Pergamon

Wat. Sci. Tech. Vol. 37, No. 11, pp. 47–55, 1998.
© 1998 IAWQ. Published by Elsevier Science Ltd.
Printed in Great Britain. All rights reserved
0273–1223/98 $19·00 + 0·00

PII: S0273-1223(98)00315-1

FORMULATING AND TESTING A RAIN SERIES GENERATOR BASED ON TIPPING BUCKET GAUGES

Karsten Arnbjerg-Nielsen*, Henrik Madsen** and Poul Harremoës*

* *Department of Environmental Science and Engineering, Building 115, Danish Technical University, DK-2800 Lyngby, Denmark*
** *Department of Mathematical Modelling, Building 321, Danish Technical University, DK-2800 Lyngby, Denmark*

ABSTRACT

Changing needs of the society also change the goals of hydrological calculations. The "design storm" concept cannot provide adequate answers to the complex problems on many time scales that society is facing today. This paper present a novel approach to generate artificial rain data with a time resolution sufficient for urban hydrology. The model is based on Wolds process of intervals, ie. it generates waiting times observed at a tipping bucket gauge conditional on the previous waiting time. The model is verified both with respect to average and extreme values and by testing it on urban drainage problems and it is concluded that the model appears to have captured to properties of rainfall well. © 1998 IAWQ. Published by Elsevier Science Ltd. All rights reserved

KEYWORDS

Markov chains; rainfall generator; simulation; urban hydrology; statistics; prediction.

INTRODUCTION

Occurrence of rainfall interferes with many basic functions of society. Changing needs of society also change the goals of hydrological calculations. When the main objective was simple design of conduits in towns, simple approximations such as the "design storm" concept proved to be sufficient.

The recent shift of focus to sustainability and environmental protection requires that larger and more complex systems are analyzed and thus the need for long and accurate historical rain series arise. The use of single design storms is too simplistic an approach; the detrimental effects may occur on many different time scales. The variation in the rain input must reflect what has been observed historically, from peak intensities lasting only a few minutes right down to variations in annual precipitation. Thus historical rain series have become necessary in the analysis of urban hydrology.

Many historical rain series are too short to enable reliable estimates of extreme values of detrimental effects. In this paper a stochastic model which can be used for generating artificial rainfall data is presented. The model can be used to generate rain events which can be seen as an extrapolation of the historical events. The focus is on establishing a rainfall generator that has the following properties: First of all it must have a Markovian structure, meaning that predictions can be constructed based on recent rain information. Thus the same model structure can be used to make long artificial time series and to make on-line predictions of the rain intensity. Secondly, it must retain the statistical properties of rainfall on all relevant time scales in urban hydrology. Following the above discussion, the statistical properties must be retained on time scales from minutes to years. Third, but not least, it should be able to use data recorded by means of automatic tipping bucket gauges, since these gauges have become the de facto standard rain gauge for urban hydrology applications.

MODEL IDENTIFICATION

Recently several rain series generators have been presented. Most of these are based on Neymann-Scott type models (Cowpertwait, 1991; Onof *et al.*, 1994). These models are most frequently based on hourly or even daily data and in order to obtain a suitable time scale the generated series have been disaggregated. However, this approach has proved to be less adequate for urban hydrology studies due to a substantial bias in the simulated series (Foundation for Water Research, 1992a, 1992b). Finally, these models have a non-Markovian structure and these are thus not useful for making short term predictions.

Markov chain models

Mathematically and computationally Markov models are simpler than the Neymann-Scott type models. Chapman (1994) cites more than 15 applications to rainfall data of Markov chain or Markov renewal models, the majority being Markov chain models on daily rainfall.

Two studies have been reported using high resolution rain data. Srikanthan and McMahon (1983) have estimated a Markov chain model based on Australian data with a resolution of six minutes, and conclude that the model performs satisfactorily. Their model consists of a two state Markov chain governing transitions between dry and wet days. On wet days a second-order non-stationary Markov chain dependent on the type of wet day is used to generate hourly data. For wet hours another set of Markov chains are used, depending on the type of wet hour and the month. The generated intensities are adjusted to obtain the observed hourly totals. The Markov chain approach is also used by Tan and Lee (1993), but based on the waiting times of a tipping bucket gauge. They use a matrix with 30 states. The states are chosen by using constant increments on a log scale. They conclude that a simulated series complies reasonably well with the original one. However, there is a tendency for the artificial series to give more heavy and shorter rainfalls.

