Estimating the probability of exceeding a given river flow threshold

M. Lefebvre

Abstract. Estimating the risk that a certain river flow will exceed a given threshold in the short term is an important problem in statistical hydrology. Using two stochastic processes with jumps, the author and Bensalma were able to derive formulas to estimate that risk. In an application to the Delaware River, both models proposed by the authors performed quite well. However, in order to obtain even more accurate estimates of the probability of threshold exceedance, exogenous variables, such as the amount of precipitation observed and forecast, should be incorporated into the models. One possible way to determine how to incorporate exogenous variables is to make use of linear regression. Using this technique, much more precise estimates are obtained.

M.S.C. 2010: 60J70, 60K40.
Key words: Diffusion process with jumps; filtered Poisson process; forecasting; linear regression.

1 Introduction

A basic problem in statistical hydrology is to forecast the flow of a river for the next day or the next few days. These forecasts are important, in particular, for dam managers who must decide whether to keep the water to produce electricity, or to release some water in order to avoid possible flooding. Examples of real hydrographs are presented in Fig. 1.

In previous papers, the author and Bensalma used a filtered Poisson process (see [7]) to model the flow $X(t)$ of a river at time t:

$$X(t) = \sum_{n=1}^{N(t)} Y_n \, e^{-(t-\tau_n)/c} \quad (X(t) = 0 \text{ if } N(t) = 0), \tag{1.1}$$

where the τ_n's are the arrival times of the events of the Poisson process $N(t)$. The times $T_n = \tau_n - \tau_{n-1}$ between the successive events are i.i.d. exponential random variables. Moreover, $Y_1, Y_2, \ldots$ are i.i.d. random variables, and are also independent of $N(t)$. Here, the Y_n's are taken to be exponentially distributed. See Fig. 2. This

Applied Sciences, Vol.19, 2017, pp. 76-83.

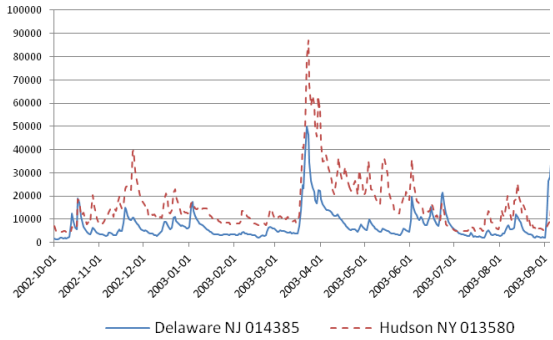

Figure 1: Examples of real hydrographs.

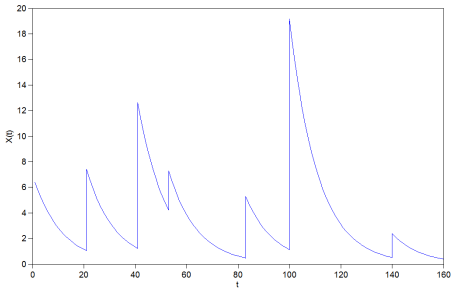

Figure 2: Example of a trajectory of a filtered Poisson process having the response function defined in (1.1).

type of stochastic process has also been used, in particular, by Kelman [2], Koch [3] and Konecny [4].

Next, in mathematical finance, diffusion processes with jumps have been widely used to model asset prices. A diffusion process with jumps $X_D(t)$ is defined by (see [1])

$$X_D(t) = D(t) + \sum_{n=1}^{N(t)} Y_n \quad \text{for } t \geq 0, \tag{1.2}$$

in which $D(t)$ is a diffusion process and $N(t)$ is a Poisson process (independent of $D(t)$), and where the Y_n's are defined as above. Thus, a compound Poisson process is added to a diffusion process. Fig. 3 presents an example of a (simplified) trajectory of a diffusion process with (positive) jumps.

Now, another important problem in statistical hydrology consists in estimating the probability that a certain flow will exceed a given threshold, either in the short or long term. This probability of threshold exceedance is needed, in particular, in hydrological risk analysis and in engineering design. In the case of the long-term probability of exceedance, people often estimate such a probability from historical data, making using of the concept of return period. However, according to some authors, this concept of return period must be reconsidered because of climate change (see [8]).

