LEAD TIME VS. ACCURACY IN HYDROLOGY FORECASTS

By

I. KONTUR

Department of Water Management, Institute of Water Management and Hydraulic Engineering, Technical University, Budapest

Presented by Prof. Dr. I. V. NAGY

(Received: December 15, 1981)

1. Introduction

The development of water management and the rise of demands on water resources have been accompanied by the improvement of methods and tools of hydrological forecasts. The development of forecasting methods was much forwarded by the advent of computers, and the arise of automatic measuring and data collecting networks was a jump forward in the field of data collection and transfer. Unified forecasting systems covering water networks connected territories and even countries to solve hydrology problems. In Hungary, a decisive part of systematic hydrological forecasts is made by the *Institute of Hydrography* (*Scientific Research Centre of Water Management* VITUKI).

Hydrology forecasts are expected, among others, to involve not only a preset lead time (time advantage) but different lead times keeping various fields of water management (flood control, navigation, water utilization etc.) in mind. Accuracy — confidence — of simultaneous forecasts with different lead times will of course be not the same, and according to observations, with increasing lead time forecast accuracy will decrease, a question to be insisted on in the following. Hydrology forecasts have always some water management purpose and a wide range of uses, with different lead times and economy impacts of measure proposals relying on these forecasts. Water management decisions relying on forecasts will only be correct if accessible to economy ponderations. In uncertain surroundings — such as hydrology processes — evaluability is based on the indication of the rate of uncertainty, the forecast error — difference between the real and the forecast value. The concerned processes being random, stochastic ones, the standard deviation of forecast errors will be the statistic characteristic of the rate of uncertainty. Process z_t has been plotted in Fig. 1. A forecast issued at time t at a lead time l will be $\hat{z}_t(l)$. The difference between this latter and the real value z_{t+l} is the forecast error:

$$e_t(l) = z_{t+l} - \hat{z}_t(l)\,. \tag{1}$$

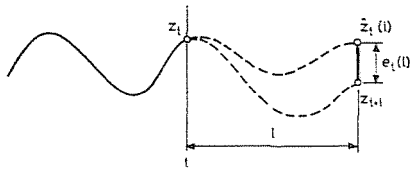

Fig. 1. Interpretation of forecast lead time and forecast error

The expected value of the square of forecast errors is the variance:

$$\sigma_e^2(l) = \underset{t}{E}\{e_t^2(l)\}, \tag{2}$$

where $\underset{t}{E}\{\cdot\}$ is the symbol of expected value formation. The rate of forecasting accuracy or uncertainty will be expressed by the variance according to (2) (or its square root, the standard deviation). (Remind that the accuracy decreases, while the forecast uncertainty increases to the sense, with increasing $\sigma_e(l)$.)

Knowledge of the $\sigma_e(l)$ value permits to trace an arbitrary strip of confidence around the forecast $\hat{z}_t(l)$:

$$P\{\hat{z}_t(l) - u_{\varepsilon/2} \cdot \sigma_e(l) < z_{t+l} < \hat{z}_t(l) + u_{\varepsilon/2} \cdot \sigma_e(l)\} = 1 - \varepsilon. \tag{3}$$

Thus, the probability of the real value to lie between the given limits is exactly $1-\varepsilon$. $u_{\varepsilon/2}$ in Eq. (3) is the value of the normal distribution function at $\varepsilon/2$. (Errors may generally be assumed to be of normal distribution, else the normal distribution function has to be replaced by another distribution function to the sense.)

Here only statistic models will be analyzed, leaving forecast lead time vs. accuracy problems of hydraulic or so-called physical models out of consideration. The excellent book on forecasting by BOX and JENKINS [1] has been relied on: examples on *ARMA* models are hydrologic applications in this country. Lead time-accuracy relationship of the general linear regression models arose from the extension of the conventional regression calculus. Finally, let us notice that the solution of the problems to be presented has largely been facilitated by the questions raised by *Dr. András Szöllősi-Nagy* in the domain of forecasting lead time accuracy.

