Mathematical Problems in Engineering, 2002, Vol. 8(6), pp. 553–562 Taylor & Francis
Taylor & Francis Group

Modeling and Forecasting the Peak Flows of a River

MARIO LEFEBVRE*

Département de mathématiques et de génie industriel, École Polytechnique de Montréal, C.P. 6079, Succursale Centre-ville, Montréal, Québec, Canada H3C 3A7

(Received 10 November 2001)

A stochastic model is found for the value of the peak flows of the Mistassibi river in Québec, Canada, when the river is in spate. Next, the objective is to forecast the value of the coming peak flow about four days in advance, when the flow begins to show a marked increase. Both the stochastic model proposed in the paper and a model based on linear regression are used to produce the forecasts. The quality of the forecasts is assessed by considering the standard errors and the peak criterion. The forecasts are much more accurate than those obtained by taking the mean value of the previous peak flows.

Key words: Stochastic modeling; Lognormal distribution; Linear regression; Correlation coefficient; Peak criterion; Hydrology

AMS Subject Classification: Primary: 62P30; Secondary: 86A05

1 INTRODUCTION

The problem of forecasting the value of the flow of various rivers and/or hydrological basins in Québec, Canada, has been considered by Labib *et al.* [4] and by the author (see Lefebvre [7], for instance), in particular. Their objective was to forecast the flow up to seven days ahead. They compared the results obtained by making use of various stochastic models to those obtained with a deterministic model known as PREVIS (see Refs. [1, 3, 5, 6]), which is currently used by a number of companies in Canada. It was found that relatively simple stochastic models could outperform PREVIS, which requires the evaluation of 18 parameters, for forecasts up to three and sometimes four days in advance. However, PREVIS generally produces more reliable forecasts from five days ahead.

Next, Lefebvre [8] tried to model the maximum flow of the Mistassibi river during each of the months of April, May and June, as well as to forecast the maximum flow in May, based on the observed maximum flow in April. This three-month period is the time when the river is in spate and the maximum value of the flow in May is also most of the time the maximum flow over the three-month period.

In the present paper, instead of trying to forecast the maximum flow in May, based on the maximum flow in April (which is very often observed on April 30th), we will attempt to forecast the value of the various peaks of the river flow during the period when the river is in spate. In some years, two or even three peaks that could cause flooding were observed. So,

* E-mail: mlefebvre@polymtl.ca

ISSN 1024-123X print; ISSN 1563-5147 online © 2002 Taylor & Francis Ltd

the problem is different from the one considered in Lefebvre [8]. Moreover, it was found in Lefebvre [8] that the maximum flow in May usually happens around May 15th and that the correlation between the maximum in April and that in May is rather weak. Therefore, it is difficult to make use of the observed maximum flow in April to forecast the maximum flow in May with high accuracy.

Here, we approach the problem of forecasting the peak flows of the Mistassibi river in a different way. More precisely, we will seek to forecast an oncoming peak flow about four days before its occurrence. Indeed, in most cases the river flow shows a marked increase at least three days before a peak. Our objective is to arrive at quite accurate forecasts of these peaks quickly enough so that the persons in charge can take action to prevent flooding if it is deemed necessary.

Other authors have tried to forecast peak flows of rivers, as well as the time of occurrence of these peak flows. Rosbjerg [9], in particular, has proposed a model and an estimator for the maximum flow (see also Ref. [2]). However, Rosbjerg's estimator depends on the correlation coefficient of two consecutive peaks. In our case, there are many years for which there is but a single peak during the whole spring season. Hence, we cannot make use of the formula developed by Rosbjerg.

In Section 2, stochastic models will be found for the peak flows of the Mistassibi river and for the river flows on the previous days. Next, in Section 3 we will make use of the models obtained in Section 2 to forecast the peak flows of the river. As will be seen, even better forecasts will in fact be produced by another model, based on linear regression. Finally, a few remarks will conclude this work in Section 4.

