J. Great Lakes Res. 18(1):236–243
Internat. Assoc. Great Lakes Res., 1992

STATISTICAL ANALYSIS AND PREDICTABILITY OF LAKE ERIE WATER LEVEL VARIATIONS

Victor Privalsky
Water Problems Institute
USSR Academy of Sciences
Sadovaya-Chernogryazskaya, 13/3
Moscow K-64, USSR

ABSTRACT. Statistical predictability and spectra of mean monthly and annual water levels (MMWL and MAWL) of Lake Erie at Cleveland, Ohio, 1860–1988, are studied within the framework of the Kolmogorov-Wiener theory of extrapolation using AR modeling and the theory of non-stationary product random processes in order to assess the attainable quality of least-squares predictions of water levels. MMWL are shown to possess relatively high predictability due to a strong seasonal cycle in water level variations, with predictability limits extending up to 12 months. MAWL reveal a time-dependent structure in the mean value, variance, and spectrum which can be ascribed, among other reasons, to a climatic change. Their predictability is quite low (predictability limit not more than 1 or 2 years) and cannot be improved by applying other techniques of scalar time series extrapolation. The uncertainties in water level predictions should be taken into account quantitatively when making decisions which depend upon hydrological parameters.
INDEX WORDS: Lake levels, statistical predictability, Box-Jenkins modeling, non-stationary prediction problem.

INTRODUCTION

The purpose of this study is to statistically analyze the basic properties of Lake Erie water level variations as represented by the time series of monthly water level heights at Cleveland, Ohio, from 1860 to 1988. The properties to be studied include spectra and statistical predictability parameters such as the relative prediction error and predictability limit at time steps of 1 month and 1 year. The knowledge of these properties is essential to better understand the mechanisms of water level variations and to improve the efficiency of decision-making process related to changes in water levels. The techniques of time series analysis used in this study lie mostly within the framework of Kolmogorov-Wiener's theory of extrapolation and Box and Jenkins' autoregressive modeling (including maximum entropy spectral analysis).

In accordance with the above-stated goals of the study, this article consists of two parts which contain the results of studying mean monthly and annual water levels.

TABLE 1. Major statistical parameters of mean monthly water levels, m.

Parameter	1860–1988	1860–1924	1925–1988
Linear trend*	.011	-.049	.113
Trend's rms error*	.0019	.0037	.0049
Mean value	173.92	173.91	173.94
RMS	.332/.329	.260/.226	.396/.305

*(Meters per month) × 100

MORE ABOUT THE INITIAL DATA

The time series of monthly water level (MMWL) heights of Lake Erie at Cleveland, Ohio, is shown in Figure 1. It is easily seen that the time series is not stationary with respect to the mean value; specifically, it can be assumed that each half contains statistically significant linear trends (see Table 1), probably related to changes in the inflow to the lake through the Detroit River. However, the analysis of such trends lies outside the scope of the study at its present stage. Therefore, the linear trend(s) will be removed so that the remaining

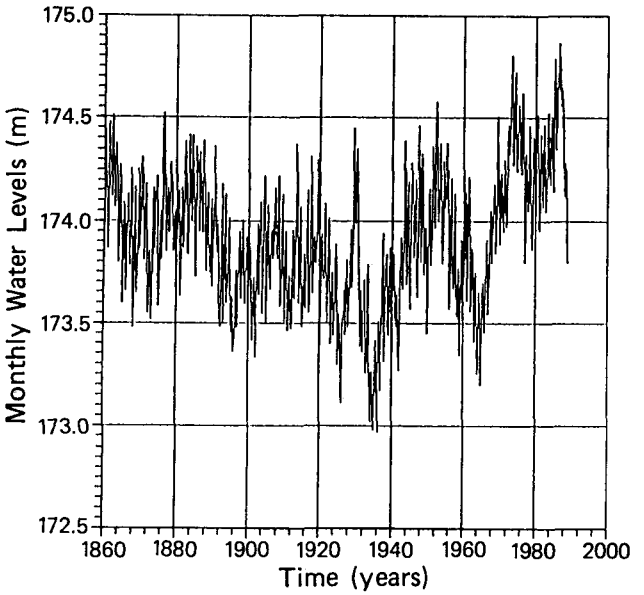

FIG. 1. Lake Erie mean monthly water levels, Cleveland, Ohio, 1860–1988.

part(s) of the time series will be regarded as stationary with respect to the mean value. Table 1, also contains some other statistical parameters of the time series.