Wold's process of intervals

There are two major problems with the Markov chain approach. First, the method uses a vast number of parameters. The model presented by Tan and Lee used 870 parameters. This large number prohibits parametric modelling of regional variation. Secondly, the parameters cannot be related to the physical properties of rainfall. However, the only successful studies of high resolution rainfall have been based on Markov chain models.

A traditional rain series generator based on a Markov chain is obtained by defining a number of states, based on an allocation of the rain intensities in a number of intervals. Then the probability of jumping from a state i to a state j, $p_{i,j}$, is estimated based on available data. The probabilities satisfy $\Sigma_j p_{ij} = 1$, and the probabilities belonging to a row of the transition matrix is thus a discrete (conditional) probability density function.

In this paper a new method is proposed which uses the fact that the sample space for the waiting times is $\Re_+$, and the discretization of the conditional density is thus avoided. Let w_{k-1} denote the waiting time number k-1 in the sequence of waiting times. The density function for the next waiting time, w_k, can be modelled conditional on the previous waiting time w_{k-1}. This type of process is known as Wold's process of intervals, due to the author who first gave a description of such a process (Wold, 1948). The process is described thoroughly by Lewis (1972).

Identification of a suitable conditional distribution function

The choice of the conditional distribution function to represent the waiting time is by no means trivial. The objective is to identify parameters which can be related to the physical properties of rainfall. By trial and error it is found, that three types of conditional densities are sufficient for adequately describing the properties. These types can be labelled corresponding to convective rain, frontal rain and no rain. Each type is modelled by a two parameter distribution function. Weighting parameters are used to identify the relative proportion of each rainfall type.

The conditional density function for w_k given the previous waiting time, w_{k-1}, is given in Eq. (1). The parameter α_c describes the probability that the next waiting time is best described by the convective rain type and α_f describes the probability that the next waiting time is best described by the frontal rain type. The probability that the next waiting time is best described by the non-rain type is thus $1 - \alpha_c - \alpha_f$. A log-Gumbel distribution, $f_1(w;\ k,\ b)$, is found to fit the convective storms well, whereas the log-normal distribution gives a good fit both to the frontal storms, $f_2(w;\ \mu_f,\ \sigma_f)$, and to the non-rainy periods, $f_3(\ w;\ \mu_n,\ \sigma_n)$. The log-Gumbel distribution arises as a special case of extreme value distribution, and the log-normal distribution as a special case of the Box-Cox transformation. In some cases it is difficult to estimate the values of all eight parameters due to optimization problems. Therefore one of the parameters in the convective rain region, k, is fixed for $w \geq 1/3$. Also, the log-normal density function in the non-rainy part of the matrix is defined only for $w > r$, where $r = 10$ is found by trial and error. For w smaller than 1/3 the probability of shifting to a type of rainfall other than the convective area is found to be close to zero and therefore only the convective region is fitted to these states. The model thus reads

$$f_{w_k|w_{k-1}}(w) = \begin{cases} f_1(w) & ;\ w_{k-1} < 1/3 \\ \alpha_c f_1(w) + \alpha_f f_2(w) & \\ \quad +(1-\alpha_c-\alpha_f) f_3(w) & ;\ w_{k-1} \geq 1/3 \end{cases} \tag{1}$$

where

$$f_1(w;\ k,b) = \begin{cases} 0 & ;\ w \leq 0 \\ \dfrac{k}{b}\left(\dfrac{b}{w}\right)^{k+1} e^{-\left(\frac{b}{w}\right)^{k}} & ;\ w > 0 \end{cases}$$

$$f_2(w;\ \mu_f,\sigma_f) = \begin{cases} 0 & ;\ w \leq 0 \\ \dfrac{1}{\sqrt{2\ \pi\ \sigma_f^2}}\, e^{-\frac{1}{2}\left(\frac{\ln(w)-\mu_f}{\sigma_f}\right)^2} & ;\ w > 0 \end{cases}$$

$$f_3(w;\ \mu_n,\sigma_n) = \begin{cases} 0 & ;\ w \leq r \\ \dfrac{1}{\sqrt{2\ \pi\ \sigma_n^2}}\, e^{-\frac{1}{2}\left(\frac{\ln(t-r)-\mu_n}{\sigma_n}\right)^2} & ;\ w > r \end{cases}$$

