REVUE DES SCIENCES DE L'EAU, Rev. Sci. Eau 16/2(2003) 255-265

Short-term hydrological forecasts using linear regression

Prévisions hydrologiques à court terme obtenues en utilisant la régression linéaire

M. LEFEBVRE

Reçu le 16 janvier 2002, accepté le 22 novembre 2002**.

RÉSUMÉ

On trouve qu'un modèle très simple pour le débit d'une rivière, obtenu en se servant de la régression linéaire, donne de meilleurs résultats, pendant une certaine période, qu'un modèle déterministe utilisé actuellement. Les comparaisons entre les deux modèles sont basées sur trois critères importants, à savoir le coefficient de corrélation, la somme des erreurs au carré, et le critère de pointe. Le modèle est utilisé pendant la période de crue de la rivière, et les prévisions hydrologiques sont effectuées jusqu'à trois jours d'avance.

Mots clés : *modélisation, loi lognormale, corrélation, erreur standard, critère de pointe.*

SUMMARY

A very simple model for the flow of a river, obtained through linear regression, is found to give better results for a certain period when compared to the deterministic model currently in use. The comparisons between the two models are based on three important criteria: the correlation coefficient, the sum of the squares of the errors and the peak criterion. The model examined was used when the river was in spate and the forecasting horizon was a three-day period.

Key words: *modeling, lognormal distribution, correlation, standard error, peak criterion.*

Département de mathématiques et de génie industriel, École Polytechnique de Montréal, C.P. 6079, Succursale Centre-ville, Montréal, Québec, Canada H3C 3A7. Téléphone: 514-340-4711 (poste 4947). Télécopieur: 514-340-4463.

Correspondance. E-mail : mlefebvre@polymtl.ca

** Les commentaires seront reçus jusqu'au 30 décembre 2003.

1 - INTRODUCTION

The Alcan company (as well as other companies in Canada) uses a deterministic model known as PREVIS (see KITE (1978), BOUCHARD and SALESSE (1986), LAUZON (1995) and LAUZON *et al.* (1997)) to forecast the flow of certain rivers and catchment basins. Its objective is to obtain reliable forecasts for up to seven days ahead. LABIB *et al.* (2000) have proposed a stochastic model for the flow based on a two-dimensional Gaussian diffusion process. They found that for one-day forecasts this model is superior to PREVIS, based on four criteria. For two-day forecasts, it is comparable to PREVIS, but it cannot compete with PREVIS for three-day forecasts. The author (see LEFEBVRE 2002a) improved the model set up from LABIB *et al.* (2000) and was able to obtain forecasts that were sometimes more precise than those produced by PREVIS for three days ahead (and even more). The author also considered a model derived from a one-dimensional lognormal diffusion process (see LEFEBVRE 2002b) to forecast river flows. Although this last model is more robust than that in LABIB *et al.* (2000) and LEFEBVRE (2002a), in that the accuracy of the forecasts deteriorates less rapidly, it could not do as well for short-term forecasts.

PREVIS needs 18 entries, such as minimum and maximum temperatures, amount of precipitation, humidity, etc., to produce its forecasts. However, if we denote the flow at time t by $X(t)$, then the model in LABIB *et al.* (2000) requires only the knowledge of $X(t-1)$ and $X(t-2)$ to generate a forecast. In LEFEBVRE (2002a), the author first drew attention to the fact that it was more realistic to consider a lognormal rather than a Gaussian model and that it was preferable to work on a logarithmic scale when it comes to computing the various comparison criteria. The author then used linear regression, first to estimate a parameter in the model, and then to find out how to best incorporate various exogenous variables into the original model. Finally, linear regression was used in the same way in LEFEBVRE (2002b).

The objective of the papers mentioned above was to propose a simple stochastic model for the variations in flow, which possesses certain properties (such as producing Gaussian forecasts, from which confidence intervals for the forecasted flow values could be computed, for example). It was also intended to be able to make use of the stochastic model to forecast the maximum flow for a given period, as well as the first time the flow will reach a given threshold. However, the primary goal of a company such as Alcan is to receive as reliable forecasts as possible, especially for the next few days. To do so, we found out that we can obtain accurate results by making use of only two variables and linear regression.