2. Lead time — accuracy relationship in ARMA models

*A*utoregressive *m*oving *a*verage *(ARMA)* models are known to often well suit description of hydrographs:

$$\begin{aligned} z_t = \varphi_1 z_{t-1} + \varphi_2 z_{t-2} + \ldots + \varphi_p z_{t-p} + a_t - \\ - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \ldots - \theta_q a_{t-q}, \end{aligned} \tag{4}$$

where

$z_t, z_{t-1} \ldots$ — hydrograph elements;

$a_t, a_{t-1} \ldots$ — independent, random *Gaussian* process (white noise) elements;

$\varphi_1, \varphi_2, \ldots, \varphi_p$ and $\theta_1, \theta_2, \ldots, \theta_q$ — hydrograph parameters.

The model according to (4) is called $ARMA(p, q)$ since it contains p autoregressive terms and q moving average terms.

Application of the so-called backward shift operator $\mathfrak{B}$:

$$z_{t-1} = \mathfrak{B}z_t,\ z_{t-2} = \mathfrak{B}^2 z_t,\ \ldots\ z_{t-k} = \mathfrak{B}^k z_t,$$

transforms (4) to:

$$\varphi(\mathfrak{B})z_t = \theta(\mathfrak{B})a_t \tag{4a}$$

where $\varphi(\mathfrak{B})$ and $\Theta(\mathfrak{B})$ are polynomials of operator $\mathfrak{B}$:

$$\varphi(\mathfrak{B}) = 1 - \varphi_1\mathfrak{B} - \varphi_2\mathfrak{B}^2 - \ldots - \varphi_p\mathfrak{B}^p$$

and

$$\theta(\mathfrak{B}) = 1 - \theta_1\mathfrak{B} - \theta_2\mathfrak{B}^2 - \ldots - \theta_q\mathfrak{B}^q.$$

z_t may be written as infinite-termed sum of random pulses a_t:

$$z_t = a_t + \psi_1 a_{t-1} + \psi_2 a_{t-2} + \ldots =$$

$$= a_t + \sum_{j=1}^{\infty} \psi_j a_{t-j} = (1 + \sum_{j=1}^{\infty} \psi_j \mathfrak{B}^j)a_t = \psi(\mathfrak{B})a_t, \tag{5}$$

where

$\psi_0 = 1, \psi_1, \psi_2, \ldots$ — weights of the "white noise" process;

$\psi(\mathfrak{B})$ — transfer function of the linear filter relating z_t to a_t.

Confrontation of (4a) and (5) shows an unambiguous relationship for determining weights

$$\varphi(\mathfrak{B}) \cdot \psi(\mathfrak{B}) = \theta(\mathfrak{B}), \tag{6}$$

of importance in what follows.

To determine the error variance $\sigma_e^2(l)$ of forecasts with different lead times, let z_{t+l} be written in form (5):

$$z_{t+l} = (a_{t+l} + \psi_1 a_{t+l+1} + \ldots + \psi_{l-1} a_{t+1}) + (\psi_l a_t + \psi_{l+1} a_{t-1} + \ldots) =$$

$$= e_t(l) + \hat{z}_t(l), \tag{7}$$

that is, the sum of the first l terms is exactly the forecasting error at time t,

of a lead time l. The expected value of the forecasting error is zero, since also the process a_t has been assumed the same; furthermore:

$$\sigma_e^2(l) = \underset{t}{E} \{ e_t^2(l) \} = (1 + \psi_1^2 + \psi_2^2 + \ldots + \psi_{l-1}^2) \cdot \sigma_a^2, \tag{8}$$

utilizing that set a_t is a "white noise" process. As a matter of fact, Eq. (8) is the wanted relationship.