2 STOCHASTIC MODELS

The observed flows of the Mistassibi river are available to us for the period from 1953 to 1994. However, due to numerous missing values for the first years, we decided to limit our study to the years 1963 to 1994. Over this time period, during the months of April, May and June, we have identified 54 occurrences when the river flow has had a daily increase of at least $90\,\mathrm{m}^3/\mathrm{s}$, leading to a peak flow in the following days. The data are presented in Tables I and II. We have included the value of the flow before the large increase (Flow), the size of the increase (Increase), the value of the flow one and two days after the increase (Flow2 and Flow3), and finally the value of the ensuing peak flow (Max) as well as the number of days elapsed until the peak flow (*N*). Moreover, Table I presents the data for the years 1963–1979, while Table II does so for the years 1980–1994.

Remark The value of $90\,\mathrm{m}^3/\mathrm{s}$ was chosen so that the peak flow could be forecasted with enough accuracy and early enough to advise the administrators to take action if needed. A $50\,\mathrm{m}^3/\mathrm{s}$ increase, for instance, leads to too many "false alarms" or lack of precision, whereas a $150\,\mathrm{m}^3/\mathrm{s}$ increase as a warning signal would entail missing some peak flows or leaving too little time to the administrators.

First, we find that the peak flows occurred on average approximately 3.5 days after the $90\,\mathrm{m}^3/\mathrm{s}$ (or more) increase. Therefore, if we could produce accurate enough forecasts of the peak flows when this large increase is observed, it would leave a few days to act in order to prevent floodings.

Next, we have tested the hypothesis that the variables Flow, Flow2, Flow3 and Max in Tables I and II combined follow a Gaussian distribution, as well as the variable

$$\text{Flow1} := \text{Flow} + \text{Increase};$$

TABLE I Data for the years 1963–1979.

Date	*Flow*	*Increase*	*Flow2*	*Flow3*	*Max*	*N*
63/05/19	660	150	949	971	971	3
64/05/01	510	136	745	801	1240	11
65/05/09	220	131	487	575	728	5
65/05/17	731	161	1000	1030	1030	3
66/05/17	219	109	459	711	1010	6
67/05/27	425	105	617	674	668	4
68/04/22	411	99	600	711	1000	5
69/05/18	580	97	750	818	818	3
69/06/04	663	96	799	793	813	4
70/05/01	382	170	878	997	997	3
70/05/17	430	97	682	714	739	4
70/06/11	402	139	651	595	651	2
71/05/10	597	94	756	841	960	5
72/05/15	268	97	515	671	1060	7
73/04/24	300	91	467	504	748	7
73/05/04	773	164	1050	1030	1050	2
74/05/12	408	158	733	892	1030	6
74/06/01	1080	130	1300	1320	1350	5
75/05/04	272	96	411	476	762	9
75/06/01	405	110	578	561	578	2
76/04/29	484	99	614	674	1010	6
76/05/12	685	300	1250	1300	1300	3
76/05/18	1350	130	1530	1560	1560	3
77/04/24	125	97	419	504	544	5
77/05/07	580	100	680	629	680	1
77/05/17	782	113	991	1000	1000	3
77/05/23	997	93	1150	1150	1150	2
78/05/09	438	136	757	988	1160	8
78/06/13	417	336	985	1070	1070	3
79/04/27	449	285	1090	1380	1480	4
79/05/13	584	149	896	862	896	2
79/06/12	448	100	575	573	701	7

TABLE II Data for the years 1980–1994.

Date	*Flow*	*Increase*	*Flow2*	*Flow3*	*Max*	*N*
80/04/25	246	121	419	489	911	9
81/05/06	437	105	621	614	621	2
81/05/14	853	132	1130	1170	1230	4
82/05/07	550	133	871	1050	1200	5
82/06/01	391	116	685	689	689	3
83/04/30	600	163	905	939	961	5
83/05/14	584	135	882	960	960	3
84/04/25	321	100	515	583	986	8
85/05/17	528	134	837	950	1100	5
85/06/02	696	132	826	763	826	1
86/04/27	493	158	857	1010	1290	7
87/04/02	352	97	414	373	449	2
87/04/19	361	90	504	530	580	5
89/05/01	115	96	250	289	854	10
90/05/11	635	130	884	833	884	2
91/05/01	437	110	654	691	691	3
92/05/09	422	124	665	720	1280	6
92/05/18	1170	170	1380	1320	1380	2
93/04/13	188	98	340	302	616	7
93/05/04	528	159	979	1250	1300	4
93/06/02	531	102	692	660	692	2
94/05/06	284	143	591	646	848	8

TABLE III p-Values associated with the normality tests.

