Appropriate spatial sampling of rainfall for flow simulation

XIAOHUA DONG, C. MARJOLEIN DOHMEN-JANSSEN & MARTIJN J. BOOIJ

Water Engineering & Management, Faculty of Engineering Technology, University of Twente, PO Box 217, NL-7500 AE Enschede, The Netherlands

x.h.dong@ctw.utwente.nl

Abstract The objective of this study is to find the appropriate number and location of raingauges for a river basin for flow simulation by using statistical analyses and hydrological modelling. First, a statistical method is used to identify the appropriate number of raingauges. Herein the effect of the number of raingauges on the cross-correlation coefficient between areally averaged rainfall and discharge is investigated. Second, a lumped HBV model is used to investigate the effect of the number of raingauges on hydrological modelling performance. The Qingjiang River basin with 26 raingauges in China is used for a case study. The results show that both cross-correlation coefficient and modelling performance increase hyperbolically, and level off after five raingauges (therefore identified to be the appropriate number of raingauges) for this basin. The geographical locations of raingauges which give the best and worst hydrological modelling performance are identified, which shows that there is a strong dependence on the local geographical and climatic patterns.

Key words flow simulation; HBV model; precipitation; raingauge network design; spatial sampling

Echantillonnage spatial de la pluie approprié pour la simulation d'écoulements

Résumé L'objectif de cette étude est de trouver le nombre et les localisations de pluviomètres au sein d'un basin versant appropriés pour simuler l'écoulement, en s'appuyant sur des analyses statistiques et de la modélisation hydrologique. Tout d'abord, une méthode statistique est utilisée pour déterminer le nombre approprié de pluviomètres. Pour cela, l'effet du nombre de pluviomètres sur le coefficient de corrélation croisée entre la moyenne spatiale des précipitations et le débit est analysé. Puis le modèle global HBV est utilisé pour étudier l'influence du nombre de pluviomètres sur les performances de la modélisation hydrologique. Le bassin de la Rivière Qingjiang, comportant 26 pluviomètres, est utilisé comme cas d'étude. Les résultats montrent que le coefficient de corrélation croisée, ainsi que la performance de la modélisation, augmentent de manière hyperbolique, jusqu'à une valeur asymptotique au-delà de cinq pluviomètres (qui est donc considéré comme étant le nombre approprié de pluviomètres) pour ce bassin. Les localisations géographiques des pluviomètres qui donnent les meilleures et les pires performances de modélisation sont identifiées, ce qui montre qu'il existe une forte dépendance par rapport aux caractéristiques géographiques et climatiques locales.

Mots clefs simulation d'écoulement; modèle HBV; précipitation; définition de réseau de pluviomètres; échantillonnage spatial

INTRODUCTION

As the understanding of the physical principles behind the hydrological processes related to flow simulation practices goes deeper and becomes more thorough, hydrological models become even more sophisticated and therefore demand large rainfall data sets as input. New technologies are developed in order to obtain more distributed data (both spatially and temporarily), such as satellite imaging and weather radar remote sensing, to meet the requirements of these advanced hydrological models. But

the question remains: does one really need such complicated models? As a consequence of using these models, is it really necessary to set up expensive data acquisition systems to obtain more detailed data to feed them? Even though these models possibly improve the flow simulation results, one may still doubt if it is worth the expense. In practical hydrological applications, compromises have to be made to the existing model and data collection system. The reality is that most rainfall recording systems in use are still point-measuring raingauges. This limits the river basin manager's choice of models, and the lumped and semi-distributed models are still the most prevailing ones. Therefore, it is still useful to know what are the appropriate rainfall data for such a model. In this study, the appropriate number and location of raingauges for a lumped HBV model (SMHI, 2003) is investigated.

The methodology for determining the appropriate spatial sampling strategy of rainfall depends on pre-existing conditions of raingauge network in the river basin: (a) ungauged, (b) gauged with not enough raingauges, and (c) a dense network exceeding the requirement. The methodology presented here deals with the third condition where rational network reduction is necessary. This is achieved in two steps. The first step is based solely on the statistical analysis of recorded rainfall and discharge data. The statistical characteristics analysed here are: (i) variance of areally averaged rainfall and (ii) cross-correlation of areally averaged rainfall and discharge. Their relationships to the number of raingauges are explored. The studying of the variance of areally averaged rainfall refers to the "variance reduction" phenomenon reported by Yevjevich (1972). The authors extended the idea to study the effect of the number of raingauges on the cross-correlation between areally averaged rainfall and discharge, deduced the theoretical relationship between the cross-correlation coefficient and the number of raingauges, and expected that the increased cross-correlation coefficient between areally averaged rainfall and discharge will improve the performance of hydrological model for flow simulation. Therefore, the second step is to verify the idea obtained in the previous step by applying the HBV physics-based hydrological model.

The objective of spatial rainfall network design for flow simulation is to determine the effect of spatial rainfall sampling (both the number and locations) on the uncertainty of estimated precipitation or on hydrological variables computed from estimated precipitation series (Bras *et al.*, 1988). So far, this objective has been mainly achieved through one of two approaches: (a) theoretical modelling of rainfall processes, or (b) use of real rainfall data observed from raingauge networks or weather radar.

The general idea of the first approach is, first of all, to derive the statistical characteristics of rainfall patterns of the river basin studied. Then, a stochastic rainfall model is constructed based on the derived statistics to create synthesized stochastic rainfall series retaining the same statistical features as the real rainfall regime. Finally, different raingauge network scenarios are used to sample the synthesized rainfall fields to investigate the sampling effects on the uncertainty of rainfall estimates and hydrological variables (usually flow rates) computed from the rainfall estimates.