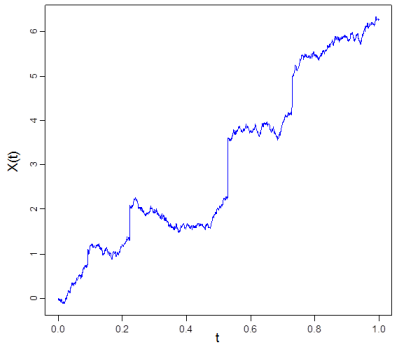

Figure 3: Simplified example of a trajectory of a diffusion process with positive jumps.

In Lefebvre and Bensalma [5], the objective was to estimate the short-term probability of threshold exceedance, namely, the probability of exceedance on the day following the current observed flow. They obtained estimates based on the two stochastic processes with jumps defined above. It turned out that the jump part of these processes was essential to obtain accurate forecasts.

2 Probability of exceeding a given threshold

2.1 Filtered Poisson process

In Lefebvre and Bensalma [5], the following formula for the value of the river flow at time $t+1$, given the observed flow at time t, based on the filtered Poisson process $X(t)$, was derived:

$$X(t+1) \mid \{X(t)=x\} = x\,e^{-1/c} + \sum_{n=N(t^+)}^{N(t+1)} Y_n\, e^{-(t+1-\tau_n)/c}. \tag{2.1}$$

Next, the authors obtained an approximate formula for

$$p(x,y) := P[X(t+1) \geq x+y \mid X(t)=x]. \tag{2.2}$$

They wanted to estimate this probability when x is large, so that if y too is large then there is a high risk of flooding. They found that

$$p(x,y) \simeq \lambda e^{-\lambda} \int_0^1 \exp\left\{-\mu\left[x\left(1-e^{-1/c}\right)+y\right] e^{(1-u)/c}\right\} du. \tag{2.3}$$

The various parameters that appear in the previous formula will be estimated in Section 3 in the case of the Delaware River. Moreover, the probabilities derived from Eq. (2.3) will be compared with the corresponding frequencies observed over an 11-year period, when the current flow x is (approximately) equal 10000 cubic feet per second. We fixed the value of x because, in particular, the parameter λ (the rate of the Poisson process) is likely to depend on x

2.2 Wiener process with jumps

Next, assume that $D(t)$ is a Wiener process with (negative) drift parameter θ and dispersion parameter σ^2. We can write that $D(t) \mid D(0)$ has a Gaussian distribution with mean $D(0) + \theta t$ and variance $\sigma^2 t$.

Let

$$p_D(x, y) := P[X_D(t+1) \geq x + y \mid X_D(t) = x]. \tag{2.4}$$

We find that

$$p_D(x, y) = e^{-\lambda} Q\left(\frac{y-\theta}{\sigma}\right) + \lambda e^{-\lambda} e^{-\mu y} \int_{-\infty}^{\infty} e^{\mu z} \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left\{-\frac{(z-\theta)^2}{2\sigma^2}\right\} dz, \tag{2.5}$$

where $Q(z) := P[\mathrm{N}(0,1) > z]$. Contrary to $p(x, y)$, this expression for $p_D(x, y)$ does not depend (explicitly) on x. Remember, however, that we assumed that the various model parameters depend on x.

2.3 Linear regression

Finally, for comparison purposes, the probability $p(x, y)$ will also be estimated by making use of linear regression. That is, we assume that

$$X_R(t+1) = a\, X_R(t) + b + \epsilon, \tag{2.6}$$

where $\epsilon \sim \mathrm{N}(0, \sigma^2)$. We can state that $X_R(t)$ is an autoregressive process of order 1, which is denoted by AR(1). It is a simple matter to obtain the following formula for the probability $p(x, y)$ based on this model:

$$p_R(x, y) = P[\mathrm{N}(a\,x + b, \sigma^2) \geq x + y] = Q\left(\frac{(1-a)x + y - b}{\sigma}\right). \tag{2.7}$$

3 Implementation of the various models

The three models defined in Section 2 will be used to estimate the probability that the flow of the Delaware River, which is an important river located in the United States, will increase by at least y cubic feet per second in one day, for various values of y, when its current value is equal to x.