Variable	*p-Value*
Flow	0.003
Flow1	0.023
Flow2	0.358
Flow3	0.212
Max	0.410

TABLE IV p-Values associated with the normality tests applied to the logarithms of the variables in Table III.

Variable	*p-Value*
LnFlow	0.153
LnFlow1	0.490
LnFlow2	0.892
LnFlow3	0.507
LnMax	0.433

we have used the Anderson-Darling test (as well as the Ryan-Joiner test actually). The p-values of the tests, that is, the smallest values of the level α of the tests that can be used to reject normality, are shown in Table III.

We see that, apart from the variables Flow and Flow1, the normality assumption cannot be rejected with a small α. However, we have also applied the Anderson-Darling test to the natural logarithms of the same variables. The corresponding p-values are given in Table IV.

It is obvious that the lognormal distribution is a better model than the Gaussian distribution for the data.

Then, we have computed the correlation coefficient of LnMax and each of the variables LnFlow, LnFlow1, LnFlow2 and LnFlow3 (see Tab. V).

As could be expected, the correlation coefficient of LnMax and the natural logarithm of the observed flow increases when the observed flow is closer to the maximum.

In the next section, the various Flow variables will be used to try to forecast the peak flows as accurately as possible. For the moment, we are looking for a stochastic model for the peak flows. Based on what precedes, we can state that the *natural logarithm* of the peak flow seems to follow a *Gaussian* distribution with mean and standard deviation approximately equal to 6.8148 and 0.2814. Similarly, the variables LnFlow, LnFlow1, LnFlow2 and LnFlow3 also seem to have a Gaussian distribution with means and standard deviations given in Table VI.

TABLE V Correlation coefficients of LnMax and the logarithms of the flows.

Variable	*Correlation coefficient*
LnFlow	0.505
LnFlow1	0.578
LnFlow2	0.669
LnFlow3	0.782

TABLE VI Means and standard deviations of the variables LnFlow, LnFlowl, LnFlow2 and LnFlow3.

Variable	*Mean*	*Standard deviation*
LnFlow	6.1313	0.4949
LnFlowl	6.3936	0.4058
LnFlow2	6.5744	0.3704
LnFlow3	6.6467	0.3657

Finally, given that the Gaussian distribution is a good model for the variable LnMax, it is a simple matter to compute estimates of the flow values that the peak flow exceeds with probability 1% or 0.1%, etc.

3 FORECASTING THE PEAK FLOW VALUES

We now turn to the problem of forecasting the value of the peak flow, based on the observed flow, together with the value of the 90^+ m^3/s increase. Moreover, because the peak flow actually occurs, on average, about 3.5 days after the large flow increase has been noticed, it is interesting to try to forecast the value of the peak flow once the actual flows one and even two days after this large flow increase are known. We will make use of the data in Table I, that is for the years 1963–1979, to arrive at an estimator, and then we will attempt to forecast the peak flows that have been observed over the years 1980 to 1994.

A first formula can be obtained by remembering that if the random vector (X_1, X_2) has a bivariate Gaussian distribution, then

$$E[X_2|X_1 = x] = \mu_{X_2} + \rho_{X_1,X_2}\frac{\sigma_{X_2}}{\sigma_{X_1}}(x - \mu_{X_1}),$$

where μ_{X_i} is the mean of X_i, σ_{X_i} is its standard deviation and ρ_{X_1,X_2} is the correlation coefficient of X_1 and X_2. Hence, computing the various quantities needed to estimate the mean of the variable Max, based on the value of the variable Flow, we obtain the following formula (see Tabs. VII and VIII):

$$\text{Gaus} := \exp\left\{6.8344 + 0.556\left(\frac{0.2652}{0.4932}\right)(\text{LnFlow} - 6.1674)\right\}.$$

TABLE VII Means and standard deviations of the variables LnFlow, LnFlowl, LnFlow2, LnFlow3 and LnMax using the data in Table I.