Determination of the error variance of the forecast of lead time l is seen to need the first $l - 1$ elements of the infinite set $\psi_1, \psi_2, \ldots$. The ψ values may be obtained from (6), or, in particular:

$$\begin{aligned}(1 - \varphi_1 \mathfrak{B} - \varphi_2 \mathfrak{B}^2 - \ldots - \varphi_p \mathfrak{B}^p)(1 + \psi_1 \mathfrak{B} + \psi_2 \mathfrak{B}^2 + \ldots) = \\ = 1 - \theta_1 \mathfrak{B} - \theta_2 \mathfrak{B}^2 - \ldots - \theta_q \mathfrak{B}^q.\end{aligned} \tag{6a}$$

Equating coefficients of operators $\mathfrak{B}$ with identical exponents yields algebraic equations for determining:

$$\begin{aligned} \psi_0 &= 1 && \rightarrow \psi_0 = 1 \\ -\varphi_1 \mathfrak{B} + \psi_1 \mathfrak{B} &= -\theta_1 \mathfrak{B} && \rightarrow \psi_1 = -\theta_1 + \varphi_1 \\ -\varphi_1 \psi_1 \mathfrak{B}^2 - \varphi_2 \mathfrak{B}^2 + \psi_2 \mathfrak{B}^2 &= -\theta_2 \mathfrak{B}^2 && \rightarrow \psi_2 = -\theta_2 + \varphi_2 + \varphi_1(-\theta_1 + \varphi_1) \\ &\vdots \\ &\text{etc.} \end{aligned} \tag{9}$$

In general, recursive formulae may be written for ψ:

$$\begin{aligned} \psi_0 &= 1 \\ \psi_1 &= \varphi_1 - \theta_1 \\ \psi_2 &= \varphi_1 \psi_1 + \varphi_2 \psi_0 - \theta_2 \\ \psi_3 &= \varphi_1 \psi_2 + \varphi_2 \psi_1 + \varphi_3 \psi_0 - \theta_3 \\ &\vdots \\ \psi_j &= \varphi_1 \psi_{j-1} + \varphi_2 \psi_{j-2} + \ldots + \varphi_p \psi_{j-p} - \theta_j \\ &\vdots \\ &\text{etc.} \end{aligned} \tag{10}$$

For $\mathbf{j} > q$, $\theta_j = 0$.

Figure 2 is the scheme of recursive calculation, separately for cases $p > q$ and $p < q$.

Calculation of ψ is seen to need only parameters φ and θ of the $ARMA$ model.

It should be noticed that it is a pure moving average model, for $MA(q)$, the set of ψ is finite and $\psi_j \equiv \theta$, thus $\psi_{q+1} = \psi_{q+2} = \ldots = 0$.

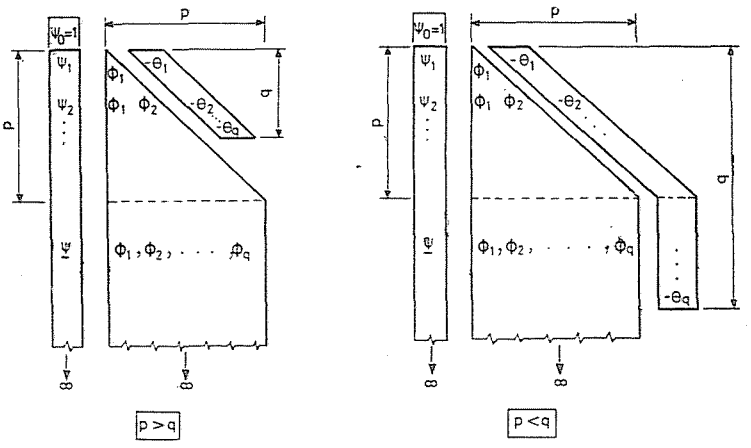

Fig. 2. Scheme of recursive calculation of coefficients for $p < q$ and $p > q$

At the same time, for $l \geq q$, $\sigma_e(l)$ equals the standard deviation σ_z of the process.

3. Lead time – accuracy relationship for a general linear regression model

In the practice of hydrology forecasts, it is often insufficient to apply a single *ARMA* model for describing the processes, since the examined phenomenon (e.g. stage) is influenced by several, interdependent phenomena (e.g. stages at upstream gauging stations, precipitation etc.). Hydrology practice mostly applies linear regression models, at a generally close approximation of reality, suiting rough estimation even for very nonlinear phenomena.