The observed daily values of the flow of the Delaware River can be found on the Web site of the U.S. Geological Survey (http://nwis.waterdata.usgs.gov). The Montague station (no. 01438500) was chosen. The data set consists of the observed flow values over an 11-year period: from 23 January 2003 to 22 January 2014. Once the parameters have been estimated, the point estimates of the probabilities $p(x, y)$, $p_D(x, y)$ and $p_R(x, y)$ will be calculated and compared with the observed frequencies of the events considered in this paper.

As we mentioned above, we chose the value $x = 10000$. This flow value is already quite high. Moreover, if the flow increases, for instance, by (at least) 5000 cubic feet per second in one day, then the risk of flooding could be high. We could not choose a value of x much greater than 10000, because we need enough observations to obtain

reliable frequencies. In fact, we did not find enough observations for which x was exactly equal to 10000 ft^3/s in the data set. For this reason, we decided to include the 106 observations that belong to the interval $[10000, 11000)$.

Lefebvre and Bensalma [5] obtained the following point estimates:

$$\begin{aligned}\hat{p}(10000, y) &\simeq 0.264 e^{-0.264} \\ &\times \int_0^1 \exp\left\{-\left(\frac{1}{8465}\right)\left[10000\left(1 - e^{-1/7.72}\right) + y\right] e^{(1-u)/7.72}\right\} du, \end{aligned} \tag{3.1}$$

$$\begin{aligned}\hat{p}_D(10000, y) &\simeq e^{-0.264} Q\left(\frac{y + 1269.1}{662.6}\right) + 0.264 e^{-0.264} e^{-(1/7933)y} \\ &\times \int_{-\infty}^{\infty} e^{(1/7933)z} \frac{1}{\sqrt{2\pi}(662.6)} \exp\left\{-\frac{(z + 1269.1)^2}{2(662.6)^2}\right\} dz \end{aligned} \tag{3.2}$$

and

$$\hat{p}_R(10000, y) = Q\left(\frac{(-1.61)10000 + y + 16031}{6180.49}\right) = Q\left(\frac{y - 69}{6180.49}\right). \tag{3.3}$$

3.1 Numerical results

The various values derived from the previous formulas when $y = 5000$, 10000 and 20000 are presented in Table 1. The corresponding observed frequencies in the data set are also given in the table. Looking at the numbers in Table 1, we must conclude

Table 1: Numerical values of the various point estimates and corresponding observed frequencies

y	$\hat{p}(10000, y)$	$\hat{p}_D(10000, y)$	$\hat{p}_R(10000, y)$	Frequency
5000	0.0926	0.0923	0.2125	0.0943
10000	0.0493	0.0492	0.0540	0.0472
20000	0.0140	0.0140	0.0006	0.0189

that $\hat{p}(10000, y)$ and $\hat{p}_D(10000, y)$ are practically equal, which is rather surprising because the underlying models are quite different. We can also conclude that both $\hat{p}(10000, y)$ and $\hat{p}_D(10000, y)$ are good approximations to the corresponding frequencies. However, the autoregressive model was unable to provide accurate point estimates of the probabilities considered (if we compare these point estimates with the corresponding frequencies).

Remember that to compute the frequencies, all observed flow values that belong to the interval $[10000, 11000)$ were used. Therefore, if we could have considered only the observations that are exactly equal to 10000, the estimates $\hat{p}(10000, y)$ and $\hat{p}_D(10000, y)$ would probably have been even better.

4 Obtaining more accurate estimates

In order to improve the results presented in the previous section, one should try to incorporate exogenous variables into the models. Probably the most important such exogenous variable is precipitation. The problem is to determine exactly how to incorporate the amount of precipitation, observed or forecast, into a purely stochastic model. A possible answer is to appeal to linear regression to do so. The technique will be illustrated with the filtered Poisson process.

Let $P(t)$ be the amount of precipitation (in inches) observed at time t and $P(t+1)$ be the amount of precipitation forecast for time $t+1$. By making use of linear regression, we find that

$$X(t+1) = -364 + 0.965\,X(t) + 4102\,P(t) + 1224\,P(t+1).$$

Based on this equation, we can estimate the probability $p(10000, y)$ more accurately, as can be seen in the next tables.

