VOL. 2, NO. 6, DECEMBER 2007 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences

www.arpnjournals.com

STATISTICAL MODELS FOR THE ERRORS IN PRECIPITATION FORECASTS

Mario Lefebvre[1] and Liliane Guilbault[1]
[1]Department of Mathematics and Industrial Engineering, École Polytechnique, Montréal, Québec, Canada
E-mail: mlefebvre@polymtl.ca

ABSTRACT

We consider the problem of finding statistical models for the absolute value of the errors made by the *Hydro-Québec* company in their precipitation forecasts (based on the ones produced by *Environment Canada*) for the current day. We find that a generalised Pareto distribution provides a very good fit to the data. We also divide the data set into various cases, in particular the case when at least 10mm of rain (or water equivalent of snow) were forecasted.

Keywords: precipitation, forecast, model, pareto distribution, lognormal distribution, goodness-of-fit test, *p*-value.

INTRODUCTION

People responsible for the management of large dams rely on accurate temperature and precipitation forecasts to make important decisions concerning the stocking or releasing of water, given the impact of these decisions both upstream and downstream. In particular, managers must endeavour to avoid possible flooding. We analyse the errors in precipitation forecasts for the Brotkord Station of the Gatineau River basin, in Canada. The data cover the one-year time period from 29 May 2003 to 28 May 2004.

Up to now, few researchers have been interested in modeling the errors generated by precipitation forecasts (Accadia *et al.*, 2003 and Johnson and Olsen 1998). Actually, as shown by some authors, including McBride and Ebert (2000), the verification of precipitation forecasts is already a difficult task. As mentioned by Déqué (2003) of the Centre National de Recherches Météorologiques (CNRM) in France, the unpredictable character of atmospheric movements implies that the forecasts can only be made for a few days in advance. To measure the success of a forecast, Déqué mentions the use of a spatio-temporal correlation coefficient between the forecasts and the observations. Within a training course on climate simulators in a research center, Rousseau (2004) tried to model the forecasting errors of Météo France by making use of linear regression models to determine temporal links between the forecasting errors. In our case, we will instead find statistical models that fit the data well in various situations.

In the present study, we will look for statistical models by performing goodness-of-fit tests of these models to the data. We will propose a general statistical model, as well as more specific models, which all fit well the observations for the station selected, for various subsets of the data set.

RESULTS AND DISCUSSION

We define the variables

W = difference between the observed and forecasted precipitation for the current day;
X = absolute value of W;
$Y = X$, given that $X > 0.2$ mm;
$Z = Y - 0.2$.

First, we would like to find an acceptable model for the random variable X for the entire data set that consists of 228 observations (there were some missing data). However, because the data representing the observed precipitation provided by *Hydro-Québec* were rounded to 0.2 mm when they actually were in the interval (0, 0.2], we consider instead the random variable Y, which denotes the *significant* errors, namely the ones larger than 0.2 mm. Moreover, to obtain a variable defined in the interval $(0, \infty)$, we must subtract 0.2 from Y.

As a model for Z, we propose a generalised Pareto distribution. That is, we suppose that

(1) $$f_Z(z; c, \theta) = c(\theta - 2)(1 + cz)^{1-\theta}$$

for $z > 0$. To estimate the unknown parameters c (> 0) and θ (> 2), we can use the method of moments:

(2) $$\hat{\theta} = 2 + \frac{2s_Z^2}{s_Z^2 - \bar{z}^2} \quad \text{and} \quad \hat{c} = \frac{1}{\bar{z}(\hat{\theta} - 3)}.$$

Notice that the standard deviation s_Z of the observations must be larger than their mean $\bar{z}$, and $\hat{\theta}$ must be larger than 3 for the method to apply. From the 228 observations, we obtain that $\hat{\theta} = 6.74$ and $\hat{c} = 0.038$.

We performed a Pearson's chi-square goodness-of-fit test (see Hines and Montgomery 1990, for example) of this distribution to the data. We obtained a p-value equal to 0.554, which is very good. Furthermore, because the test takes for granted that the observations constitute a particular random sample of Z, we also conducted the same test with a subset of almost uncorrelated observations, and the model was again accepted with a large enough p-value.