Thus, the root mean square (RMS) estimates, after deleting the linear trends from the two parts of the time series (shown in the denominators), suggest that the time series as a whole cannot be regarded as stationary. In fact, the hypothesis of stationarity should be rejected at a significance level of 0.01 according to the variance ratio test.

Other statistical properties of the time series such as its correlation function or spectral density and predictability parameters have also undergone a change from the first to the second half of the time series, though less pronounced.

With this in mind, we will first study the properties of the second half of the time series regarded as a sample of a stationary random process by analyzing its properties at relatively small time scales (up to about 1 year).

FORECASTING MONTHLY WATER LEVELS

Probability Distribution Function

In the present context, the information about the probability distribution function (PDF) of the time series is necessary in order to (a) estimate the applicability of linear Gaussian models, and (b) calculate the confidence bounds for the extrapolation function of water levels. When the seasonal trend is removed, MMWL have a probability distribution which is close to Gaussian. Though the presence of the trend will definitely affect this result, we will still assume, for our problem, that its PDF is close to Gaussian. This assumption will not be valid for solving problems which depend heavily upon the choice of a probability distribution function.

It should also be noted that the second half of the time series, from 1925 to 1988, is not stationary because its properties are also time-dependent. However, the changes are relatively slow and, as we are interested in the small time scale (monthly) properties, this phenomenon will be ignored at this stage.

Spectral Density and Statistical Predictability

The linear parametric autoregressive-moving average (ARMA) models introduced by Box and Jenkins (1976) are widely used to describe hydrological time series in the time domain (Salas *et al* 1980). However, the maximum entropy approach in spectral analysis which leads to autoregressive (AR) approximations to the time series is practically unknown in hydrology. It is not used in Bras and Rodriguez-Iturbe (1985), for example. We will summarize some of its properties below (Jaynes 1982).

1. The approach allows one to choose a stationary random process for a given Gaussian time series to which this series belongs with the highest possible probability.
2. It is most effective in studying short time series, which is usually the case in hydrology and climatology.
3. The maximum entropy spectral estimation leads to an immediate and simple solution of the time series least-squares linear prediction problem within the framework of Kolmogorov-Wiener theory.
4. As the number of AR parameters which describe the time series' properties is usually small, the resulting estimates of its spectrum, correlation function, and predictability parameters are relatively reliable statistically.

Note that these properties hold for multi-variate time series as well, and thus the approach is effective in studying the relations between different time series (Privalsky 1988a).

Consider first the estimate of water levels spec-

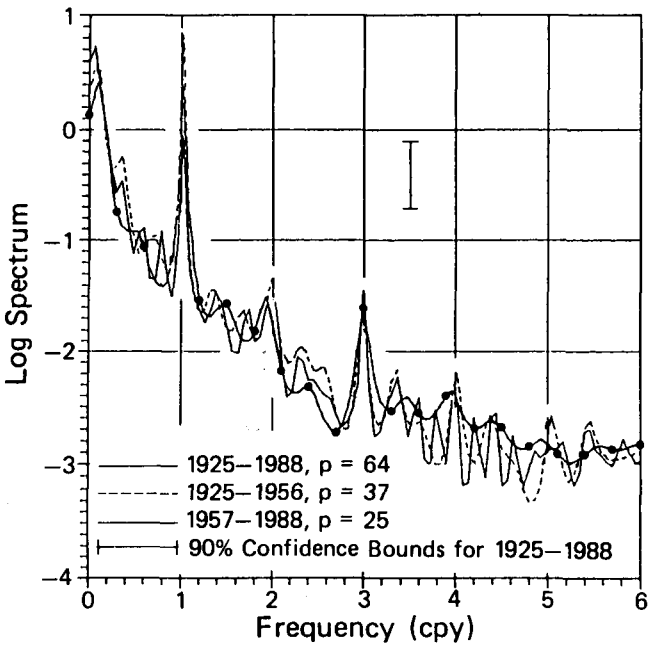

FIG. 2. Spectrum of mean monthly water levels at Cleveland, Ohio, 1925–1988.