Using this model formulation very few parameters are used compared to the traditional Markov chain approach, see Figure 1. What is even more important is the fact that each of the parameters of the model can be interpreted physically. This is shown in Figure 2, where the estimated parameters for one rain series are shown. Although only one parameter, k_l, has been fixed, the other parameters develop smoothly and intuitively correct. Note, that for w between 0.3 and 60 minutes two conditional distribution functions are fitted. One corresponds to rainfall with accumulated volumes larger than 3 mm and the other corresponds to rains that do not (yet) contain 3 mm rain. Note, that especially the probability of staying in the frontal region is much larger if the threshold of 3 mm is exceeded. This incorporates the physical knowledge that, given that the rain event contains much volume the chance that the rain event will stop now is lower than average.

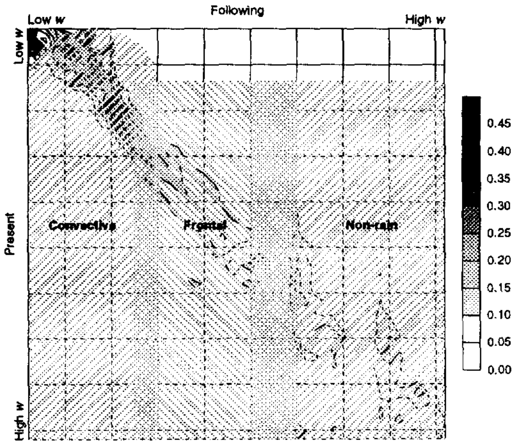

Figure 1. Identification of physically related distributions compared to the Markov chain approach.

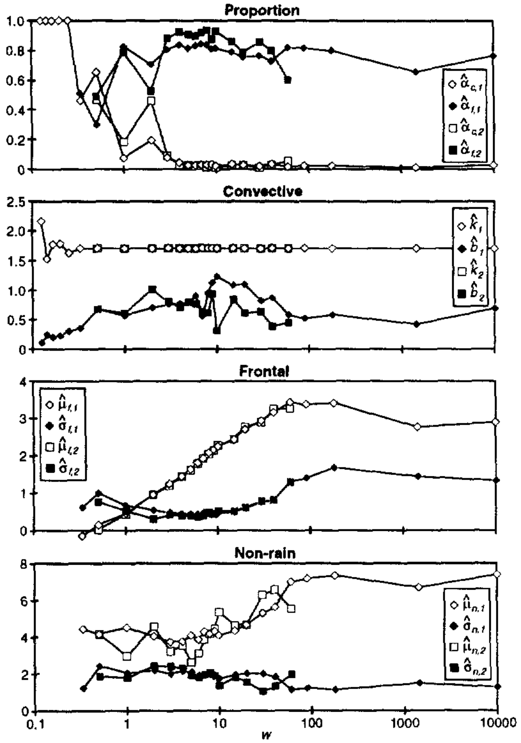

Figure 2. The parameter estimates for gauge 20211. Subscripts 1 and 2 refer to the conditional distribution functions being below and above the threshold value of 3 mm, respectively.

TESTING THE MODEL

It is important to identify the proper test statistics used to compare the artificial rain series with the data. The test statistics should relate to and encompass all the different aspects of urban storm drainage. Since the model is estimated based on the average properties of rainfall the test statistics should primarily focus on extreme values.

Extreme values of test statistics

Upstream peak flows in small and large urban catch ment can be evaluated by means of extreme values of maximum average 10 and 60 minute intensities. This analysis corresponds to the traditional intensity-duration-frequency curve principle. The two statistics check convective rain and heavy frontal storms. In urban hydrology common design return periods are from one to five years.