The aim of the present paper is not to find the best possible model to forecast the flow of the Mistassibi river; but rather, to be able to predict as accurately with almost rudimentary models as with much more sophisticated models. We do not claim that a very simplistic model is always able to compete with complex ones. However, here we compare the linear regression model to a model requiring 18 entries (PREVIS) and to another model involving stochastic differential equations and diffusion processes. We found that the linear regression model produced the best hydrological forecasts. If the linear regression model had done almost as well as the other models considered, it would have already been noteworthy.

The formulas used to make forecasts are presented in the next section, as well as the numerical results. Concluding remarks follow in Section 3.

2 – FORECASTING EQUATIONS AND NUMERICAL RESULTS

Because the availability (to us) of the forecasts produced by PREVIS was limited to the Mistassibi river in Québec for the years 1993 to 1995, we will concentrate on this river and this time period. A map of Alcan's Saguenay-Lac-Saint-Jean hydroelectric system, which includes the Mistassibi river, is provided at the end of this paper.

Since it is especially important to provide Alcan with reliable forecasts during the period of spate, our study will be confined to this period. In the case of the Mistassibi river, it was found (see LEFEBVRE 2002a) that the river is in spate from around the end of March until late in May. Thus, the period of interest was the 51 days from March 29th to May 18th during the years 1993-1995. Two hydrographs of the Mistassibi river for the 100-day period from March 29th to July 6th for the years 1992 and 1993 are given in figures 1 and 2.

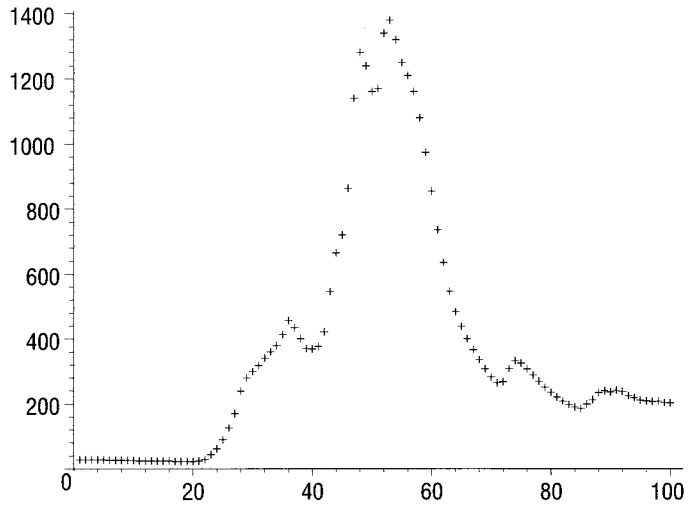

Figure 1 Hydrograph of the Mistassibi river from March 29th to July 6th 1992.

As mentioned in the Introduction, we will work on a logarithmic scale as far as the forecasts and the computation of the various criteria are concerned. This is due to the fact that the flow of the Mistassibi river varies from around 30 m^3/s

to more than 1000 m^3/s during the period of spate. Large variability in the flow can unduly influence the comparisons between the various models so that it distorts the reality.

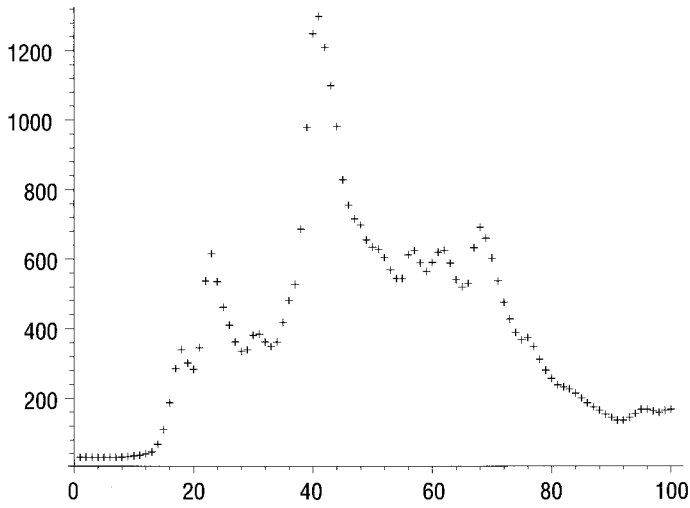

Figure 2 Hydrograph of the Mistassibi river from March 29th to July 6th 1993.