trum obtained by the maximum entropy technique. The best AR model according to four criteria including Akaike's AIC proved to be the model of order $p = 64$ [i.e., AR(64)]. Its spectrum is shown in Figure 2. The most important features of the spectrum seem to be its fast decrease with frequency, the presence of seasonal trend, and its harmonics. This means that the predictability of the time series will be rather high at small lead times due to the strong seasonal trend, while the long-range prediction will depend upon the presence of high-energy low-frequency components rather than upon any "cycles." Note that the changes of the spectrum from 1925–1956 to 1957–1988 are not large.

Several measures of statistical predictability can be introduced, the most important one being the relative prediction error, or RPE, $d(\tau) = D(\tau)/D(\infty)$, where $D(\tau)$ is the mean square prediction error at lead time τ. In the stationary case, $D(\infty)$ coincides with the time series variance σ^2. Thus, $d(\tau)$ is a monotonically non-decreasing function of τ satisfying the inequality $0 \leq d(\tau) \leq 1$. The quality of prediction decreases with growing $d(\tau)$ and, therefore, another measure of predictability is given by the lead time τ_γ which corresponds to a given value of $d(\tau) \sim 1$. Predictions at higher lead

TABLE 2. Parameters of seasonal multiplicative SAR(2,5) model.

Parameter	Estimate	Standard Error
ϕ_1	1.248	.035
ϕ_2	-.292	.035
ϕ_1	.168	.036
ϕ_2	.201	.036
ϕ_3	.164	.036
ϕ_4	.090	.037
ϕ_5	.195	.036
MEAN	173.80 m	.10 m
θ	1.40 m	
σ_a	.067 m	

times, $\tau > \tau_\gamma$, are regarded as too inaccurate. The choice of τ_γ depends upon the specific problem at hand. It should play an important role in decision-making. This value is called the limit of statistical predictability (Privalsky 1983) or predictability horizon (Parzen and Newton 1984). The correlation coefficient $\rho(\tau)$ between the actual and predicted values of water levels is: $\rho(\tau) = [1 - d(\tau)]^{1/2}$.

As the time series of monthly water levels contains a strong seasonal trend, the class of AR models to be fitted to it should contain a seasonal operator. We chose to approximate the time series with seasonal multiplicative AR models because the moving average operator may cause computational instability. The best model of this type for monthly levels, x_t, proved to be SAR(2,5). The first digit in the parentheses is the order of the non-seasonal AR operator and the second is the seasonal one with period $s = 12$ months:

$$\left[1 - \sum_{j=1}^{2} \phi_j B^j\right] \left[1 - \sum_{j=1}^{5} \Phi_j B^{js}\right] x_t = \theta + a_t. \quad (1)$$

The model's parameters are shown in Table 2 where $\sigma_a^2 = D(1)$ is the variance of the innovation sequence a_t which coincides with the 1-month mean square prediction error.

The predictability measures are shown in Figure 3. As seen from the figure, RPE $d(1)$ is rather small for both models and remains below the 0.5–0.6 level up to lead time $\tau = 12$ months. Respective values of correlation coefficient are $\rho(1) \cong .97$ while $\rho(12)$ stays between 0.6 and 0.7. These predictability properties are about the same for the AR(64) model.

If, as is usually done in geophysics, we assume $\gamma = 0.9$, the limit of statistical predictability will