Here, some examples of the first approach are presented. Krajewski *et al.* (1991) and Azimi-Zonooz *et al.* (1989) used a Monte Carlo method to study the rainfall sampling effect on the basin response using a distributed catchment model. A space–time stochastic model was built to generate synthetic rainfall data, which were consequently sampled by synthetic raingauge networks at varying densities. Rainfall data sampled from a hypothetical scenario with high resolution were regarded as the

"ground truth" and used as a benchmark for comparison with other sampling schemes. The results indicate higher sensitivity of basin response with respect to the temporal resolution than to the spatial resolution of the rainfall data. However, in this study, attention is paid only to the spatial sampling of rainfall on flow simulation. St-Hilaire *et al.* (2003) used a rainfall interpolation method (kriging) as a means to estimate the spatial distribution and variance of rainfall. The results revealed a more refined spatial distribution of rainfall during important rainfall events, and the variance was reduced with a denser network. Tarboton *et al.* (1987) and Bras *et al.* (1988) investigated the effect of rainfall sampling strategy on the basin response. The index of the effectiveness of the sampling strategies is defined as the variance of the error of estimated streamflows. This was related to the physical properties of the basin through parameterization. Two stochastic rainfall models were used to generate rainfall, and a state space approach was used to provide a minimum variance linear estimate of flow from a rainfall event, using rainfall and runoff measurements combined. The results obtained related the variance of the estimation error to the measurement strategy and basin (and rainfall) parameters, which is useful in the design of measurement networks.

For the methods used by Krajewski *et al.* (1991) and Azimi-Zonooz *et al.* (1989), discharge data corresponding to the synthesized rainfall data are clearly not available and, hence, it is impossible to investigate the rainfall sampling effect on flow forecasting results from the model. Therefore, this method is not applicable to the present research. If there are very few raingauges, the kriging methods mentioned above (St-Hilaire *et al.*, 2003) are useful to position the sites of new additional raingauges. As stated above, the purpose of this study is the opposite, that is to reduce the density of the existing raingauge network to an appropriate degree, which makes the kriging approach inapplicable. The method developed by Tarboton *et al.* (1987) and Bras *et al.* (1988) is promising for the network design as they defined the sampling strategy as the triplet of (a) number of raingauges, (b) rainfall measurement interval and (c) discharge measurement interval, which is very practical in real network design. They used stochastic rainfall generators to create synthetic rainfall series, and a linear model to estimate the runoff from synthesized rainfall. Also, a hypothetical river basin was used to check the effectiveness of the sampling strategies. However, their method is not used here for two reasons. First, observed rainfall and runoff data will be used, because the method will be applied to a real river basin. Second, in the works of Tarboton *et al.* (1987) and Bras *et al.* (1988), the rainfall was sampled randomly without taking into account the geographical influence on the sampling results, which is one of the purposes of this research.

The second approach, which uses high resolution rainfall data to determine the appropriate spatial sampling scheme of rainfall, is realized under the pre-condition that a dense raingauge network or weather radar, which can provide high resolution precipitation data, already exists. These dense data are used as the representative of the "ground truth". This "ground truth" precipitation field is re-sampled and these precipitation estimates are compared to the "ground truth" situation to investigate the sampling effect on precipitation estimates or hydrological variables derived from precipitation estimates.

Tsintikidis *et al.* (2002) applied statistical methods to quantify the uncertainty associated with the estimation of precipitation for an existing raingauge network and, furthermore, tried to identify the possible sites of additional gauges to reduce the

precipitation interpolation errors. Kriging is also used to interpolate the point rainfall measurements to grid-averaged rainfall series over the catchment. The locations of the additional raingauges are selected such that the greatest reduction in estimation error is obtained. In contrast to the work carried out by Krajewski *et al.* (1991), the analysis of Tsintikidis *et al.* (2000) was based on real observations with an hourly time interval, and the proposed gauge network is appropriate for short-time flood forecasting applications. Duncan *et al.* (1993) used radar-measured rainfall data with a half-hour temporal resolution to study the effect of gauge sampling density on the accuracy of streamflow predictions. Ten sampling densities were used. For each density, hydrographs were computed for a large number of randomly sampled spots (spatially). The results show that, for increased gauge density, the standard deviation of the predicted hydrograph falls off as a power law. Bradley *et al.* (2002) followed a similar method of using radar-estimated precipitation to design raingauge networks. Their approach differs from that of Duncan *et* al. (1993) in that, instead of using the hypothetical sampling point rainfall directly, they used a stochastic model to simulate gauge observations based on the areal-average precipitation for each radar grid cell. The stochastic model accounts for sub-grid variability of precipitation within the cell and gauge measurement errors. The results indicate that errors of network estimation for hourly precipitation are extremely sensitive to the uncertainty in sub-grid spatial variability. Georgakakos *et al.* (1995) studied the effect of the number of raingauges (from 1 to 11) on the simulation performance (cross-correlation coefficient between observed and simulated flow) in two American river basins with an area of about 2000 km^2. The results revealed that the cross-correlation coefficient increased considerably until five raingauges were reached. Therefore they concluded that 11 raingauges are more than adequate to represent mean areal precipitation over the catchments for their research purpose (the linkage of catchment climatology and hydrology to time scale).

The approach used by Tsintikidis *et al.* (2002) is not applicable to the present study because a dense raingauge network is already available in the study area; therefore, adding new gauges and identifying their locations is not expected to be necessary. The methodologies used by Duncan *et al.* (1993) and Bradley *et al.* (2002) are not applicable either, due to the lack of radar-measured rainfall data in this case. The present research will be similar to the method used by Georgakakos *et al.* (1995) in the sense that the relationship between the number of raingauges and the performance of flow simulation is to be explored. It is different from their work in the sense that: (a) the effect of the spatial sampling on the cross-correlation coefficient between mean areal precipitation (not the simulated flow) and the observed flow is studied and (b) the geographical location of the raingauges is taken into account.