To predict the logarithm of the flow of the Mistassibi river k days ahead of time t, we will use the following forecasting equation:

$$\ln \widehat{X(t)} = c_0 + c_1 \ln X(t-k) + c_2 \ln X(t-k-1)$$

for k=1,2,3, where $X(t)$ is the observed flow at time t (an instantaneous flow observed every day) and the hat denotes the predicted value. Furthermore, c_0, c_1, and c_2 are the constants obtained through linear regression.

The criteria retained to carry out the comparisons between the competing models are the same as in LABIB *et al.* (2000). For example, the correlation coefficient r between the forecasted and observed logarithms of the flows, the sum SSQ of the squares of the forecasting errors and the peak criterion defined by

$$PC_k = \frac{\left[\sum_{t=1}^{N} \left(\ln \widehat{X_k(t)} - \ln X(t)\right)^2 \ln^2 X(t)\right]^{1/4}}{\left[\sum_{t=1}^{N} \ln^2 X(t)\right]^{1/2}} \qquad (1)$$

where N denotes the number of flow values greater than 1/3 of the mean peak flow over the period of interest and k is for the number of days ahead for which the forecast was produced. LABIB *et al.* (2000) used the standard error $STD = (SSQ/50)^{1/2}$ rather than SSQ and they also considered a fourth criterion, namely the Nash criterion given here by :

$$NC_k = 1 - \frac{\sum_{t=1}^{51} \left(\ln \widehat{X_k(t)} - \ln X(t)\right)^2}{\sum_{t=1}^{51} \left(\ln X(t) - < \ln X(t) >\right)^2},$$

where < > denotes the mean value. It is important to note that the denominator in the previous formula does not depend on the forecasting model, therefore, the model with the best value of *SSQ* will also have the best Nash criterion. For this reason, this fourth criterion will not be computed in the present paper.

The forecasting equations obtained through linear regression were computed with the 255 points of data from the flood season (March 29th to May 18th) for the five-year period 1988 to 1992 (taken all at once). These results are summarized below, together with the corresponding coefficients of determination (R^2):

$$\ln \widehat{X(t)} = 0.0534 + 1.75 \ln X(t-1) - 0.755 \ln X(t-2),\ R^2 \cong 0.997 \qquad (2)$$

$$\ln \widehat{X(t)} = 0.139 + 2.31 \ln X(t-2) - 1.33 \ln X(t-3),\ R^2 \cong 0.988 \qquad (3)$$

$$\ln \widehat{X(t)} = 0.238 + 2.75 \ln X(t-3) - 1.79 \ln X(t-4),\ R^2 \cong 0.972 \qquad (4)$$

Remarks.

i) The Durbin-Watson statistic, to test whether the errors are uncorrelated, was approximately 2.02. Since there was a sampling size of 255, we may conclude that the errors are indeed uncorrelated.

ii) We also computed the autocorrelation coefficients of the residuals for lags from 1 to 10. They are given by: -0.011; -0.060; 0.079; 0.010; 0.037; -0.007; -0.023; 0.030; -0.069 and -0.055. We see that all the autocorrelation coefficients were small and there was not a pattern.

iii) Finally, the partial correlation coefficient between ln $X(t)$ and ln $X(t-2)$, with ln $X(t-1)$ held fixed, was approximately equal to –0.735, which is logical, given the very strong positive correlations (0.997 and 0.988) between ln $X(t)$, ln $X(t-1)$ and ln $X(t-2)$.

The performance of the forecasting equations (2)-(4), with respect to the correlation coefficient and the sum of the squares of the errors, is summarized in tables 1-3 for the years 1993 to 1995, respectively. The accuracy of the forecasts was compared to that of the forecasts produced by PREVIS and by the stochastic model proposed in LEFEBVRE (2002a) (an improvement of the model set up in LABIB *et al.* (2000), as mentioned above). In these tables, the subscript *L* stands for linear regression, *S* for the stochastic model and *P* for PREVIS.