The general regression model of lead time l is of the form:

$$z_{t+l} = b_1(l)x_{1,t} + b_2(l)x_{2,t} + \ldots + b_n(l)x_{n,t} + e_t(l), \tag{11}$$

where $x_{1,t}, x_{2,t}, \ldots, x_{n,t}$ are variables involved in forecasting that may include target variable values z preceding time t (e.g. $x_{1,t} = z_{t-1}, x_{2,t} = z_{t-2}, \ldots$ etc.) so that the model includes also autoregressive terms. Including terms of the "white noise" process among independent variables (e.g. $x_{n,t} = a_t, x_{n-1,t} = a_{t-1}, \ldots$ etc.) the general linear regression model contains also a moving average. In the special case where the general linear regression model contains only the quoted autoregressive and moving average terms, (11) tends to the *ARMA* model in (4).

Estimation of parameters of a general linear regressive model is obtained by minimizing the square sum of deviations $e_t(l)$:

$$\hat{\mathbf{b}}(l) = \mathbf{R}_{xx}^{-1} \cdot \mathbf{r}_{xz}^{*}(l) \tag{12}$$

where

$\underset{(n,1)}{\hat{\mathbf{b}}(l)}$ — vector of estimated parameters $\hat{b}_1(l) \ldots, \hat{b}_n(l)$;

$\underset{(n,n)}{\mathbf{R}_{xx}}$ — correlation matrix determined from hydrographs $x_{1,t}, \ldots, x_{n,t}$;

$\underset{(n,1)}{\mathbf{r}^*_{xz}(l)}$ — vector formed of so-called distorted correlations $\frac{\sigma_z}{\sigma_{x_j}} r_{x_j z}(l)$;

σ_z — standard deviation of process z_t;

σ_{x_j} — standard deviation of process $x_{j,t}$ $(j = 1, 2, \ldots, n)$;

$r_{x_j z}(l)$ — correlation between $x_{j,t}$ and z_{t+l}.

Now, the expected value of the variance becomes:

$$\underset{t}{E}\{e_t^2(l)\} = \sigma_e^2(l) = \sigma_z^2\big(1 - \mathbf{r}'_{xz}(l)\,\mathbf{R}_{xx}^{-1}\mathbf{r}_{xz}(l)\big), \tag{13}$$

$\mathbf{r}_{xz}(l)$ being a vector formed of correlations $r_{x_j z}(l)$ $(n, 1)$.
Introducing notation:

$$\varrho(l) = \mathbf{r}'_{xz}(l) \cdot \mathbf{R}_{xx}^{-1} \cdot \mathbf{r}_{xz}(l), \tag{14}$$

formula of the forecast accuracy becomes:

$$\sigma_e^2(l) = \sigma_z^2\big(1 - \varrho(l)\big). \tag{13a}$$

Thus, it is sufficient to examine the development of $\varrho(l)$;
if $l \to \infty$, $\varrho(l) \to 0$, hence variance of the forecast error tends to the process variance, quite harmonizing with the practical approach.

Special cases of the development of $\varrho(l)$ have been analyzed in [2, 3].

4. Examples and applications

Calculation of the lead time vs. accuracy relationship will be presented on two sets of data. One is the forecast of daily stages in the *Felsőberecki* section of *Bodrog* river. The other data set comprises the monthly mean discharges at the *Szeged* section of the *Tisza* river. A yearly set of the Bodrog river stages ($N = 365$) has been examined. The Tisza set contained ten years of deviations of monthly discharges from long-time averages ($N = 120$). Autocorrelation functions of Bodrog stages and of Tisza discharges have been plotted in Figs 3a and b, respectively.