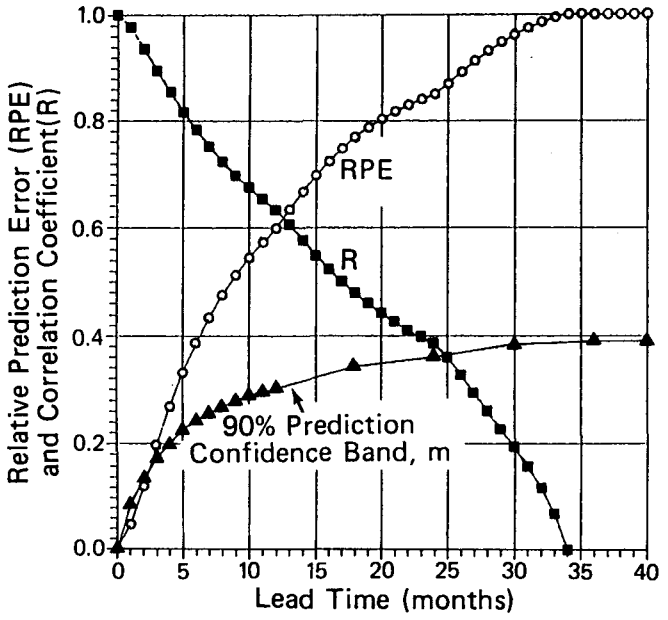

FIG. 3. Relative prediction error (RPE) and correlation coefficient (R) between the actual and predicted mean monthly water levels of Lake Erie at Cleveland, Ohio, 1925–1988, model SAR(2, 5).

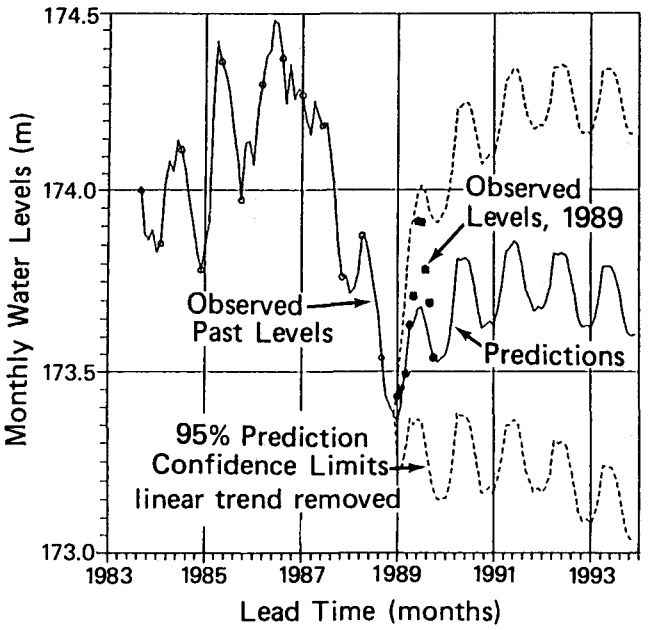

FIG. 4. Forecasting Lake Erie mean monthly water levels at Cleveland, Ohio, 1925–1988, model SAR(2, 5).

extend to 26 months and over 40 months for SAR(2, 5) and AR(64) models, respectively. Thus, monthly water levels of Lake Erie possess relatively high statistical predictability properties.

Several remarks should be made at this point. First, due to the estimation errors, the actual predictability will be lower (Box and Jenkins 1976). Consequently, the AR(64) model which contains 64 coefficients against 7 coefficients for the SAR(2, 5) model should probably be rejected. Second, the absolute $100(1 - \alpha)\%$ prediction error

$$\epsilon_\alpha(\tau) = \pm \mu_{\alpha/2}{}^{d}(\tau)\sigma \qquad (2)$$

where α, the significance level, is actually rather large. For $\alpha = 0.1$, it equals $+/- 0.1$ and $+/- 0.3$ m at lead times $\tau = 1$ and $\tau = 12$ months, respectively (Fig. 3). Finally, it should be remembered that no other linear (or non-linear, assuming Gaussianity) predictions based upon the past values of this time series can lead to smaller prediction errors. An example of actual predictions shown in Figure 4 confirms that MMWL at Cleveland can be predicted rather accurately up to lead times $\tau = 10$ months.

In summing up the results of the study at this stage, it can be said that:

(a) as long as we are interested in short-term (several months) statistical properties and predictions of Lake Erie water levels at Cleveland, Ohio, it seems advisable to use only the second part of the time series;

(b) within the framework of stationary seasonal models, the time series is best approximated by a multiplicative seasonal AR model SAR(2, 5) plus a linear trend;

(c) complete and rigorous solution of respective prediction problems within the framework of the Kolmogorov-Wiener theory reveals relatively high statistical predictability of this process. Respective confidence intervals should serve as a basis for decisions concerning engineering problems related to water level variations.