Table 1 *Performance of the various models for the year 1993.*

Tableau 1 Performance des différents modèles pour l'année 1993.

k	r_L	r_S	r_P	SSQ_L	SSQ_S	SSQ_P
1	0.996	0.996	0.889	0.618	0.645	24.723
2	0.980	0.979	0.872	3.507	3.834	17.5913^3
3	0.955	0.953	0.866	8.019	9.175	18.1263^3

Table 2 *Performance of the various models for the year 1994.*

Tableau 2 Performance des différents modèles pour l'année 1994.

k	r_L	r_S	r_P	SSQ_L	SSQ_S	SSQ_P
1	0.998	0.998	0.994	0.250	0.251	1.150
2	0.991	0.990	0.993	1.238	1.338	1.565
3	0.977	0.976	0.991	3.039	3.378	2.087

Table 3 *Performance of the various models for the year 1995.*

Tableau 3 Performance des différents modèles pour l'année 1995.

k	r_L	r_S	r_P	SSQ_L	SSQ_S	SSQ_P
1	0.998	0.998	0.989	0.252	0.256	1.461
2	0.992	0.992	0.984	0.929	0.996	2.029
3	0.981	0.980	0.979	2.239	2.381	2.568

Looking at the numbers in tables 1-3, we must conclude that the forecasts produced by the model obtained through linear regression are the most accurate. Although the difference in the values of the correlation coefficients is sometimes rather small, the linear regression model is more accurate than the stochastic model systematically in the case of the *SSQ* criterion, which is the more important criterion. As for PREVIS, it had poor results in the year 1993. Actually, its performance is even worse than what appears in Table 1 because the superscripts in the SSQ_P column denote the number of forecasts produced by PREVIS that had to be discarded because they were negative. However, in 1994 and in 1995, PREVIS did very well. We see that neither the stochastic model nor the linear regression model was able to beat PREVIS for $k = 3$ in 1994. Nevertheless, we may conclude that the linear regression model is worth considering when it comes to forecasting the Mistassibi river flows during the springtime.

It could be argued that the two criteria used so far, namely the correlation coefficient and the sum of the squares of the errors, favor the linear regression model. Due to this, we also decided to consider the peak criterion, as described above. This criterion is useful to judge the quality of the forecasts during the flood season.

In 1993, there were four peaks between March 29th and May 18th: 340 m^3/s, 616 m^3/s, 385 m^3/s and 1300 m^3/s, for a mean peak flow of

660.25 m^3/s. According to the PC criterion, we must discard all the observed flows below approximately 220 m^3/s. Therefore, we must eliminate all the data from March 29th to April 13th, so that $N = 35$ in the formula (1).

In 1994, there were two peak flows, averaging 515 m^3/s. This time, we had to discard the period from March 29th to April 19th, and from April 24th to April 29th, giving a value of N equal to 23.

Finally, in 1995 there were three peak flows (if we count the flow on May 18th which was the maximum flow over the entire period of interest), averaging approximately 694 m^3/s. We eliminated the flows from March 29th to April 23rd, thus obtaining $N = 25$.

The various values of the peak criterion are shown in table 4 for the linear regression model and PREVIS. The closer to zero the numerical value is, the better the model performed.

Table 4 *Numerical values of the peak criterion for the years 1993 to 1995.*

Tableau 4 Valeurs numériques du critère de pointe pour les années 1993 à 1995.

	1993		*1994*		*1995*	
k	PC_L	PC_P	PC_L	PC_P	PC_L	PC_P
1	0.05644	0.08705	0.04566	0.06897	0.04281	0.07332
2	0.08646	0.09205	0.06843	0.07387	0.06827	0.08032
3	0.10541	0.09524	0.08566	0.07863	0.08914	0.08539

We see that the linear regression model performed better than PREVIS for one and two-day forecasts in every year. However, PREVIS did slightly better than the linear regression model systematically (even for 1993) for $k = 3$.