The results are presented of raingauge network rationalization (discarding the redundant gauges and positioning the remaining ones) for application in flow simulation in the mountainous Qingjiang River basin in China, using data collected from an existing dense raingauge network. The essential elements of this approach are the computation of the sampling effect on the variance of precipitation estimation and on the correlation between estimated rainfall and discharge at the outlet of the area of interest. The validation of the results from the statistical analyses is done by applying a lumped HBV model. The sampling effect on the variance of precipitation is originally formulated by Yevjevich (1972), and has been used by Krajewski *et al.* (1991) and

Tsintikidis *et al.* (2002). The present research extends the idea to investigate the sampling effect of precipitation on streamflow simulation. The next section describes the region of interest and the data used in the study; "Methodology" gives the outline of the methods which are used; "Statistical analysis" describes the theory behind the variance reduction effect of the estimated rainfall and the increase in the correlation between rainfall and discharge time series, caused by an increasing number of raingauges; and "Hydrological modelling" explains how the HBV model is applied to validate the results from the statistical analysis. The results of these approaches are presented followed by discussion of the results and, finally, conclusions drawn.

STUDY AREA AND DATA DESCRIPTION

The study area is the area upstream of Yuxiakou in the Qingjiang River basin in China as shown in Fig. 1. The whole basin is located in the south of the Three Gorges area of the Changjiang (Yangtze) River. The Qingjiang River joins the Changjiang about 100 km downstream of the Three Gorges Dam (which is still under construction). The length of the main river channel is 423 km, with an overall head difference of 1439 m. The basin area is 17 000 km^2 (the study area upstream of Yuxiakou, indicated in grey in Fig. 1, is 12 209 km^2), of which 34% is forested and 13% is agricultural land. It is a mountainous river basin, with an average altitude of about 1500 m. Most of the river channel is banked with steep valleys (depths ranging from 200 to 1000 m), with narrow river widths and steep slopes, leading to very quick hydrological responses to rainfall events. The basin is located in the subtropical zone. The local climate is heavily influenced by monsoon winds blowing from the south, bringing heavy rainfall in the summer. The annual mean precipitation reaches 1400 mm and the annual mean discharge at the outlet of the basin is 464 m^3 s^{-1}. The annual mean temperature is 16°C, the annual mean relative humidity 70–80% and the annual mean evaporation 820 mm. There is a remarkable difference in runoff among different seasons: 76% of the total runoff volume occurs during the flood season (April–September) and 63% of the

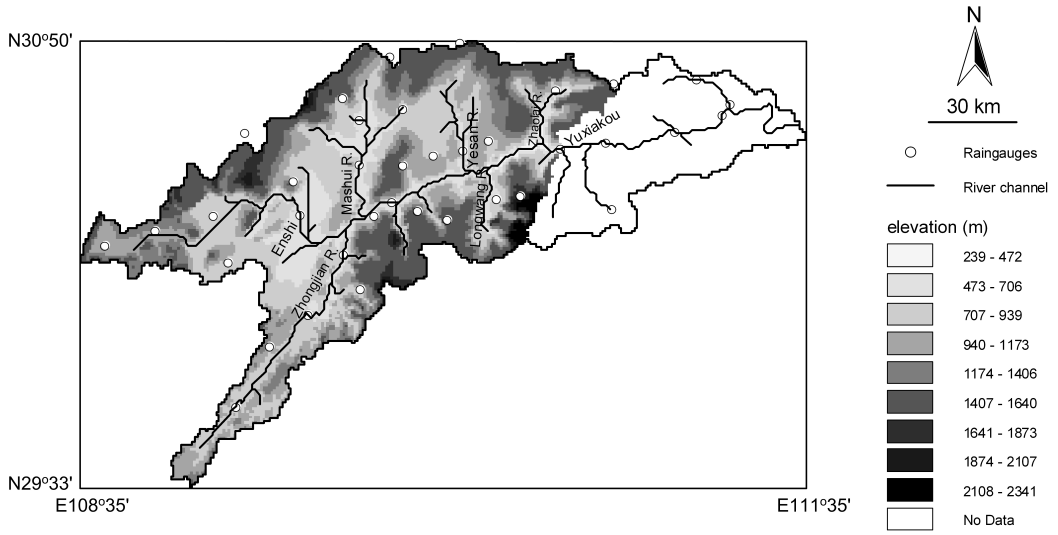

Fig. 1 Qingjiang River basin in the midstream area of Changjiang River, China.

flooding events occur in June and July (QHDC, 1998). To summarize, the Qingjiang River is a well forested and quickly responding mountainous river.

Two types of data are used in this study: (a) hydrological data, including precipitation, discharge and evaporation; and (b) land-use data. In total, 10 years of hydrological data are used, from 1989 to 1995, and from 1997 to 1999. The data from 1996 are missing. The precipitation data were obtained from a network consisting of 26 raingauges as shown in Fig. 1, and were measured with an interval of 6 hours. Of these 26 raingauges, only half operate all year round, while the other half only operate in the flood season (April–September). To create continuous records for all raingauges, the 13 all-year-round raingauges are used as reference gauges. The raingauges with missing data in winter will use the data measured at the nearest reference gauge. This replacement of data is assumed to be reasonable on the condition that, first of all, a spatially lumped HBV model will be used to simulate the streamflow. Therefore, this nearest neighbour approach for the winter season will not have substantial effects on the results. Secondly, the average runoff in winter contributes only 24% to the annual runoff at the outlet of the basin. Therefore it is assumed that adopting the neighbouring measurements will have a minor influence on the overall simulation results. Evaporation data measured using evaporation pans (type E-601) located at Yuxiakou were used as one of the inputs. The measuring interval of evaporation data is one day and data are downscaled to 6 h in order to be consistent with those of the rainfall data. Although there are several flow gauges in the study area, only the data from Yuxiakou are used here. Its measuring interval is also 6 h. The area is categorized into only two types of land use: forest (4089 km^2) and field (8120 km^2), because the study area will be treated as one sub-basin in the HBV model.