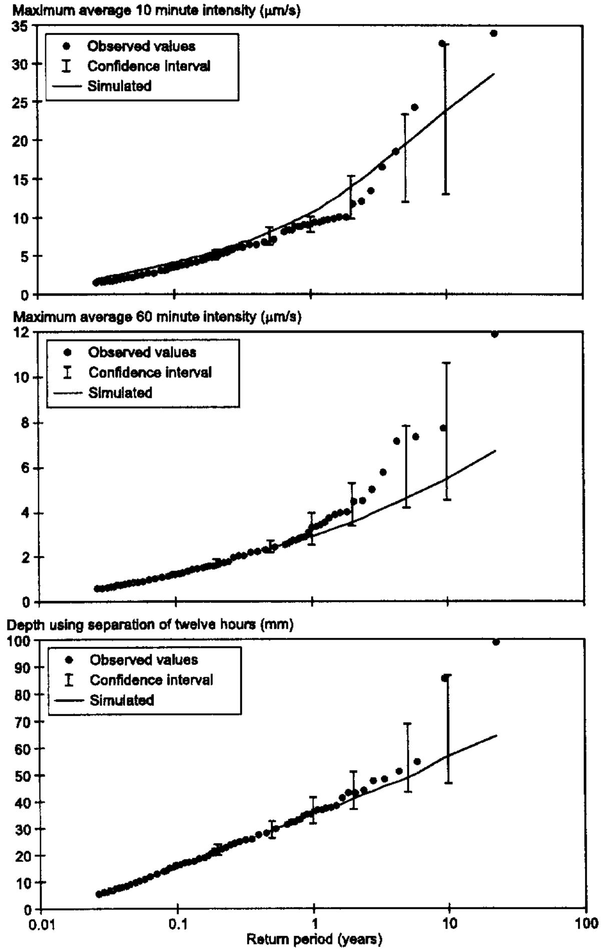

Figure 3. Extreme values of the observed and the simulated series at gauge 20211. The simulated series are the average of 100 simulations using the same number of transitions as the original series. The vertical lines indicate confidence limits estimated by means of a bootstrap estimation method (Arnbjerg-Nielsen *et al.*, 1994) for small return periods and a Partial Duration Series model (Madsen *et al.*, 1997) for the larger return periods.

Total volumes of rain are important when evaluating detention ponds with small interceptor capacities. There is, however, a conflict in the assumption of a negligible interceptor capacity and a separation between events of only one hour; there is a rather large probability of a new event starting before the detention pond is empty. A rough estimate of typical detention ponds and interceptor capacities shows that events should be separated by approximately 12 hours and that the relevant interceptor capacities could be neglected. The fourth test statistic is thus extreme depths using a separation between events of 12 hours. Meteorologically, the third and fourth statistics check frontal storms and intermittent rain. Roughly, common design return periods are from ten times per year to a few years in urban hydrology.

The fifth test statistic is the annual average precipitation. When making long-term simulations and/or generating series longer than the observed one it is important that a summary statistic such as the annual average precipitation is also in accordance with the observed series.

As can be seen on Figure 3 the model perform well for all of these test statistics. The simulated annual average precipitation is 638 mm which can be related to the observed value of 646 mm.

Testing on urban catchment

Since the rainfall generator is intended for use in urban hydrology applications the best test of the performance of the artificial series is to route the artificial and original rain series through a model of test catchment to study their relative performance. A real catchment from Virum in Copenhagen is chosen for the analysis. The catchment has a total area of 42.1 ha and the paved area is 13.55 ha. An analysis of the uncertainty of various detrimental effects from the catchment has previously been published by Arnbjerg-Nielsen and Harremoës (1996). A plot of the catchment is shown in Figure 4.

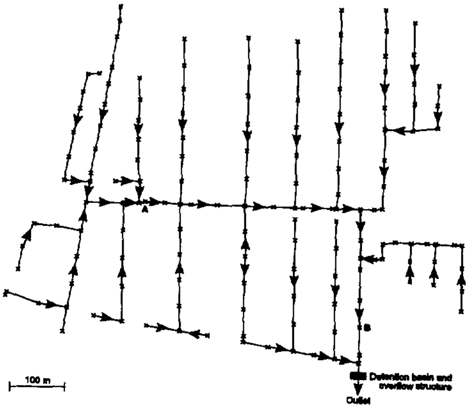

Figure 4. Sewer pipes in the catchment in Virum, Copenhagen. The paved area is 13.55 ha. The overflow structure and the detention pond at the outlet of the system have been designed using a criterion of a maximum of ten discharges per year.