Variable	*Mean*	*Standard deviation*
LnFlow	6.1674	0.4932
LnFlowl	6.4294	0.4075
LnFlow2	6.6181	0.3435
LnFlow3	6.7053	0.3136
LnMax	6.8344	0.2652

TABLE VIII Correlation coefficients of LnMax and the logarithms of the flows using the data in Table I.

Variable	*Correlation coefficient*
LnFlow	0.556
LnFlow1	0.635
LnFlow2	0.718
LnFlow3	0.853

Similarly, the forecasts based on LnFlow1 and on the observed flows one and two days after the $90^{+}\,\mathrm{m}^3/\mathrm{s}$ increase are respectively

$$\mathrm{Gaus1} := \exp\left\{6.8344 + 0.635\left(\frac{0.2652}{0.4075}\right)(\mathrm{LnFlow1} - 6.4294)\right\},$$
$$\mathrm{Gaus2} := \exp\left\{6.8344 + 0.718\left(\frac{0.2652}{0.3435}\right)(\mathrm{LnFlow2} - 6.6181)\right\},$$

and

$$\mathrm{Gaus3} := \exp\left\{6.8344 + 0.853\left(\frac{0.2652}{0.3136}\right)(\mathrm{LnFlow3} - 6.7053)\right\}.$$

To assess the quality of the forecasts produced by the estimators above, we will compute the correlation coefficients of the forecasts and the observed peak flows. However, two more important criteria are the standard error defined by

$$\mathrm{STD} = \left[\frac{\mathrm{SSQ}}{n-1}\right]^{1/2},$$

where SSQ is the sum of the squares of the forecasting errors and $n = 22$ in our case, and the peak criterion:

$$\mathrm{PC} := \frac{\left[\sum_{i=1}^{22}(\hat{X}_k - X_k)^2 X_k^2\right]^{1/4}}{\left[\sum_{i=1}^{22} X_k^2\right]^{1/2}}, \tag{1}$$

in which X_k is the kth observed peak flow and $\hat{X}_k$ is the corresponding forecasted peak flow. This last criterion enables us to measure the capacity of the estimator for forecasting very high flows, which is really important. Moreover, we can show that the peak criterion PC is in the interval [0,1] and that the closer to 0 the quantity PC is, the better the forecasts are.

Next, to get a better idea of the quality of the forecasts produced by the formulae Gaus, Gaus1, etc., we will also compute the values of the criteria above obtained with other estimators. A "naive" estimator of the next peak flow is the average value of all the previous peak flows observed. Therefore, we can estimate the first peak flow in Table II by computing the average value of the 32 peak flows in Table I; then, we estimate the second peak flow in Table II by computing the mean of the 33 previous peak flows, etc. We denote this estimator by MEAN.

Since the maximum flow occurs on average between four and five days after the variable Flow has been observed, another simple estimator of the peak flow is given by

$$\mathrm{LIN1} := \mathrm{Flow} + 5 \times \mathrm{Increase}.$$

Remarks (1) Notice that this estimator doesn't make use of the data in Table I. (2) Because the mean value of the variable N is approximately equal to 4.6, we could have set LIN1 = Flow + 4.6 × Increase instead. However, as will be seen below, this estimator isn't very reliable and its performance is actually worse when we replace 5 by 4.6.