Three examples — $AR(1)$, $AR(2)$, and $ARMA$ (1, 1) — will be presented for the application of $ARMA$ models, (No pure moving average model suits these hydrographs: one- and two-step autocorrelations r_1 and r_2 are outside the range of pure moving average models.) Parameters φ_1; φ_1, φ_2; and φ_1, θ_1 of

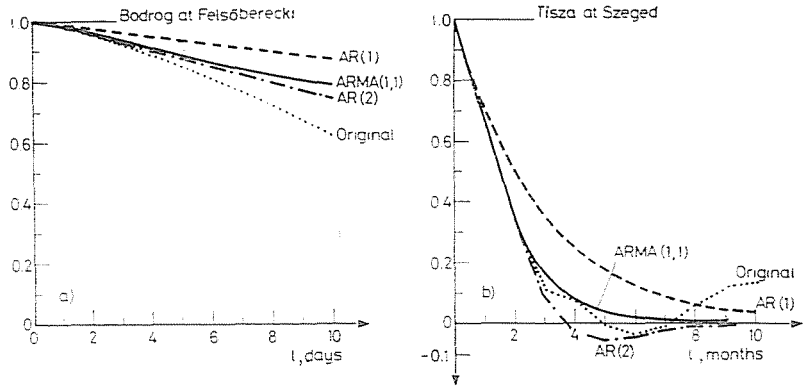

Fig. 3. Autocorrelation functions *a)* for daily stages of the *Bodrog* at Felsőberecki; *b)* for monthly mean discharges of the *Tisza* at Szeged

the three kinds of models are seen in *Table* 1 for both hydrographs. So are values of autocorrelation factors r_1 and r_2, standard deviations σ_z of the hydrographs, as well as calculated values of σ_a.

Table 1

Model	Bodrog—Felsőberecki $r_1 = 0.986$, $r_2 = 0.962$, $\sigma_z = 145$ [cm]				Tisza—Szeged $r_1 = 0.7065$, $r_2 = 0.3349$, $\sigma_z = 250 \left[\frac{m^3}{s}\right]$			
	φ_1	φ_2	θ_1	σ_a	φ_1	φ_2	θ_1	σ_a
AR(1)	0.986	— —	—	24.178	0.7065	—	—	176.928
AR(2)	1.34757	—0.3667	—	22.492	0.9381	—0.32828	—	167.178
ARMA(1.1)	0.9756	—	—0.450	22.070	0.474	—	—0.545	163.037

Identified models $AR(1)$, $AR(2)$, and $ARMA(1, 1)$ were relied on in calculating theoretical autocorrelation functions, also seen in Figs 3a and 3b. Confrontation with the autocorrelation functions obtained from the original hydrograph shows model autocorrelation functions to be approximative, and obviously, only one- or two-step autocorrelation functions to coincide.

On the basis of model parameters (φ, θ), recursion formula (10) was applied to calculate the ψ_j value, followed by determining the $\sigma_e^2(l)$ value for all the six models (three for Bodrog, and three for Tisza) according to Eq. (8).

In view of the different hydrographs examined, while results of forecast lead time vs. accuracy relationships had to be confronted, rather than the standard deviation or variance of deviations, their ratio to the process variance σ_z^2:

$$\sigma_e^2(l)/\sigma_z^2$$

was examined.

Fig. 4 shows the trend of $\sigma_\varepsilon^2(l)/\sigma_z^2$ during $l = 1, 2, 3, \ldots, 13$ days for the three models referring to the *Felsőberecki* section of *Bodrog* river. Although there is often a slight difference between models, the quality order where $\sigma_\varepsilon^2(l)/\sigma_z^2$ is the least, is for $l = 1$ day: $ARMA$, $AR(2)$, $AR(1)$; for $l = 2, 3, \ldots,$ 10 days: $AR(1)$, $ARMA$, $AR(2)$, and for $l < 10$ days: $ARMA$, $AR(2)$, $AR(1)$. By the way, if the forecast accuracy is to be rated by the correlation index, then, on the basis of $\sigma_\varepsilon^2(l)/\sigma_z^2$:

$$R = \sqrt{1 - \frac{\sigma_\varepsilon^2(l)}{\sigma_z^2}}. \tag{15}$$

Also square values of the correlation indices have been indicated in Fig. 4.