FORECASTING MEAN ANNUAL WATER LEVELS

Even after the piecewise linear trend is removed from the time series of mean annual water levels (MAWL), 1860 to 1988 (solid curve in Fig. 5), the resulting random sequence cannot be regarded as stationary because its variance is obviously time-dependent. This phenomenon is related to changes in the lake's water budget constituents which may be caused, among other factors, by a climatic

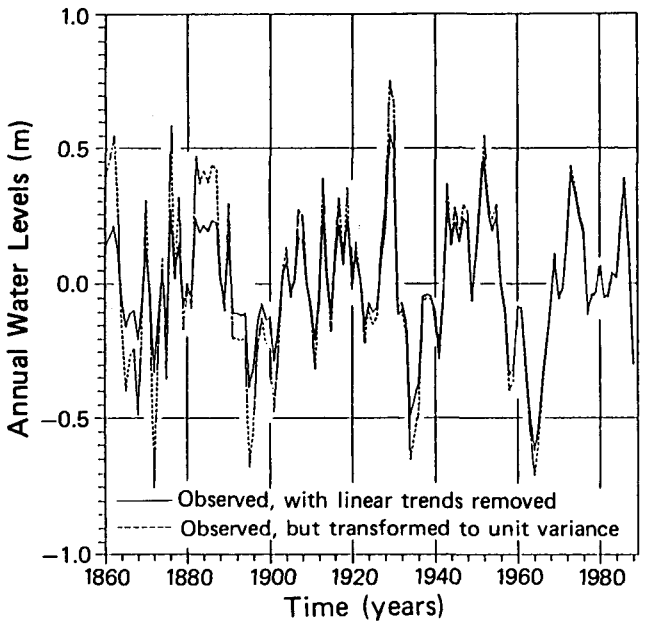

FIG. 5. Lake Erie mean annual water levels, Cleveland, Ohio, 1860–1988. Linear trends are removed.

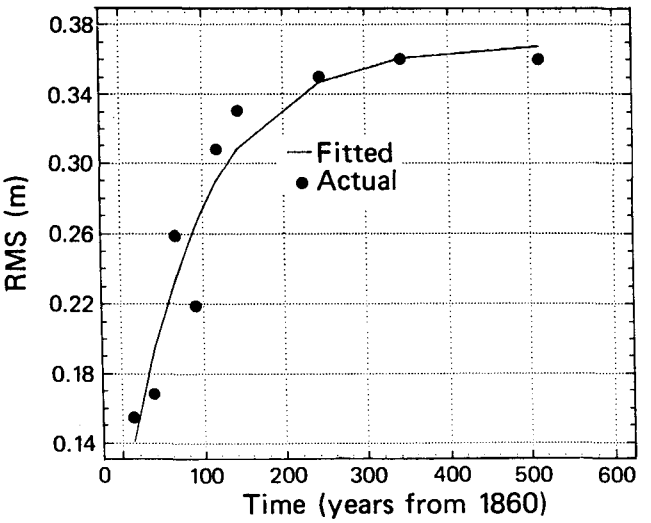

FIG. 6. Plot of fitted model for RMS of MAWL of Lake Erie at Cleveland, Ohio.

change. As long as we deal with water level data exclusively, no physically reasonable explanation for this change of properties can be suggested. Leaving aside possible reasons for this change, we can and should study this time series as a sample of a non-stationary random process. In order to do this, the nature of its non-stationarity must be defined.

Whatever the reason for the change of the process' structure can be, its variance can hardly increase indefinitely. It is reasonable to assume that we are now in a transition period from one climatic regime to another; consequently, the new stationary regime will eventually be reached at some distant time in the future. However, as we will be interested in time scales which do not exceed a decade, the parameters of the new regime which cannot be deduced from the time series of water levels are not important for our problem. It seems safe to assume that we are dealing with a product non-stationary process of the form

$$x_t = \alpha(t)u_t \tag{3}$$

where u_t is a stationary random process with zero mean and unit variance, $\alpha(t)$ a given "slow" function of time. As in the previous discussion, the latter can be written in the form

$$\alpha(t) = A - Be^{-Ct}, \quad t \geqq 0, \tag{4}$$

so that the variance of water levels equals $(A - B)^2$ and A^2, as t tends to zero and infinity, respectively.