In a similar way, once the values of Flow2 and Flow3 are known, we can estimate the coming peak flow by

$$\text{LIN2} := \text{Flow1} + 4 \times (\text{Flow2} - \text{Flow1})$$

and

$$\text{LIN3} := \text{Flow2} + 3 \times (\text{Flow3} - \text{Flow2}).$$

Finally, a technique that has been used successfully by the author in previous papers is that of linear regression. The various regression equations obtained with the data in Table I are the following:

$$\text{Max} = 414 + 0.550 \times \text{Flow} + 1.86 \times \text{Increase} := \text{Reg1},$$
$$\text{Max} = 368 - 0.599 \times \text{Flow} - 0.69 \times \text{Increase} + 1.27 \times \text{Flow2} := \text{Reg2}$$

and

$$\text{Max} = 248 + 0.724 \times \text{Flow} + 0.983 \times \text{Increase} - 1.85 \times \text{Flow2} + 1.93 \times \text{Flow3} := \text{Reg3}.$$

Remark These regression equations are quite reliable. Indeed, the corresponding coefficients of determination R^2 are given by 55.3%, 62.0% and 85.7% respectively. Moreover, if we include the variable N into the model, we get a coefficient of determination (approximately) equal to 92.9%. Of course, the value of N is not known when we want to forecast the next peak flow, so it cannot be used in the forecasting formulae.

First, Table IX presents the correlation coefficients of Max and the various estimators proposed above.

We see that the LIN and Reg estimators have the best correlation coefficients, the Gaus estimators being a little less correlated with the variable Max. Although the important criteria are the standard error and the peak criterion, as mentioned above, we may immediately conclude that the MEAN estimator is really not reliable at all.

Next, we compute the standard errors obtained with each of the estimators in Table IX. The results are shown in Table X.

TABLE IX Correlation coefficients of Max and its various estimators.

Estimator	*Correlation coefficient*
MEAN	−0.389
LIN1	0.666
LIN2	0.749
LIN3	0.825
Gaus	0.498
Gaus1	0.555
Gaus2	0.667
Gaus3	0.779
Reg1	0.638
Reg2	0.734
Reg3	0.837

TABLE X Standard errors (in m^3/s) obtained with the various estimators of Max.

Estimator	*Standard error*
MEAN	274.8
LIN1	310.7
LIN2	310.3
LIN3	239.2
Gaus	234.6
Gaus1	225.4
Gaus2	203.6
Gaus3	180.9
Reg1	208.9
Reg2	184.2
Reg3	162.8

TABLE XI Peak criteria obtained with various estimators of Max.

Estimator	*Peak criterion*
MEAN	0.2451
Gaus	0.2611
Gaus1	0.2254
Gaus2	0.2632
Gaus3	0.2378
Reg1	0.2184
Reg2	0.2040
Reg3	0.1977

Table X enables us to discard the LIN estimators, since the corresponding Gaus and Reg estimators are clearly better.

Finally, another important criterion is the peak criterion defined in (1). We computed this criterion for the estimators MEAN, Gaus and Reg, as shown in Table XI.

Surprisingly, the naive estimator MEAN has a smaller value of PC than Gaus and Gaus2. At any rate, it is now obvious that the best forecasts are produced by the Reg estimators. Indeed, the Reg estimators always have the smallest standard errors and the smallest values of the peak criterion. We see that, compared to the accuracy of the MEAN estimator, there is a decrease of 24% to over 42% in the standard error.

The Gaus estimators do quite well as far as the standard error is concerned; however, they don't seem to be able to forecast (very) high flows very well, which is a serious flaw. Notice that these estimators are all based on a bivariate Gaussian distribution, whereas Reg$i$ involves $i + 1$ variables, for $i = 1, 2, 3$.

4 CONCLUDING REMARKS

In this paper, a simple stochastic model was first proposed for the peak flows of the Mistassibi river, in Québec, during the months of April to June, that is, when the river is in spate. We saw in Section 2 that the variable Max seems to have a lognormal distribution (although a Gaussian distribution would also have been acceptable for this variable). This model enables us to compute flow values that will only be exceeded with a given (small) probability. Confidence intervals could also be obtained from this model for the mean peak flow.