Table 2: Values of $\hat{p}(10000, 5000, P(t), P(t+1))$, compared with $\hat{p}(10000, 5000) = 0.0926$ and with the frequency 0.0943

$P(t)$	$P(t+1)$	$\hat{p}(10000, 5000, P(t), P(t+1))$
0	0	0.0859
0	1	0.1007
1	0	0.1467
1	1	0.1721

Table 3: Values of $\hat{p}(10000, 10000, P(t), P(t+1))$, compared with $\hat{p}(10000, 10000) = 0.0493$ and with the frequency 0.0472

$P(t)$	$P(t+1)$	$\hat{p}(10000, 10000, P(t), P(t+1))$
0	0	0.0447
0	1	0.0525
1	0	0.0764
1	1	0.0896

Looking at these results, we can conclude that our estimates of the probability $p(10000, y)$ are much more precise when we take precipitation into account. In particular, we see that $\hat{p}(10000, y)$ is much higher when both $P(t)$ and $P(t+1)$ are positive. We also see that the amount of precipitation observed at time t increases the value of $\hat{p}(10000, y)$ more than the same amount of precipitation forecast for time $t+1$.

Finally, we can consider the variable $P(t)$ as deterministic. However, $P(t+1)$ is actually random. Indeed, the amount of precipitation forecast for time $t+1$ can obviously be wrong. Let us write

$$P(t+1) = \hat{P}(t+1) + E,$$

Table 4: Values of $\hat{p}(10000, 20000, P(t), P(t+1))$, compared with $\hat{p}(10000, 20000) = 0.0140$ and with the frequency 0.0189

$P(t)$	$P(t+1)$	$\hat{p}(10000, 20000, P(t), P(t+1))$
0	0	0.0122
0	1	0.0143
1	0	0.0207
1	1	0.0243

where $\hat{P}(t+1)$ is the actual forecast, and E is the forecasting error.

The author and Guilbault [6] worked on the problem of finding a suitable statistical distribution for the forecasting errors in the case of precipitation. Based on their work, we can state that, given that it will be positive, E has a certain lognormal distribution. This distribution can be used to obtain an explicit expression for $\hat{p}(x, y)$, given that $E > 0$.

It is particularly important to use this expression to estimate the risk that the river flow will exceed a given threshold on the next day, even if $\hat{P}(t+1)$ is equal to 0 or is very small. Indeed, it is possible that the actual amount of precipitation will be much larger than the forecast $\hat{P}(t+1)$. Taking the random nature of $\hat{P}(t+1)$ into account, one can be more prudent in estimating the risk of flooding.

5 Concluding remarks

In this note, we saw how the estimate of the probability that a river flow will exceed a given threshold can be made more precise. We can obtain even more accuracy if we take into account the forecasting error for the amount of precipitation at time $t+1$. Finally, we could add other exogenous variables to the stochastic models, such as the maximum and minimum temperatures observed at time t, and forecast for $t+1$.

Acknowledgements. This research was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC).

References

[1] R. Cont, P. Tankov, *Financial Modelling with Jump Processes*, Chapman & Hall, CRC Press, 2004.

[2] J. Kelman, *A stochastic model for daily streamflow*, J. Hydrol. 47 (1980), 235-249.

[3] R.W. Koch, *A stochastic streamflow model based on physical principles*, Water Resour. Res. 21 (1985), 545-553.

[4] F. Konecny, *On the shot-noise streamflow model and its applications*, Stoch. Hydrol. Hydraul. 6 (1992), 289-303.

[5] M. Lefebvre, F. Bensalma, *On the probability of exceeding a given river flow threshold*, Hydrolog. Sci. J. 61 (2016), 2205-2212.

[6] M. Lefebvre, L. Guilbault, *Statistical models for the errors in precipitation forecasts*, ARPN J. Eng. Appl. Sci. 2 (2007), 27-28 (online).

[7] E. Parzen, *Stochastic Processes*, Holden-Day, San Francisco, 1962.
[8] J. Salas, J. Obeysekera, *Revisiting the concepts of return period and risk for nonstationary hydrologic extreme events*, J. Hydrol. Eng. 19 (2014), 554-568.

Author's address:

Mario Lefebvre
Department of Mathematics and Industrial Engineering,
Polytechnique Montréal, C.P. 6079,
Succ. Centre-ville, Montréal, Québec H3C3A7, Canada.
E-mail: mlefebvre@polymtl.ca