Hence, the linear regression model was able to forecast flow values better than PREVIS up to two days ahead, during the period when the river flow was high, which is really important.

It is interesting to check how both models performed when we discard all the flows except the peak flows for each year and we compute the peak criterion with $N = 4$, 2 and 3 data, respectively. The computations were made for $k=1$ and $k = 2$ and are shown in table 5.

Table 5 *The peak criterion computed with only the peak flows for the years 1993 to 1995.*

Tableau 5 Le critère de pointe calculé à partir des débits de pointe seulement pour les années 1993 à 1995.

	1993		*1994*		*1995*	
k	PC_L	PC_P	PC_L	PC_P	PC_L	PC_P
1	0.05382	0.10943	0.04567	0.06897	0.04325	0.06774
2	0.08313	0.10439	0.06843	0.07387	0.06075	0.07332

Again, the simple linear regression model was always better than PREVIS at forecasting peak flows one and two days ahead. However, the number of data used to compute the peak criterion was so small that we cannot state very strong conclusions.

Finally, we justified the logarithmic transformation of the data because it is recommended to use such a transformation to reduce the effect on the comparison criteria of a few poor forecasts of the flow values when the flow is very large. We could have used another transformation to attain this objective. For example, we recomputed the correlation coefficients and the sum of squares SSQ obtained with PREVIS and the linear regression model when the square root transformation was applied to the data instead. The results are presented in tables 6 and 7 for the years 1994 and 1995. We see that the conclusions are practically the same as with the logarithmic transformation.

Table 6 *Performance of PREVIS and the linear regression model for the year 1994 with the square root transformation.*

Tableau 6 Performance de PRÉVIS et du modèle de régression linéaire pour l'année 1994 avec la transformation racine carrée.

k	r_L	r_P	SSQ_L	SSQ_P
1	0.997	0.990	18.517	89.948
2	0.987	0.988	86.076	112.89
3	0.972	0.986	184.80	136.12

Table 7 *Performance of PREVIS and the linear regression model for the year 1995 with the square root transformation.*

Tableau 7 Performance de PRÉVIS et du modèle de régression linéaire pour l'année 1995 avec la transformation racine carrée.

k	r_L	r_P	SSQ_L	SSQ_P
1	0.998	0.987	17.584	100.03
2	0.990	0.983	79.587	130.08
3	0.976	0.979	172.71	156.10

3 – CONCLUDING REMARKS

The forecasting equations for the flow of the Mistassibi river, obtained through linear regression, gave unexpected results. The linear regression model performed better than PREVIS for one and two-day forecasts, for the three years considered, and could surely compete with PREVIS for three-day forecasts.

In order to obtain even more accurate forecasts, various options exist. First, we could use more data (more years) to compute the forecasting equations. However, there is no guarantee that increasing the number of data would improve the accuracy of the forecasts.

Another way to improve the accuracy of the forecasts is to add at least one exogenous variable to the model, as was done in LEFEBVRE (2002a). One such variable is the temperature on the day when the flows are to be forecasted. Because this piece of information is only available to us for the years 1993 and 1994, we had to limit ourselves to those two years. The forecasting equation for $k = 1$ is the following:

$$\ln \widehat{X(t)} = 0.118 + 0.00481\ T(t-1) + 1.69 \ln X(t-1) - 0.715 \ln X(t-2),$$

where $T(t-1)$ is the average temperature on day $t-1$.

The values of the correlation coefficient r and of the sum of squares SSQ obtained with this forecasting equation are shown in Table 8.

Table 8 *Performance obtained by adding temperature to the model for $k = 1$ for the years 1993 and 1994.*

Tableau 8 Performance obtenue en incluant la température dans le modèle pour $k=1$ pour les années 1993 et 1994.