Next, we turned to the problem of forecasting the peak flows of the river once an important increase in the daily flow has been observed. We found in Section 3 that a very good estimator of the coming peak flow is obtained by making use of linear regression. Of course, the more observations of the actual flow we have, the more accurate the forecasts are. Since the peak flow occurs on average approximately three and a half days after an increase of 90 or more m^3/s, we considered the estimators based on the flows before and after the large increase, as well as the flows one and two days after the large increase. Although the estimator based on all these variables at the same time is naturally the most accurate, it might often be too late to act if we wait two days after having noticed a large increase in the daily flow. Therefore, the other estimators (Reg1 in particular) are also useful. The administrators can always react and change their decision if the forecast produced by Reg2 is less alarming than that provided by Reg1, for instance.

Now, to improve further the accuracy of the forecasts, we could try to incorporate at least another variable into the regression equation, such as the amount of precipitation on the day when the large flow increase was observed. This was done in Lefebvre [8] without much success; however, as mentioned previously, there is not enough correlation between the maximum flow in April and that in May. In the present case, knowing the amount of precipitation a few days before the peak flow could prove to be useful.

Finally, it has been found in other works that another way of improving the forecasts is to take the arithmetic mean of the forecasts produced by various estimators. Here, if we define

$$\mathrm{Ave3} = \frac{\mathrm{MEAN} + \mathrm{LIN3} + \mathrm{Gaus3} + \mathrm{Reg3}}{4},$$

we find that the correlation coefficient of Max and Ave3 is 0.827, the standard error computed with this estimator is $156.1\ \mathrm{m}^3/\mathrm{s}$, and the peak criterion is 0.1951. These values of the standard error and of the peak criterion are better than those obtained with any single estimator. So, even if some estimators (such as MEAN) considered individually do not provide reliable forecasts, they can nevertheless be used to improve the quality of the forecasts. Actually, if we define instead

$$\mathrm{Ave3} = \frac{\mathrm{LIN3} + \mathrm{Gaus3} + \mathrm{Reg3}}{3},$$

we find that the correlation coefficient of Max and Ave3 increases to 0.829; however, the standard error and the peak criterion also increase, to $177.6\,\mathrm{m}^3/\mathrm{s}$ and 0.2042 respectively. Thus, rather surprisingly, it is better here to leave MEAN in the formula.

Acknowledgements

This work was supported by the Natural Sciences and Engineering Research Council of Canada.

References

[1] Bouchard, S. and Salesse, L. (1986) *Amélioration et structuration du système de prévision hydrologique à court terme PRÉVIS*, Rep. RH-86-01, Group of Hydraulic Resources, ÉÉQ, SÉCAL, Jonquière, Québec, Canada, 1986, pp. 1–31.

[2] Gupta, V. K., Duckstein, L. and Peebles, R. W. (1976) On the joint distribution of the largest flood and its time of occurrence, *Water Resour. Res.*, **12**, 295–304.

[3] Kite, G. W. (1978) Development of a hydrological model for a Canadian watershed, *Can. J. Civ. Engrg.*, **5**, 126–134.

[4] Labib, R., Lefebvre, M., Ribeiro, J., Rousselle, J. and Trung, H. T. (2000) Application of diffusion processes to runoff estimation, *J. Hydro. Engrg.*, **5**, 1–7.

[5] Lauzon, N. (1995) Méthodes de validation et de prévision à court terme des apports naturels, *Master's thesis*, Département de génie civil, École Polytechnique, Montréal, Québec, Canada, 1995.

[6] Lauzon, N., Birikundavyi, S., Gignac, C. and Rousselle, J. (1997) Comparaison de deux procédures d'amélioration des prévisions à court terme des apports naturels d'un modèle déterministe, *Can. J. Civ. Engrg.*, **24**, 723–735.

[7] Lefebvre, M. (2002) Using a lognormal diffusion process to forecast river flows, *Water Resource Research*, **38**.

[8] Lefebvre, M. (2001) Au sujet du debit maximal de la rivière Mistassibi durant la période printanière, *Can. J. Civ. Engrg.*, **28**, 1041–1045.

[9] Rosbjerg, D. (1987) On the annual maximum distribution in dependent partial duration series, *Stochastic Hydrology Hydraul.*, **1**, 3–16.