$\sigma_\varepsilon^2(l)/\sigma_z^2$ values calculated for models $AR(1)$, $AR(2)$ and $ARMA$ (1, 1) fitted to monthly mean discharges of the *Tisza* river have been plotted in Fig. 5. With increasing lead times, the forecast accuracy is seen to steeply

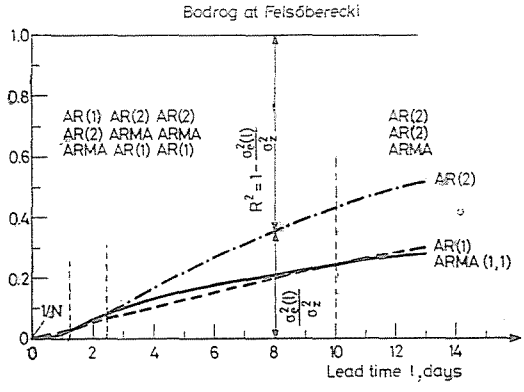

Fig. 4. Lead time to uncertainty relations for different models referring to *Bodrog* stages at Felsőberecki

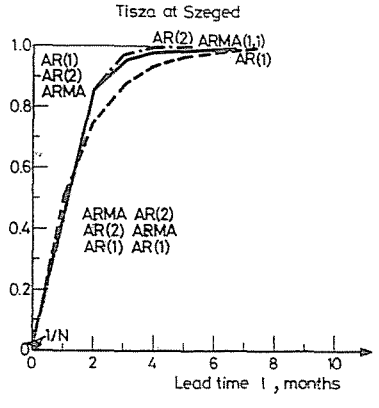

Fig. 5. Lead time to uncertainty relations for different models referring to monthly mean discharges of the *Tisza* at Szeged

decrease, in conformity with expectations from the confrontation of autocorrelation functions of both processes.

Again, the quality order of forecast models varies with the lead time; for $l = 1$ month: $ARMA$, $AR(2)$, $AR(1)$; for $l = 2$ months: $AR(1)$, $AR(2)$, $ARMA$; $l = 2, 3, \ldots,$ months: $AR(1)$, $ARMA$, $AR(2)$.

Nevertheless, the forecast model contains almost no information excess in the case of four to five months of lead time.

Finally, let us consider the application of the general linear regression model for different lead times, referring to daily stage data at the *Felsőberecki* section of the *Bodrog* river. Forecasting involved the following stations:

[001]: Bodrog, Felsőberecki
[002]: Bodrog, Sárospatak
[140]: Bodrog, Bodrogszerdahely
[101]: Latorca, Nagykapos
[122]: Laborc, Vaján.

For instance:

$$[001]_{t+l} = a_0(l) + b_1(l)\,[101]_t + b_2(l)\,[122]_t + b_3(l)\,[002]_t + e_t(l).$$

In case of Model (A):

$$\mathbf{R}_{xx} = \begin{bmatrix} 1.0 & 0.946 & 0.946 \\ 0.946 & 1.0 & 0.916 \\ 0.964 & 0.916 & 1.0 \end{bmatrix}$$

its inverse being:

$$\mathbf{R}_{xx}^{-1} = \begin{bmatrix} 27.710 & -8.495 & -13.147 \\ -8.495 & 9.537 & -0.547 \\ -13.147 & -0.547 & 14.175 \end{bmatrix}.$$

Furthermore:

$$\mathbf{r}_{xz}(1) = \begin{bmatrix} 0.972 \\ 0.955 \\ 0.973 \end{bmatrix}; \quad \mathbf{r}_{xz}(2) = \begin{bmatrix} 0.955 \\ 0.938 \\ 0.944 \end{bmatrix}; \quad \mathbf{r}_{xz}(3) = \begin{bmatrix} 0.930 \\ 0.910 \\ 0.910 \end{bmatrix}.$$

Accordingly, from (13) and (14):

$$\varrho(1) = 0.974 \qquad \varrho(2) = 0.930 \qquad \varrho(3) = 0.876$$

$$\sigma_e(1) = 23.38\ [\mathrm{cm}]; \quad \sigma_e(2) = 38.36\ [\mathrm{cm}]; \quad \sigma_e(3) = 51.06\ [\mathrm{cm}].$$