METHODOLOGY

The identification of the appropriate spatial sampling method of rainfall for flow simulation is carried out using both statistical analyses and a specific hydrological model. The most common way of performing flow simulation is by running a model (most models are physics-based). The design of such a model consists of different steps: (a) identification of possible physical processes taking place in the rainfall–runoff transformation, and mathematical description of these processes; (b) use of input–output data pairs to adjust the parameters in the model to reduce the discrepancy between computed and recorded output time series (calibration); (c) use of new input–output data series in the calibrated model to see if the model performs well in a new situation (validation); and (d) operational use of the calibrated and validated model, in which the previous forecasting results are sometimes updated (either continuously or irregularly) with new data.

For appropriate flow simulation, the appropriate resolutions of input data have to be known. The spatial and temporal resolutions of rainfall data are closely related, mutually affected and both have a large influence on flow simulation results. Here, the effect of temporal resolution of rainfall is not considered and the full focus is on the spatial sampling effect. The appropriate spatial sampling of rainfall could be determined by the application of a physics-based model, that is, by simply enumerating all the possible combinations of raingauges, aggregating their rainfall time series into

areal averages time series, which are subsequently fed into the model to see which combination gives good enough simulation of the rainfall–runoff relationship. This method is conceptually straightforward, but practically very difficult to implement, because calibrating and running a physics-based hydrological model is very time-consuming, and the number of possible combinations of raingauges may be enormous.

Fortunately, the range of raingauge combinations which are most likely to lead to appropriate rainfall–runoff modelling can be narrowed down, by looking at the statistical characteristics of rainfall and discharge time series. Then, a physics-based model (HBV) can be used to test if the statistically superior combinations of raingauges can give better flow simulation results than other combinations. The statistical methods proposed in this section are initiated by two research questions: (a) What is the effect of an increasing number of raingauges on the statistics of areally-averaged rainfall time series, i.e. what is the effect on the variance of the areally-averaged rainfall? Since what is really of interest is the discharge time series in the river channel, this leads to the next question: (b) How does the change in the statistics of areally averaged rainfall influence the relationship between rainfall and discharge time series if the number of raingauges increases? i.e. what is the effect on the cross-correlation value between areally averaged rainfall and discharge?

STATISTICAL ANALYSIS

Variance reduction due to the increase in the number of raingauges

The most distinct effect of the increase in the number of raingauges on the areally averaged rainfall series is the reduction of its variance. Variance in the rainfall time series provides one estimate of the variability of the rainfall at a location or of a region. It is probably neither the only nor even the most useful indicator of variability because of the skewness of the distribution of the precipitation records. If more raingauges are averaged together, the skewness of the areally averaged rainfall time series decreases. Therefore, the effect of skewness on the derived information about variability (variance) decreases with an increasing number of raingauges. First, the relationship between the variance of station measurements and that of areally averaged rainfall is established, using the method adapted from Yevjevich (1972). Then, this methodology is extended to study the effect of an increasing number of raingauges on the rainfall–runoff cause–effect relationship, because all the methodologies used in this study are oriented towards flow simulation as the final objective. The variance considered here is derived from full time series, in which dry days are not removed.

Assume rainfall is gauged at n points in an area, and the length of the records is N (with whatever measuring interval). Under the assumption that the rainfall process recorded in the area is ergodic and homogeneous in space, the variance of the areally averaged rainfall can be formulated as (after Yevjevich, 1972):

$$s^2 = \frac{\overline{s_j^2}}{n}[1 + \bar{r}(n-1)] \tag{1}$$

where:

$$\overline{s_j^2} = 1 \Big/ \left(n \sum_{j=1}^{n} s_j^2 \right) \tag{2}$$

$$s_j^2 = \frac{\sum_{i=1}^{N} (x_{ij} - \overline{x_j})^2}{N} \tag{3}$$

$$\overline{x_j} = \frac{\sum_{i=1}^{N} x_{ij}}{N} \tag{4}$$

$$\bar{r} = \frac{\sum_{j=1}^{n-1} \sum_{i=j+1}^{n} r_{ij}}{C_n^2} = \frac{2 \sum_{j=1}^{n-1} \sum_{i=j+1}^{n} r_{ij}}{n(n-1)} \tag{5}$$

where $\overline{s_j^2}$ is the mean of the station variance; s_j^2 is the variance of the *j*th raingauge; x_{ij} is the rainfall data recorded at the *i*th time point and the *j*th raingauge; $\overline{x_j}$ is the mean of the *j*th raingauge; r_{ij} is the sample product–moment correlation coefficient between rainfall series of gauges *i* and *j*; and $\bar{r}$ is the arithmetic mean of the correlation coefficients of all bi-combinations of the raingauges.

According to equation (1), it is expected that the variance of the areally averaged rainfall will decrease hyperbolically with an increasing number of raingauges *n* as shown in Fig. 2. For *n* approaching infinity, equation (1) shows that the variance of areally averaged rainfall is a linear function of the average point variance and the average correlation coefficient in the area. Rodriguez-Iturbe & Mejia (1974) showed that, for a stationary isotropic spatial random field, the average correlation coefficient can be calculated using a distribution function for the distance between any two points randomly chosen in the area. This can be used to calculate, for any area, the variance

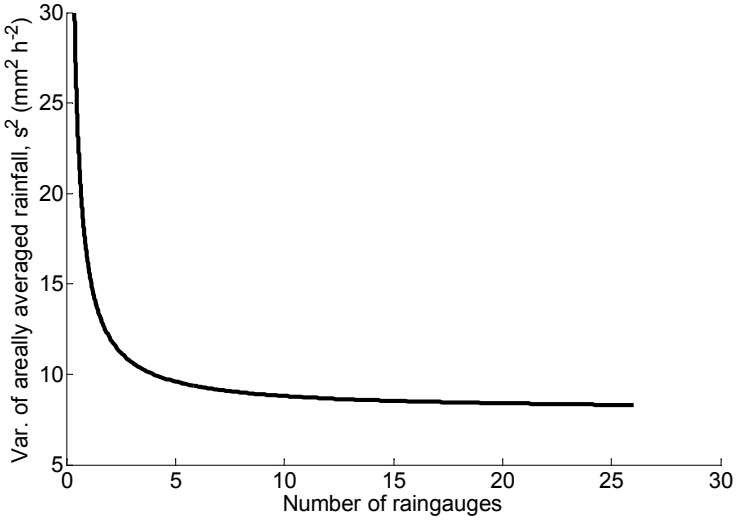

Fig. 2 Effect of variance reduction.

of areally averaged rainfall based on the average point variance and the correlation length for rainfall (see e.g. Booij, 2002).