Parameters A, B, and C can be estimated through short-term (for example, over five 25-year intervals) estimates of water level variance. The results of this procedure are shown in Figure 6. The values of respective parameters are: A = .37 (+/- .02), m, B = 0.26 (+/- .03) m, and C = .010 (+/- .003) year^{-1} with the determination coefficient, $R^2 \cong .92$. (RMS of respective estimates are given in the parentheses.) It should be noticed here that though the value (A - B) can be chosen on the basis of the existing water level observations, the value A which characterizes the water level variance in a distant future has been chosen arbitrarily as we have no information about the effect of climatic change upon water levels. An estimate of A can be obtained by further developing the approach used by Croley (1990) to evaluate the impact of a climatic change upon the hydrological regime of the Great Lakes Basin and then applying the techniques developed by Privalsky (1988b).

Dividing the time series of annual water levels by $\alpha(t)$, one arrives at the time series of variance-stationary process, μ_t, (dashed curve in Fig. 5), which can then be analyzed statistically. The best AR approximation to this series is

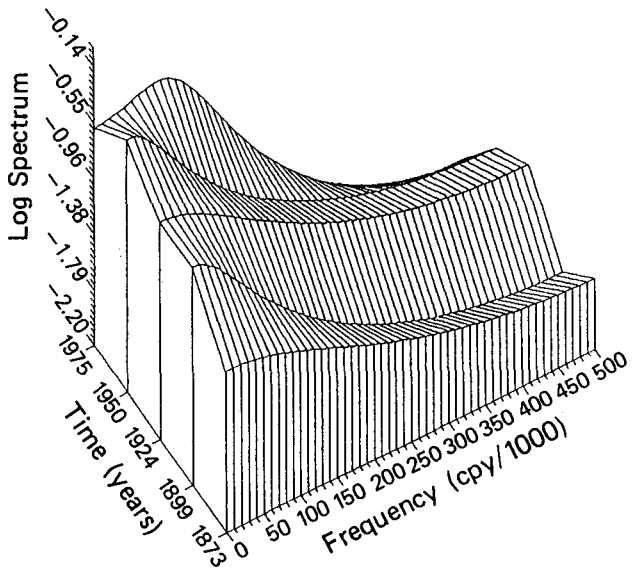

FIG. 7. Time-dependent spectrum of MAWL of Lake Erie at Cleveland, Ohio.

$$\mu_t - \phi_1\mu_{t-1} = \epsilon_t \tag{5}$$

where $\phi = 0.61$, ϵ_t is a sequence of identically distributed and mutually independent random variables with zero mean and variance $\alpha^{2\epsilon} = 1 - \phi_{12}$. Now the properties of non-stationary model (3) with $\alpha(t)$ and μ_t given by equations (4) and (5) can be studied analytically through respective non-stationary spectra (Bendat and Piersol 1986). Keeping the purpose of this research in mind, we will only discuss the estimated time-dependent spectrum of MAWL and the predictability problem.

As seen from Figure 7, the evolution of MAWL spectrum with time is not very prominent and can probably be explained just by the sampling variability of individual estimates. (This refers, in particular, to a broad peak at $f \cong 0.2$ cpy in the spectrum estimate for the last 25 years.) Generally, the spectrum is rather flat and the energy is decreasing with growing frequency. This means that the statistical predictability of MAWL will be rather small even for the unit lead time $\tau = 1$ year. Its quantitative measure can be obtained in the following manner. Substituting $t + \tau$ for t and rewriting equation (5) in the form

$$\mu_t = (1 - \phi_1 B)^{-1}\epsilon_t \tag{6}$$

where B is the backward shift operator, then substituting equations (4) and (6) into equation (3) leads, after some simple algebra, to the following expression for future water levels $x_{t+\tau}$:

$$x_{t+\tau} = [A - Be^{-C(t+\tau)}](\epsilon_{t+\tau} + \phi_1\epsilon_{t+\tau-1} + \phi_1^2\epsilon_{t+\tau-2} + \ldots) =$$

$$= [A - Be^{-C(t+\tau)}] \sum_{j=0}^{\infty} \psi_j\epsilon_{t+\tau-j} \tag{7}$$

where $\psi_j = \phi_1^j$. As ϵ_t is a zero mean white noise sequence, its least-squares predictions, $\epsilon_t(\tau) = 0$ for $\tau > 0$, while previous values ϵ_{t-k} for $t \leq k$ can be calculated from previous predictions and observations (Box and Jenkins 1976). Therefore, least-squares predictions $x_t(\tau)$ of x_t at lead times τ are

$$x_t(\tau) = [A - Be^{-C(t+\tau)}] \sum_{j=\tau}^{\infty} \psi_j\epsilon_{t+\tau-j} \tag{8}$$