The $\varrho(l)$ and $\sigma_e(l)$ values are calculated in the same manner for any lead time. Lead time vs. forecast error has been plotted in Fig. 5, together with models

(B) $[001]_{t+l} = a_0(l) + b_1(l)\,[140]_t + b_2(l)\,[140]_{t-1} + b_3(l)\,[001]_t + e_t(l)$,

(C) $[001]_{t+l} = a_0(l) + b_1(l)\,[140]_t + b_2(l)\,[140]_{t-1} + e_t(l)$,

(D) $[001]_{t+l} = a_0(l) + b_1(l)\,[140]_t + b_2(l)\,[001]_t + e_t(l)$,

and also the forecast accuracy of model $AR(2)$ examined before:

$$(G)\quad [001]_{t+l} = a_0(l) + b_1(l)\,[001]_t + b_2(l)\,[001]_{t-1} + e_t(l)$$

has been analyzed by means of the relationship for cases of the generalized linear regression model.

According to Fig. 5, model accuracies follow different trends with increasing lead times, though with insignificant deviations but hinting to the possible need to change the forecast model structure with the variation of lead time. The deviation between two ways of calculation of model $AR(2) = (G)$ results from the implicit assumption that the process is perfectly described by model $AR(2)$, correlations r_3, r_4, . . . etc., dependent on r_1 and r_2 give no further information. The reality is, however, different, taken by the relationship for

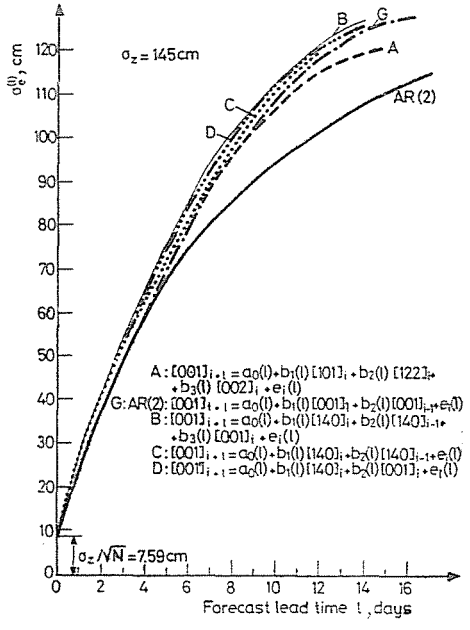

Fig. 6. Lead time to forecast error relations for different linear regression models on *Bodrog* stages at Felsőberecki

the generalized linear regression model into consideration, permitting, at the same time, to reckon to a degree with the uncertainty arising from the model identification.

Summary

Forecast lead time vs. accuracy relationships have been examined. Forecast accuracy has been described in terms of the expected value of the square of the deviation between the forecast and the real value. A relationship has been given for the lead time vs. accuracy of autoregressive moving average *(ARMA)* models. Variance of the forecast error may be ex-

pressed in terms of parameters φ and θ of $ARMA$ models for different lead times. An accuracy to lead time relationship has been established for the general linear regression model. Variance of the forecast error may be obtained from auto- and cross-correlations.

Examples have been presented on the application of the theoretical relationships to forecast daily stages of the *Bodrog* river, and monthly mean discharges of the *Tisza* river at *Szeged*.

References

1. Box, G. E. P.—Jenkins .G. M.: Time Series Analysis, Forecasting and Control. Holden-Day, San Francisco 1970.
2. Kontur, I.: Forecast Lead-Time vs. Accuracy Relationships. IInd National Congress of the Hungarian Hydrological Society, Pécs, July 1—2, 1981. Edited by the Federation of Technical and Scientific Societies.
3. Kontur, I.: Forecast Lead-Time vs. Accuracy Relationship in the Case of ARMA Models.* Hidrológiai Közlöny (in press).
4. Searle, S. R.: Linear Models. John Wiley. New York 1971.

Senior Assistant Dr. István Kontur, H-1521, Budapest

* In Hungarian