The choice of detrimental effects is chosen to be the same as in Arnbjerg-Nielsen and Harremoës, since the overall uncertainties are known for these effects. The reasons for choosing these effects are discussed in detail in their paper. The detrimental effects are: Annual number of CSOs, yearly discharges of phosphorus loads and flooding of manholes. The calculations are summarized in Table 1.

Yearly discharges

The calculated annual number of overflows from CSO structures from the artificial series is close to the number calculated using the original series. There is a tendency for the calculated yearly discharges to be smaller when based on the artificially generated series rather than the observed series. However, following Arnberg-Nielsen and Harremoës a 90% central confidence interval for the calculated annual phosphorus load is [9 kg/yr, 27 kg/yr]. The calculated phosphorus load using the simulated rain series is thus well within this interval.

Table 1. Detrimental effects calculated by means of the Odense rain series and the observed and an artificial series from gauge 20211

Rain series	20211		Odense
	observed	artificial	
Annual number of overflows (yr^{-1})	14.4	14.1	10.6
Yearly discharges, Phosphorus loads ($kg\ yr^{-1}$)	18.1	14.8	12.8
Volume per year ($m^3\ yr^{-1}$)	6647	5449	4750
Flooding, manhole A (m)	-0.64	-0.30	-0.29
Flooding, manhole B (m)	-0.44	-0.16	-0.08

Closer examination of the calculated results indicates that the smaller loads are due to too small volumes of the extreme occurences of CSO. This means that a few rains with very high volumes are not generated in the simulation scheme. This finding is supported by the plots shown in Figure 3.

Flooding in subcatchments

Calculation of the water level in the sewer system in subcatchments is very uncertain when the pipe capacity is exceeded. The water level with a return period of one year is calculated at two manholes using both the artificial series and the observed series. The confidence limits for this level at these manholes includes both flooding in the catchment as well as water levels below the top of the pipe. In particular, the plot shows that there is little difference between calculating exceedances of the top of the pipe and calculating flooding.

The water level calculated when using the artificial rain series is close to the water level calculated using the observed rain series, see Table 1. Both models predict a level between the top of the pipe and the surface of the catchment. The water level of all the tested rain series are classified into the categories: 1) flooding, 2) top of pipe exceeded, and 3) top of pipe not exceeded. For all rain series, the observed and the simulated rain series yield water levels in the same category.

FURTHER WORK

The model presented here is still under construction and the flexible formulation of the model almost guarantees that a better performance can be obtained in the future. This section is dedicated to suggestions of further testing of the model, and to the discussion of its advantages and possible improvements that can be implemented.

Previous work suggest that extreme rainfall can be divided into high intensity rain storms with short duration and long duration rainfall with heavy volume, but low intensities. Furthermore, a seasonal variation was indicated (Arnbjerg-Nielsen, 1996). The artificial rainfall series should be tested to see if the first property is retained in the present formulation of the model. Also, the seasonal variation is not included in the present formulation of the model. This can, however, be established by assuming an annual variation of the parameters.

The artificial rain series generate the same annual number of overflows from CSO structures as the observed series. However, they presently underestimate annual loads, due to a lack of ability to generate artificial rain events with very extreme volumes, i.e. the same problem as observed in the UK with the Neymann-Scott type model. The solution in the UK is to build the detention ponds somewhat larger than calculations using the artificial rain series suggest. The presented model in its present form is assumed to be well suited for urban hydrology applications and the model can replace the use of historical rain data from other regions, if a short rain series is measured at a gauge in the region.

There is a need for further parameter reduction in the model. Presently the parameters of each state are estimated separately, i.e. the model uses 269 parameters. The number of parameters can almost certainly be reduced by considering how they vary between different states. The use of fewer parameters has two major advantages: 1) it will produce more reliable estimates and 2) make regional modelling more easy.

