible. As a result of this fact the transition matrix is not symmetrical. Accordingly, the matrix in Eq. (1) is not symmetric, the transition from C to B is not equal to 4 but 3. Zero values next to the off diagonals indicate that the water levels can move only to adjacent clusters. Hence, the possible transitions are ABCD only. For instance, transition to cluster C is possible 4 times from B, 54 times from previous C and only once from D with no transition from A, (hence a total of 59 transitions). Columnar values show transition to the cluster considered from other clusters and the transition probabilities can be calculated after dividing each value in the column by the column total. Hence, the transition probability matrix $[P]$ becomes from Eq. (1) as

$$[P] = \begin{array}{c} \\ \text{A} \\ \text{B} \\ \text{C} \\ \text{D} \end{array} \begin{array}{c} \begin{array}{cccc} \text{A} & \text{B} & \text{C} & \text{D} \end{array} \\ \begin{bmatrix} 0.9932 & 0.0068 & 0 & 0 \\ 0.0042 & 0.9790 & 0.0168 & 0 \\ 0 & 0.0508 & 0.9152 & 0.0339 \\ 0 & 0 & 0.0476 & 0.9524 \end{bmatrix} \end{array} \tag{2}$$

The linear regression line that relates two successive water levels, namely, W_i and W_{i-1}, can be obtained from the cluster scatter diagram in Fig. 4 as

$$W_i = 0.9858 W_{i-1} + 1.45918 + \varepsilon_i \tag{3}$$

in which ε_i signifies the vertical random deviations from the regression line. Theoretically, these random deviations should have a Gaussian distribution function for the validity of the regression line and Fig. 7 indicates that they are normally distributed. In order to adopt Eq. (3) estimations with the cluster scatters, it is essential to take into account the following steps:

a. Because the most frequently occurring water levels are confined in cluster A, the initial state W_0 is selected randomly from the actual water levels in this cluster.

b. Decision whether there is transition to the next cluster is achieved through the transition probabilities given in matrix $[P]$ in Eq. (2). The transitions occur according to the following rules.

1. Transition to cluster A is possible only from cluster B or the level remains within the same cluster. From the transition matrix these have probabilities as 0.9932 and 0.0042 and their summation is equal to 1.0. In order to decide which one of these two clusters will be effective in the next time step, it is necessary to generate a uniform random number, ξ_i, which varies between zero and one. If $\xi_i < 0.9932$ then the water level will remain within cluster A, otherwise for $0.9932 < \xi_i < 1.0$ a transition occurs from cluster A to B. In the former case, after generating a normally distributed random number, ε_i, the new water level value is generated by the use of the clusteral regression model in Eq. (3). However, in the latter case, water level will be selected randomly from the range of water levels for cluster, B.

2. At any instant, transition to cluster B may take place from two adjacent clusters (A or C). The transitional probabilities from A and C are 0.0068 and 0.0508, respectively, with complementary probability of 0.9790 remaining within cluster B. Now the decision of transition to B will have three independent regions of the uniform distribution, namely, if $0 < \xi_i < 0.0068$ then a transition occurs from A to B or when $0.0068 < \xi_i < 0.9858$ water level remains within cluster B and finally, for $0.9858 < \xi_i < 1.0$ a transition occurs from C to B. If the water level remains within cluster B, a normal variate is generated as ε_i and the regression expression in Eq. (3) is used to predict the next water level. In the transition cases, water level is depicted randomly from the available levels.

c. Transitions to cluster C show a similar mechanism to cluster B with different transition probabilities but the same generating mechanism.

d. Finally, transition to cluster D is possible only from cluster C, in addition to remaining in the same cluster.

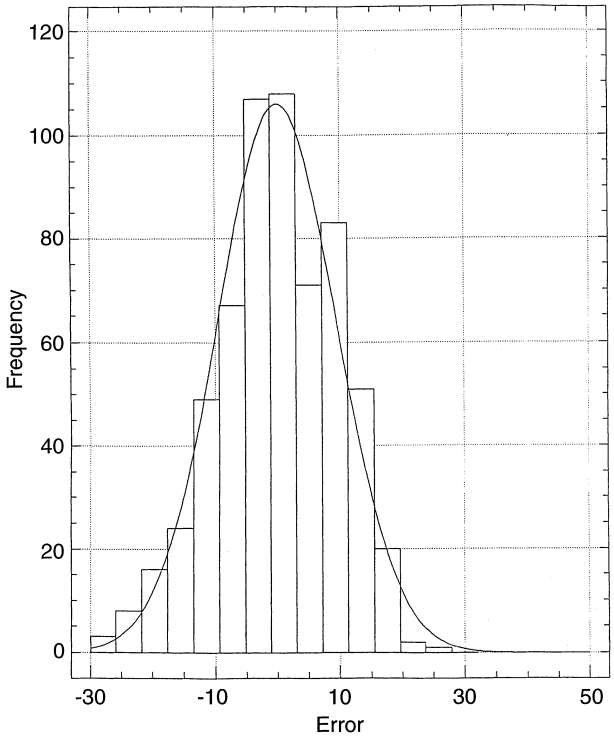

Fig. 7. Regression error distribution function

The application of all these procedures and steps to Lake Van monthly water level variations result in the development of the synthetic transition matrix $[M_s]$

$$[M_s] = \begin{array}{c} \\ \text{A} \\ \text{B} \\ \text{C} \\ \text{D} \end{array} \begin{array}{c} \begin{array}{cccc} \text{A} & \text{B} & \text{C} & \text{D} \end{array} \\ \begin{bmatrix} 287 & 2 & 0 & 0 \\ 1 & 230 & 3 & 0 \\ 0 & 4 & 56 & 2 \\ 0 & 0 & 1 & 19 \end{bmatrix} \end{array} . \tag{4}$$

Comparison of corresponding elements between the two matrices in Eqs. (1) and (4) show that they differ by less than 5% relative error. This indicates that the preservation of transition numbers as probabilities in the predicted lake levels are indistinguishable from the actual water level data. The synthetic cluster scatter diagram obtained from the use of Eqs. (2) and (3) is shown in Fig. 8 where the regression line has the form as

$$W_i = 0.978 W_{i-1} + 1.47 + \varepsilon_i \tag{5}$$

Again comparison of this expression with Eq. (3) shows that the corresponding coefficients vary by less than 5% relative error. In other words, the autocorrelation coefficient in the prediction of water levels is preserved in spite of shifts in the original data.