From variance reduction to cross-correlation

For a better understanding of how raingauge density affects the flow simulation accuracy, it is necessary to establish how the different number of raingauges influences the relationship between the areally averaged rainfall series and the discharge series. Therefore, this section will contribute to establishing the relationship between the number of raingauges and the rainfall–runoff correlation coefficient. Then, this will be further tested by a hydrological model (HBV).

The effect of the number of raingauges on rainfall–runoff modelling will be investigated without running a hydrological model, but will solely be based on input–output data series. With more raingauges, the variance of the resulting areally averaged rainfall series will decrease. As the output (discharge) series remains the same, the interrelationship between the areally averaged rainfall series and the discharge series will also be influenced by increasing the number of raingauges. The cross-correlation coefficient is used here as an indicator of the relationship between the areally averaged rainfall series and the discharge series, and the effect of the number of raingauges on the cross-correlation coefficient will be investigated. The time lag between the areally averaged rainfall and discharge is considered when calculating the cross-correlation coefficients. Results will be shown for the time lag that corresponds to the maximum cross-correlation value. The selected time lag reflects the hydrological response time between the commencement of the rainfall event and the corresponding discharge at the outlet of the area.

The expected cross-correlation between the areally averaged rainfall series x_i and discharge series y_i at the outlet of the area with time lag k can be formulated as:

$$R_k = \frac{\operatorname{cov}(x_i, y_{i+k})}{(\operatorname{var} x_i \operatorname{var} y_{i+k})^{1/2}} = \frac{\sum_{i=1}^{N-k}(x_i - \frac{1}{N-k}\sum_{i=1}^{N-k} x_i)(y_{i+k} - \frac{1}{N-k}\sum_{i=1}^{N-k} y_{i+k})}{[(N-k)s_x^2]^{1/2}[(N-k)s_y^2]^{1/2}} = \frac{\sum_{i=1}^{N-k}(y_{i+k} - \bar{y})x_i - \bar{x}\sum_{i=1}^{N-k}(y_{i+k} - \bar{y})}{[(N-k)s_x^2]^{1/2}[(N-k)s_y^2]^{1/2}} \quad (6)$$

where R_k is the expected cross-correlation coefficient for lag time k, $\operatorname{cov}(x_i, y_{i+k})$ is the covariance between x_i and y_i for lag time k, var x_i and var y_{i+k} are the variance for series x_i and y_{i+k}, respectively. Because the term $\sum_{i=1}^{N-k}(y_{i+k} - \bar{y})$ in the numerator of equation (6) equals $\sum_{i=1}^{N-k}(y_{i+k}) - (N-k)\bar{y} = 0$, it can be omitted from equation (6), leading to:

$$R_k = \frac{\sum_{i=1}^{N-k}(y_{i+k} - \bar{y})x_i}{(N-k)s_x s_y} \quad (7)$$

Because the discharge series at the outlet of the area remains the same, no matter how many raingauges are involved, the standard deviation of the discharge series s_y is

constant, and the same with the term $(y_{i+k} - \bar{y})$ in the summation of the numerator. The areally averaged precipitation, x_i, does not stay constant, but will change randomly for different combinations of raingauges. However, with the increase in the number of raingauges, the areally averaged rainfall series produced from these raingauges will converge gradually to the real situation of the rainfall event. For an individual rainfall event at a certain moment in time, the variance of its areally averaged value will decrease with increasing number of raingauges and converge to the ground truth when the number of raingauges approaches infinity. When the number of raingauges approaches infinity, the summation term $\sum_{i=1}^{N-k}(y_{i+k} - \bar{y})x_i$ in the numerator of equation (7) converges to a certain constant. To clarify the relationship between R_k and n represented by equation (7), the summation term is calculated by using the mean value of x_i, which is actually the areally averaged rainfall series aggregated from the observations of the total 26 raingauges in the area. In this case, $\left[\sum_{i=1}^{N-k}(y_{i+k} - \bar{y})x_i\right]/(N-k)s_y$ is regarded as a constant and denoted as A. According to the observed rainfall and runoff data, $A = 2$. Therefore, the value of R_k depends solely on s_x which decreases hyperbolically with an increasing number of raingauges, as revealed by equation (1).

Substituting $\left[\sum_{i=1}^{N-k}(y_{i+k} - \bar{y})x_i\right]/(N-k)s_y = A$ together with equation (1) (where $s = s_x$) into equation (7), the relationship between R_k and the number of raingauges n can be expressed as:

$$R_k = \frac{A}{s_x} = A\left[\frac{n}{\overline{s_j^2}(1+(n-1)\bar{r})}\right]^{1/2} \tag{8}$$

This implies that the cross-correlation of areally averaged precipitation and discharge will increase hyperbolically with an increasing number of raingauges as shown in Fig. 3, exhibiting a reverse behaviour compared to the $s_x - n$ relationship shown in Fig. 2. In addition to the constant A, the values of the other two constants $\overline{s_j^2}$ and $\bar{r}$ are also calculated from the rainfall observations of 26 raingauges in the study area as 16 mm^2 h^{-2} and 0.5, respectively. Substituting the values of $\overline{s_j^2}$ and $\bar{r}$ into equation (8), the value of R_k will converge to 0.71 ($R_k|_{n\to\infty}$) for this study area when the number of raingauges approaches infinity. Therefore, $R_k|_{n\to\infty}$ is the maximum cross-correlation coefficient between areally averaged rainfall and discharge that can be achieved, if one simply takes the arithmetic mean of the station rainfall as the areally averaged rainfall.