The model can perhaps be extended to generate the high volume rain events and thus improve the description of annual loads from CSO structures. The problem is probably overcome by studying different model dependencies on the accumulator variable. A first approach can be to introduce a second threshold on the accumulator variable, i.e. to have three types of rainfall, separated by accumulated volumes of, say, 3 and 10 mm. Probably the only parameter with a significant variation between the second and third state will be the probability of staying in the frontal rain type, α_f. The extension will probably produce larger volumes for high return periods.

CONCLUSIONS

Markov chain models has been used with success to generate artificial rain series with properties like the original series. The vast number of parameters in the Markov chain model is reduced by formulating an equivalent model based on Wold's process of intervals. The model predicts consecutive waiting times between tips using a conditional distribution function with eight parameters in total. The parameters in the distribution function are related to the physical processes generating the rainfall and vary depending on the present state of the rain. The state of the rain is described by the last observed waiting time and the accumulated volume of the present rain event. In the current formulation the model uses 269 parameters. Suggestions for reducing the number of parameters are given in the paper.

The artificial rain series are tested using test statistics applied directly on the rain series, as well as comparing the calculated detrimental effects when using an artificial series and the corresponding observed series. Events from an artificial rain series with a return period exceeding two years generally have lower volume than events from the observed series. The difference in volume may be due to the limited observation period for each of the individual series. However, the volume deficit is present at all the tested gauges and, therefore, it is believed that the extreme volumes are due to a physical phenomenon rather than just a limited observation period. The volume deficit in simulated rain series is most apparent in the Copenhagen region.

The number of parameters has been reduced sufficiently to generate long artificial series based on an observed series from a gauge, since one year of measurements leads to approximately 3,500 - 4,000 observations. However, for extrapolation to ungauged sites, the number of parameters should be reduced. The reduction can be obtained with the present model formulation and the approach is outlined. A method to improve the ability to describe extreme volume rainfall is also suggested.

LITERATURE

Arnbjerg-Nielsen, K. (1996). Statistical analysis of urban hydrology with special emphasis on rainfall modelling. PhD-thesis. Department of Environmental Science and Engineering, Danish Technical University, Lyngby, Denmark.

Arnbjerg-Nielsen, K. and Harremoës, P. (1996). The importance of inherent uncertainties in state-of-the-art urban storm drainage modelling for ungauged small catchments. *Journal of Hydrology*, **179**(1-4), 304-319.

Chapman, T. (1994). Stochastic models for daily rainfall. *25th Congress of the International Association of Hydrogeologists/International Hydrology and Water Resources Symposium of the Institution of Engineers.*The Institution of Engineers, Adelaide, Australia, pp. 7-12.

Cowpertwait, P. S. P. (1991). Further developments of the Neymann-Scott clustered point process for modelling rainfall. *Water Resources Research*, **27**(7), 1431-1438.

Foundation for Water Research (1992a). *Sewer Flooding Risk - Final Report*, Report FR0309. Foundation for Water Research, Bucks, UK.

Foundation for Water Research (1992b). *Testing of the Stochastic Rainfall Generator (SRG) model: Supplementary Report*, Report FR0318. Foundation for Water Research, Bucks, UK.

Lewis, P. A. W. (1972). Stochastic point processes: statistical analysis, theory, and applications. In: *Papers presented at a conference held at the IBM Research Center, Yorktown Heights, N. Y. 1971*. Wiley - Interscience, New York, USA.

Madsen, H., Rasmussen, P. F. and Rosbjerg, D. (1997). Comparison of annual maximum series and parital duration series methods for modeling extreme hydrologic events. *Water Resources Research*, **33**(4), 747-757.

Onof, C., Wheater, H. S. and Isham, V. (1994). Note on the analytical expression of the inter-event time characteristics for Bartlett-Lewis type rainfall models. *Journal of Hydrology*, **157**, 197-210.

Srikanthan, R. and McMahon, T. A. (1983). Sequential generation of short time-interval rainfall data. *Nordic Hydrology*, **1983**, 277-306.

Wold, H. (1948). On stationary point processes and Markov chains. *Skandinavisk Aktuarietidskrift*, **31**, 229-240.