Subtracting eq. (8) from eq. (7) gives prediction error $\delta_t(\tau)$ at time t and lead time τ:

$$\delta_t(\tau) = [A - Be^{-C(t+\tau)}] \sum_{j=0}^{\tau-1} \psi_j\epsilon_{t+\tau-j} \tag{9}$$

so that the prediction mean square error

$$D_t(\tau) = <\delta_t^2(t)> = [A - Be^{-C(t+\tau)}]^2 \sum_{j=0}^{\tau-1} \psi_j^2 \tag{10}$$

with the angle brackets meaning ensemble averaging. Remembering that $\psi_{j=\phi}1^j$ and $\sigma_\epsilon^2 = (1 - \phi_1^2)$, one obtains the following expression for the mean square prediction error:

$$D_t(\tau) = [A - Be^{-C(t+\tau)}]^2(1 - \phi_1^{2\tau}). \tag{11}$$

Now, the prediction algorithm in this non-stationary case will contain the following steps:

(a) calculate predicted annual water levels $x_t(\tau)$ by predicting $u_{t+\tau}$ as a first order AR process, multiply the result by $\alpha(t + \tau)$, and add the linear trend;

(b) compute mean square prediction error $D_t(\tau)$ and respective confidence limits, $x_t(\tau) = x_t(\tau) \pm \mu_{\alpha/2}D_t^{1/2}(\tau)$.

As seen from Figure 8, statistical predictability of annual water levels with the two linear trends deleted is indeed small for the next decade. Both

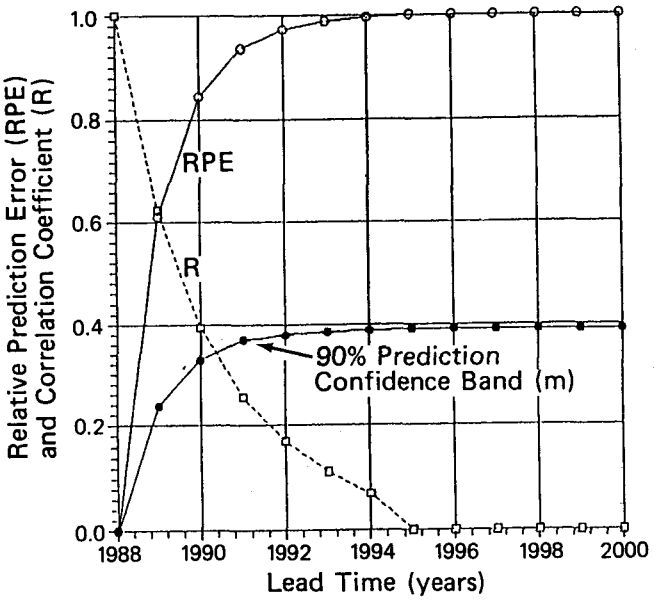

FIG. 8. Relative prediction error (RPE) and correlation coefficient (R) between the actual and predicted mean annual water levels of Lake Erie at Cleveland, Ohio, 1925–1988, non-stationary product model.

RPE and correlation coefficient at 1 year lead time equals about 0.6, while the limit of statistical predictability is reached in 2 years, by 1991.

CONCLUSIONS

1. Both mean monthly and annual water revel variations of Lake Erie at Cleveland, Ohio, from 1860 to 1988, cannot be regarded as a sample of a stationary random process. The non-stationarity is caused by the presence of a strong seasonal trend in mean monthly levels as well as by the time-dependent mean value and variance. Thus, the properties of the time series, including its statistical predictability, as well as its prediction problem, should be studied within the framework of respective non-stationary approximations.