According to Fig. 3, the correlation between an areally averaged rainfall series and a discharge series increases quickly at the beginning and levels off after a certain threshold. This implies that the similarity between the areally averaged rainfall and discharge series will also increase hyperbolically if one regards R_k as the indicator of the similarity, leading to the expectation that the mathematical mapping between the input and the output can be established more easily. This inference will be tested by a physics-based hydrological model (HBV).

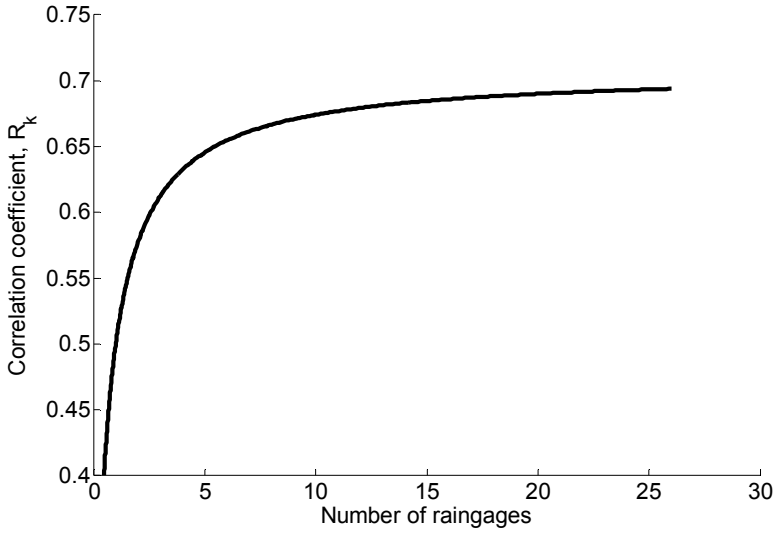

Fig. 3 Rainfall–runoff correlation coefficients *vs* number of raingauges.

Criterion for appropriateness

As shown in Fig. 3, the increase in the correlation coefficient is no longer significant beyond a certain critical number of raingauges. The improvement in the performance of hydrological modelling is expected to act in a similar way. Therefore, it is concluded that the spatial sampling density of rainfall is already "good enough" for flow simulation, if the number of raingauges is larger than this critical number. This critical number of raingauges is identified as the appropriate number of raingauges, if the first derivative of R_k with respect to n is smaller than or equal to a threshold value, chosen arbitrarily to be 0.01. Therefore, the criterion to find the appropriate number of raingauges is defined as:

$$\frac{dR_k}{\mathrm{d}n} = \frac{A\left(1-\bar{r}\right)}{2\left(\overline{s_j^2}\right)^{1/2}} \times \frac{1}{n^{1/2}\left[1+(n-1)\bar{r}\right]^{3/2}} \leq 0.01 \tag{9}$$

This leads to $n \geq 5$ and five is identified as the appropriate number of raingauges for this study area.

HYDROLOGICAL MODELLING

It is generally recognized that a simultaneous use of both statistic and physics-based (or deterministic) methods of analysis of hydrological processes is necessary to produce the best scientific and practical information for hydrology (Yevjevich, 1972). Using a statistical method independently of the physics-based one may lead to an analysis of data without a sound theoretical background. Therefore, the conceptual hydrological model HBV is used here to verify the results from the statistical analysis.

The HBV model is a conceptual, semi-distributed hydrological model developed by SMHI (Swedish Meteorological and Hydrological Institute) which is used for continuous computation of discharges at the outlet of a river basin. The model has

proven to be a rather robust tool for the assessment of the basin-scale runoff dynamics in various parts of the world (e.g. Bergström, 1995; Zhang & Lindström, 1996; Lindström *et al.*, 1997). Time series data including precipitation, air temperature and estimated potential evapotranspiration are used as inputs to calculate the river discharge (output). Observed discharge series can be used to calibrate the model. In the implementation of the model, the whole considered river basin can be divided into a number of sub-basins. Information about geographical features of the sub-basins is also needed to assemble the model, namely, the area, mean elevation and type of vegetation zones (forest, field, etc.). Each sub-basin can be calibrated separately provided that discharge data at the outlet of the sub-basin are available. The outflow of each sub-basin will be routed to the outlet of the whole basin using the Muskingum method (Linsley *et al.*, 1988) and combined with outflows from other sub-basins, taking into account delaying and damping effects. Each sub-basin model consists of six subroutines: a snow and rainfall routine, a soil routine, a fast flow routine, a slow flow routine, a transformation routine and a routing routine. These sequential sub-routines simulate the complete hydrological process from the commencement of precipitation to the formation of discharge at the outlet. Detailed descriptions can be found in SMHI (2003) and Bergström (1995).

For the application of the HBV model to check the results of the statistical analyses, the study area needs to be subdivided into a number of sub-basins. Here, the whole area upstream of Yuxiakou was treated as one sub-basin. Seven years (1989–1995) of hydrological data (precipitation, evaporation and discharge) were used to calibrate the HBV model. Precipitation and evaporation were used as input, and discharge as output. All 26 raingauges available in the area were used to obtain areally averaged rainfall for calibration. The calibrated model was used for validation. During the validation, areally averaged rainfall series were obtained from different numbers (from 1 to 26) of raingauges, and for a specific number of raingauges for different combinations of raingauges, to compare the effect of different spatial sampling strategies on the performance of the calibrated HBV model. Three years (1997–1999) of hydrological data were used for the validation (data of 1996 are missing). The evaporation and discharge data remained the same during the whole validation procedure.