Next, since the forecasting error must surely be a function of the amount p_f of forecasted precipitation, we will consider four cases separately, namely the cases when i) $p_f = 0$, ii) $0 < p_f < 5$, iii) $5 \le p_f < 10$ and iv) $p_f \ge 10$. To justify the assertion made above, we give the mean and the standard deviation (in mm) of the observations of X in the various cases considered:

- all observations: $\bar{x} = 5.18$; $s_X = 8.49$;

- forecast $p_f = 0$: $\bar{x} = 2.28$; $s_X = 6.25$;
- $0 < p_f < 5$: $\bar{x} = 3.78$; $s_X = 6.98$;
- $5 \leq p_f < 10$: $\bar{x} = 8.58$; $s_X = 10.51$;
- $p_f \geq 10$: $\bar{x} = 15.69$; $s_X = 8.81$.

a. During the time period considered, 69 times (out of 228) no precipitation at all was forecasted; 37 times, there indeed was no precipitation, while 10 times the observed precipitation was in the interval (0, 0.2]. We performed the Anderson-Darling normality test (available with the statistical software *MINITAB*) for the logarithm of the absolute values of the remaining 22 forecasting errors (minus 0.2). We obtained an excellent *p*-value of 0.748.

b. For the case when $0 < p_f < 5$, there were 108 data points. The forecasting errors were never equal to zero, but 19 times they were in the interval (0, 0.2]. A chi-square goodness-of-fit test performed with the absolute values of the significant errors showed that, as in the general case, a generalised Pareto distribution is an appropriate model for *Z*. The *p*–value is 0.303. Because it is particularly important to evaluate the risk of a thunderstorm (or a snowstorm) when the forecast calls for a large amount of precipitation, the cases when 5 to 10 mm of rain were forecasted, and that for which more than 10 mm of rain were forecasted are really crucial.

c. 25 times we had $0 < p_f < 5$. Each time, the forecasting error was larger than 0.2mm. We calculated the absolute values of the errors (minus 0.2) and conducted an Anderson-Darling normality test. With a *p*-value of 0.204, the Gaussian distribution is acceptable in that case.

d. Finally, there were also 25 days for which $p_f \geq 10$. All forecasting errors were larger than 0.2 mm. This time, we found that a lognormal distribution provides a truly excellent fit to the observations of the random variable *Z,* since the *p*-value is 0.980.

In summary, the various distributions proposed for *Z* are the following:

- all observations: generalised Pareto distribution;
- forecast pf = 0: lognormal distribution;
- 0 < pf < 5: generalised Pareto distribution;
- 5 ≤ pf < 10: Gaussian distribution;
- pf ≥ 10: lognormal distribution.

CONCLUSIONS

We considered other stations as well as forecasts for one and two days ahead. Moreover, we also studied the summer and winter seasons separately. The same conclusions as above were obtained in most cases.

We had to discard some stations because the observations of *W* were not reliable. In the case of the Brotkord Station, there were also outliers. For example, we had to eliminate a forecasting error equal to more than 58mm, which is hardly possible.

Except when $0 < p_f < 5$, there were few data points in the particular cases considered. Consequently, it would be interesting to perform the statistical tests again when new observations become available.

A related problem consists in finding statistical models for the errors made in forecasting temperature. Although this problem seems relatively easy, the classical models, such as the Gaussian or the Laplacian distribution, do not appear to be appropriate.

ACKNOWLEDGEMENTS

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

REFERENCES

Accadia C., Casaioli M., Mariani S., Lavagnini A., Speranza A., De Venere A., Inghilesi R., Ferretti R., Paolucci T., Cesari D., Patruno P., Boni G., Bovo S., and Cremonini R. 2003. Application of a statistical methodology for limited area model intercomparison using a bootstrap technique. Il Nuovo Cimento *C*, 26: 61-77.

Déqué, M. 2003. La prévision numérique à l'échelle saisonnière: que sait-on faire et que peut-on espérer ? La Météorologie, 41: 20-29.

Hines W.W., and Montgomery D.C. 1990. Probability and statistics in engineering and management science, 3rd Edition, Wiley, New York.

Johnson L.E. and Olsen B.G. 1998. Assessment of quantitative precipitation forecasts. Weather and Forecasting, 13: 75-83.

McBride J.L. and Ebert E.E. 2000. Verification of quantitative precipitation forecasts from operational numerical weather prediction models over Australia. Weather and Forecasting, 15: 103-121.

Rousseau V. 2004. Étude d'un simulateur de climat-Modélisation statistique des erreurs de prévisions de Météo France, 62 pages. Retrieved August 28, 2007 from http://www.inra.fr/internet /Departements/MIA/T//garcia/Anticipation/AvancementProjet/Stages/2004/Vanessa/rapportVanessa2.pdf