	1993	**1994**
r	0.997	0.998
SSQ	0.5908	0.2493

Comparing the numbers in table 8 to the corresponding ones in tables 1 and 2, we notice that the addition of the temperature to the linear regression model has had a positive effect in 1993; however, the value of SSQ in 1994 is only slightly smaller with the variable $T(t-1)$ incorporated into the model. This is probably due to the fact that the temperature is included in the flow variables. Therefore, the conclusion on the usefulness of the temperature is not clear. We could of course incorporate a different exogenous variable or more than only one exogenous variable. However, it was found in LEFEBVRE (2002a) that adding the amount of precipitation to the model had very little impact, again probably because precipitation effects are also included in the flow variables. Another explanation is that the relationships between precipitation or temperature and (the logarithm of the) flow are likely nonlinear rather than linear.

Next, we could also compute forecasting equations based on more than two values of the flow, namely the flow at time $t-1$ and at time $t-2$. We could try to measure the quality of the forecasts produced by a regression equation involving the flow at times $t-1$, $t-2$ and $t-3$, for example.

Finally, as was done in LEFEBVRE (2002a,b), we can consider forecasts obtained by taking the mean of the forecasts produced by PREVIS and by the linear regression model. This idea can be extended to take into account forecasts produced by any model, such as one based on neural networks. The performance of the averaged forecasts is shown in table 9 for $k = 1$ and $k = 2$.

Table 9 *Performance of the average forecasts for k = 1 and k = 2, for the years 1993 to 1995.*

Tableau 9 Performance de la moyenne des prévisions pour $k = 1$ et $k = 2$, pour les années 1993 à 1995.

	1993		1994		1995	
k	***r***	***SSQ***	***r***	***SSQ***	***r***	***SSQ***
1	0.968	6.171	0.998	0.325	0.996	0.488
2	0.961	5.369	0.996	0.705	0.992	0.998

While for $k = 1$ this procedure is not recommended, essentially because the linear regression model is superior to PREVIS, we notice that for $k = 2$ in 1994 the results are very impressive. The value of *SSQ* in that case was reduced by over 43%, compared with the smallest value of *SSQ*, namely $SSQ_L = 1.238$.

In conclusion, although it is surely possible to further improve the accuracy of the forecasts, the results obtained by making use of the regression equations presented in this paper are very satisfactory. Considering the simplicity of the model, and hence its low implementation cost, it can be considered as an alternative to or as a complement to a deterministic model such as PREVIS, at least for short-term forecasting. For seven-day forecasts, the various variables PREVIS uses come into effect and help produce quite reliable forecasts. Furthermore, no probabilistic assumptions were made in this paper. Therefore, the stochastic models proposed, in particular, by LABIB *et al.* (2000) and by LEFEBVRE (2002a,b) have other advantages. Nevertheless, as far as the accuracy of short-term forecasts is concerned, the linear regression model is the winner.

ACKNOWLEDGMENTS

This research was supported by the Natural Sciences and Engineering Research Council of Canada. The author also expresses his gratitude to the referees of this paper for their constructive comments.

REFERENCES

BOUCHARD S., SALESSE L., 1986. Amélioration et structuration du système de prévision hydrologique à court terme PRÉVIS. Groupe de Ressources Hydrauliques, ÉÉQ, SÉCAL, Jonquière, Québec, Canada, Rapport RH-86-01, pp. 1-31.

KITE G.W., 1978. Development of a hydrological model for a Canadian watershed. *Rev. Can. Génie Civ.*, 5, 126-134.

LABIB R., LEFEBVRE M., RIBEIRO J., ROUSSELLE J., TRUNG H.T., 2000. Application of diffusion processes to runoff estimation. *J. Hydrol. Eng.*, 5, 1-7.

LAUZON N., 1995. Méthodes de validation et de prévision à court terme des apports naturels. Mémoire de maîtrise, École Polytechnique, Montréal, Québec, Canada.

LAUZON N., BIRIKUNDAVYI S., GIGNAC C., ROUSSELLE J., 1997. Comparaison de deux procédures d'amélioration des prévisions à court terme des apports naturels d'un modèle déterministe. *Rev. Can. Génie Civ.*, 24, 723-735.

LEFEBVRE M., 2002a. Using a lognormal diffusion process to forecast river flows. *Water Resour. Res.* (À paraître)

LEFEBVRE M., 2002b. Geometric Brownian motion as a model for river flows. *Hydrol. Process.*, 16, 1373-1381.