Two statistical criteria are used to judge the performance of the HBV model: the coefficient of efficiency (R^2) (Nash & Sutcliffe, 1970):

$$R^2 = 1 - \frac{\sum_{i=1}^{N} (Q_{c,i} - Q_{o,i})^2}{\sum_{i=1}^{N} (Q_{o,i} - \overline{Q_o})^2} \tag{10}$$

and the relative accumulated difference between computed and observed discharge:

$$RD = \frac{\sum_{i=1}^{N} (Q_{c,i} - Q_{o,i})}{\sum_{i=1}^{N} Q_{o,i}} \tag{11}$$

where $Q_{o,i}$ is observed discharge, $Q_{c,i}$ is computed discharge, $\overline{Q_o}$ is the mean of the

observed discharge and N is the total number of observations. The value of R^2 ranges from $-\infty$ to 1 and the higher the value, the better the agreement between computed and observed discharges. The relative accumulated difference, RD is used to identify any bias in the water balance, which is particularly useful in the initial stage of the calibration.

RESULTS

The effects of the number of raingauges on the variance of areally averaged rainfall series and on the cross-correlation between areally averaged rainfall and the discharge are shown in Figs 4 and 5, respectively. To illustrate how the variance decreases as

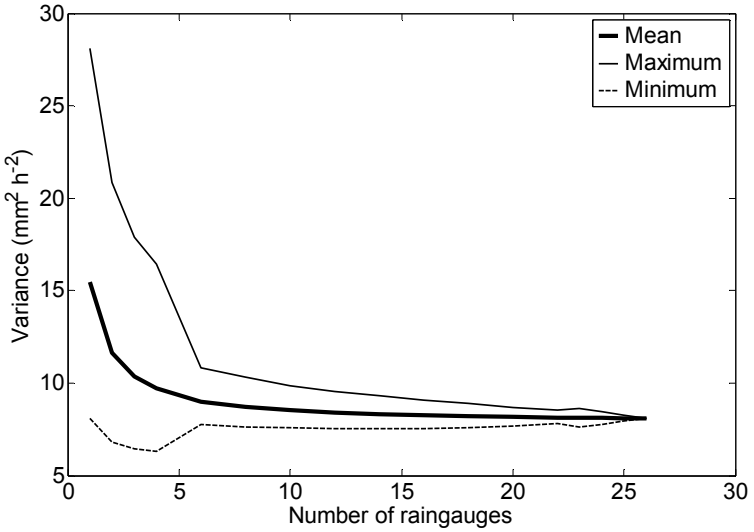

Fig. 4 Effect of the number of raingauges on the variance of areally averaged rainfall.

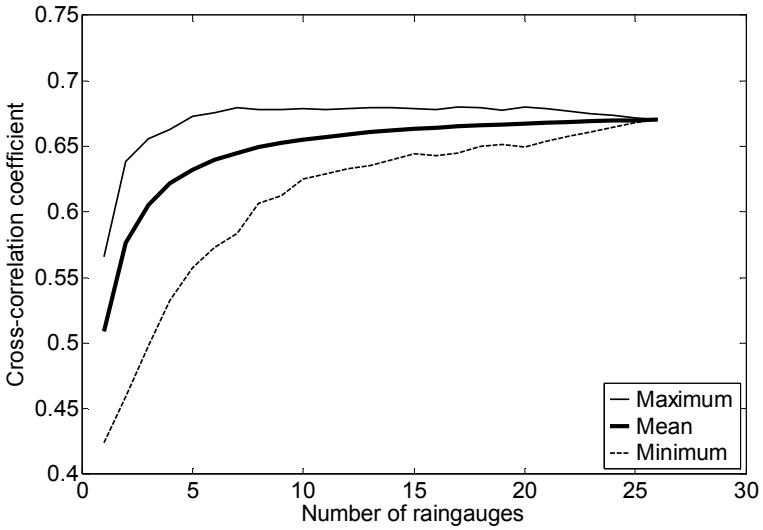

Fig. 5 Effect of the number of raingauges on the cross-correlation between areally averaged rainfall and discharge.

more time series are aggregated, s_n^2 (the variance of n-station averaged time series), is computed for the station data that fall into the area under consideration (Fig. 1). To obtain the true value of s_n^2, the variances of every combination of n stations chosen from the N available stations should be computed and their mean, maximum and minimum values taken. For some combinations, such a calculation is not computationally feasible (e.g. there are $26!/(13!13!) \approx 10^7$ such combinations when choosing $n = 13$ from 26 stations available). Therefore, when choosing n ranging from 1 to 7 and 20 to 26, all possible combinations are enumerated, and the combinations which give the maximum and minimum value of variance are selected to draw the boundary lines as shown in Fig. 4. To present the mean variance, the combination with a variance nearest to the mean variance is taken. For n ranging from 8 to 19, up to 5000 combinations are selected randomly in each case. In order to show that 5000 randomly selected combinations is already enough to produce unbiased means and most of the range (minimum to maximum) of variance and cross-correlation coefficients, the statistics calculated from 10 000 and 15 000 combinations are shown together with the 5000 combinations in Table 1. Two numbers of raingauges are chosen to do this analysis. The results revealed that the means and ranges remain essentially the same.

Table 1 The effect of the number of combinations on the statistics of areally averaged rainfall and lagged cross-correlation between areally averaged rainfall and discharge.