Conclusion

The basis of a new regression equation with clusters are presented with an application to the water level fluctuations of Lake Van, eastern Turkey. The clusteral regression method provides the best regression line in addition to the cluster occurrences and transition

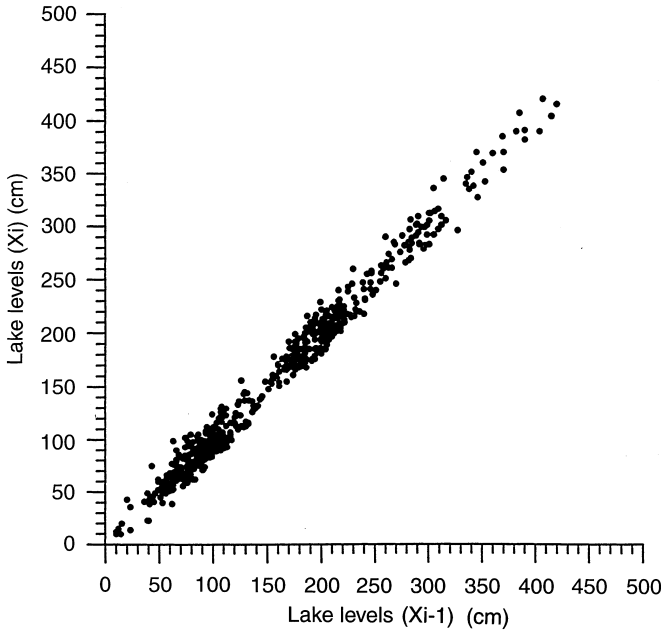

Fig. 8. Synthetic lag-one water level fluctuations

probabilities along this line. Its difference from the classical regression approach lies in the appearance of nonoverlapping clusters. The cluster regression approach preserves all the statistical parameters in addition to the autocorrelation coefficient, which is a measure of short-term persistence in lake level records. Any shifts in the data do not lead to spurious and unrealistic autocorrelation.

Acknowledgements. Topical Editor D. J. Webb thanks P. Challenor and another referee for their help in evaluating this paper.

References

Batur, E., Lake Van water budget and catchment climate, Unpublished MSc Thesis, Istanbul Technical University, Meteorology Department, 124 pp, (in Turkish), 1996.

Box, G. E. P., and G. M. Jenkins, *Time series analysis forecasting and control*, Holden Day, San Francisco, 560 pp, 1976.

Degens, E. T., and F. Kurtman, The geology of Lake Van. *The Mineral Research and Exploration Institute of Turkey,* Rep. 169, 158 pp, 1978.

Hubert, P., J. D. Carbomnel, and A. Chaouche, Segmentation des series hydrometeoraloques: application a des series de precipitation et de debits de L'Afrique de L'ouest. *J Hydrol.,* **110,** 349–367, 1989.

Iruine, K. N., and A. K. Eberthardt, Multiplicative seasonal ARIMA models for lake Erie and Lake Ontario water levels, *Water Reso. Bull.* **28** (3), 385–396, 1992.

Kadıoğlu, M., Z. Sen, and E. Batur, The greatest soda-water lake in the world and how it is influenced by climatic change, *Ann. Geophysicae,* **15,** 1489–1497, 1997.

Kempe, S., F. Khoo, and Y. Gürleyik, Hydrography of lake Van and its drainage area, pp. 30–45, *in the Geology of Lake Van,* Eds. E. T. Degen and F. Kurtman, The Mineral Research and Exploration Institute of Turkey, Rep. 169, 1978.

Kite, V., Time series analysis of Lake Erie levels. In 'Proc Great Lakes Water Level Forecasting and Statistics Symposium', Eds. H. C. Hartmann and M. J. Donalhue, Great Lake Commission, Ann Arbor, Michigan, pp. 265–277, 1990.

Loucks, E. D., Modeling the Great Lakes hydrologic-hydraulic system. Ph D Thesis, University of Wisconsin, Madison, 1989.

Mathier, L., L. Fagherazzi, J. C. Rasson, and B. Bobee, Great Lakes net basin supply simulation by a stochastic approach. *INRS-Eau Rapp Scientifique* **362,** INRS-Eau, Sainte-Foy, 95 pp, 1992.

Privalsky, V., Statistical analysis and predictability of Lake Erie water level variations, in 'Proc Great Lakes Water Level Forecasting and Statistics Symposium', Eds. H. C. Hartmann and M. J. Donalhue, Great Lake Commission, Ann Arbor, Michigan, pp 255–264, 1990.

Sneyers, R., On the use of statistical analysis for the objective determination of climate change, *Meteorol Z.* **1,** 247–256 pp, 1992.

Slivitzky, M., and L. Mathier, Climatic changes during the 20th century on the Laurentian Great Lakes and their impacts on hydrologic regime, *NATO Adv. Study Inst.,* Deauaille, France, 1993.

Vannitsem, S., and G. Demaree, Detection et modelisation des secheresses an Sahel-proposition d'une nouvelle methodologie, *Hydrol Continent* **6,** (2) 155–171 pp, 1991.