2. Taking into account the fact that the time series is relatively short, its properties should be studied, whenever possible, by using the techniques of time series analysis which are designed for dealing with short time series in both the time and frequency domains such as AR-modeling and maximum entropy spectral analysis. (This is true only when there is no strong evidence of a non-linear or long-memory behavior of the time series.)

3. The part of the MMWL time series between 1925 and 1988 can be treated, as a first approximation, as a sample of a periodically-correlated random process. The best stationary approximation to this time series seems to be a seasonal multiplicative AR sequence SAR(2, 5) which contains two nonseasonal and five seasonal coefficients and possesses relatively high statistical predictability caused by the presence of a strong seasonal trend. The limit of statistical predictability for MMWL is reached in about 2 years. As the unit lead time τ = 1 month, the correlation coefficient $\rho(1)$ between the actual and predicted MMWL exceeds 0.9 while the absolute value of respective 90% confidence bound amounts to about +/- 0.1 m. For τ = 6 and τ = 12 months, these values change to 0.8 and +/- 0.2 m and to 0.6 and 0.3 m. Respective confidence bands and probabilities must be taken into account in the decision-making process.

4. The sequence of MAWL from 1860 to 1988 can be regarded as a mixture of a piecewise linear trend and a non-stationary product random process with the stationary part represented by an AR model of order 1. The deterministic product function which describes the non-stationary part of the process can be given, as a first approximation, in the form of a logistic curve. Its physical interpretation may include the transition between two stationary states which corresponds to two different stationary climates.

5. When this structure is assumed for MAWL, its statistical predictability for the next decade proves to be low: correlation coefficient between the actual and predicted levels in 1989 is about 0.6 with the 90% confidence band of about +/- 0.2 m. The limit of statistical predictability for MAWL is reached in 2 or, at most, 3 years. This means that any engineering decision which is related to predictions of mean annual levels will necessarily contain a large degree of uncertainty. This uncertainty can hardly be diminished by any other means but predicting the changes in the elements of the lake's water budget, first of all the inflow through the Detroit River.

6. The next problem to study seems to be modeling water level variations in Lake Erie as a process created by random inputs to the dynamic system defined by the lake's water budget equation (Privalsky 1988b). Reliable predictions of inflow to the lake at small lead times can hardly be anticipated in the near future. Nevertheless, it seems possible to predict probable changes in water level statistical properties due to a given climate change which will be based upon the solution of stochastic

dynamic water budget equation and respective estimates of water budget changes in the Great Lakes basin.

REFERENCES

Bendat, J. S., and Pierson, A. G. 1986. *Randon Data. Analysis and Measurement Procedures*. New York: John Wiley and Sons.

Box, G. E. P., and Jenkins, G. M. 1976. *Time Series Analysis: Forecasting and Control*. San Francisco: Holden-Day (Revised Edition).

Bras, R. E., and Rodriguez-Iturbe, I. 1985. *Stochastic Functions and Hydrology*. Reading, Massachusetts: Addison-Wesley.

Croley, T. E. II. 1990. Laurentian Great Lakes double CO_2 climate change hydrological impacts. *Climate Change* 17:27–47.

Jaynes, E. T. 1982, On logical foundations of the maximum entropy methods. *IEEE Trans.* 70:33–51.

Parzen, E., and Newton, J. 1984. Forecasting and time series model types of Ill economic time series. In *The Forecasting Accuracy of Major Time Series Methods*, ed. S. Makridakis *et al.*, pp. 267–288. New York: Wiley and Sons.

Privalsky, V. 1983. Statistical predictability and spectra of air temperature over the northern hemisphere. *Tellus* 35A:51–59.

_____ . 1988a. Stochastic models and spectra of interannual variability of mean annual sea surface temperature in the North Atlantic. *Dynamics of Atmospheres and Oceans* 12:1–18.

_____ . 1988b. Modeling long term lake variations by physically based stochastic dynamic models. *Stochastic Hydrology and Hydraulics* 2:303–315.

Salas, J. D., Delleur, J. W., Yevievich, V., and Lane, W. I. 1980. *Applied Modeling of Hydrologic Time Series*. Chelsea, Mich.: Water Res. Publ.

Submitted: 29 January 1991
Accepted: 5 December 1991