Number of raingauges	Number of combinations	Variance of areally averaged rainfall ($mm^2\ h^{-2}$)			Lagged cross-correlation between areally averaged rainfall and discharge		
		Min.	Mean	Max.	Min.	Mean	Max.
12	5000	6.8	8.5	10.5	0.63	0.66	0.68
	10000	6.7	8.5	10.7	0.63	0.66	0.68
	15000	6.6	8.5	10.7	0.63	0.66	0.68
16	5000	7.1	8.3	9.7	0.64	0.66	0.68
	10000	7.0	8.3	9.9	0.64	0.66	0.68
	15000	7.1	8.3	10.0	0.64	0.66	0.68

As can be seen from Fig. 4, the variance decreases hyperbolically with an increase in the number of raingauges n. This confirms the expected variance reduction phenomenon of the areally averaged rainfall series as indicated in equation (1) and Fig. 2. After a certain threshold, the variance levels off to a final value s_∞^2, which implies that the effect of the variance reduction of areally averaged rainfall series is no longer significant when n is greater than a certain threshold number. As expected, the relationship between the cross-correlation of areally averaged rainfall and discharge and the number of rain stations behaves similarly to the variance reduction effect but in a reverse way (as shown in Fig. 5): the value of the cross-correlation increases hyperbolically, and levels off after the same threshold, for example, five, as suggested in the previous section. The precise selection of this threshold number of raingauges has to be done together with hydrological modelling results, which are presented below.

The HBV modelling results are shown in Figs 6 and 7. The specific combination of rainfall stations was used that gave the maximum, mean and minimum correlation values as shown in Fig. 5, and the corresponding areally averaged rainfall series was

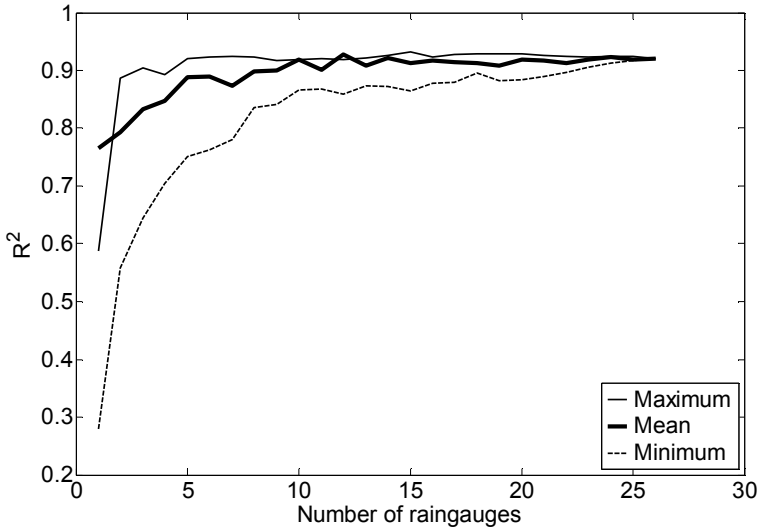

Fig. 6 Nash-Sutcliffe coefficient *vs* the number of raingauges.

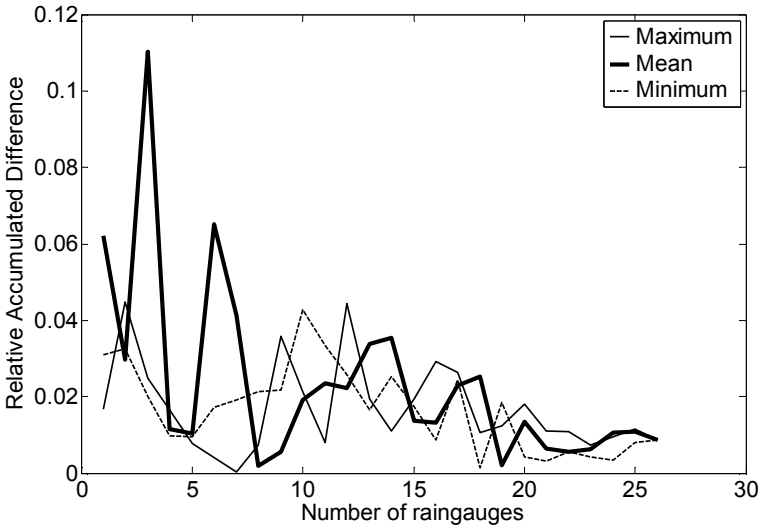

Fig. 7 Absolute value of relative accumulated difference *vs* the number of raingauges.

created. These areally averaged rainfall series were used in the simulation with the calibrated HBV model under the expectation that the higher the correlation value, the easier for the model to simulate the rainfall–runoff relationship. Figure 6 confirms this expectation because: (a) the combinations which give the maximum correlation values lead to better model performance as indicated by a higher R^2 value, compared to the lines which represent the mean and minimum correlation values; and (b) the increase of the R^2 value behaves very similarly to the increase in correlation values shown in Fig. 5; both increase hyperbolically, but level off after a certain threshold. This threshold number of raingauges can be spotted from Fig. 6, to be five. Beyond this number, a further increase in the number of raingauges will not largely improve the model performance. This confirms the finding in the previous section. Figure 7 shows the effect of the number of raingauges on the absolute value of the relative

accumulated difference of the discharge. Although the three lines do not decline in parallel as in Fig. 6, their decrease does exhibit a hyperbolic trend as does the variance reduction behaviour. The fact that the decreasing trend is much more diverse than the lines shown in Fig. 6 can be explained as follows:

(a) The relative accumulated difference (*RD*) emphasizes the capability of the model to fulfil the water balance. When the number of raingauges is small and they are selected randomly in the basin, the chance of missing certain rainfall event(s) is quite high, which then leads to big *RD* values. This is especially the case for small-scale rainfall events.
(b) The specific combination of raingauges used to carry out hydrological simulation is selected by a linear method, whereas the hydrological modelling is a nonlinear process. This discrepancy implies that the minimum and maximum calculations alluded to in Figs 6 and 7 do not fully reflect the minimum and maximum of model performance, and may be part of the reason of the poor behaviour found in Fig. 7.

Figure 8 gives an example of the geographical locations of the combinations of raingauges (from 3 to 7) which give the minimum, mean and maximum cross-correlation between areally averaged rainfall and discharge. Two characteristics of geographical distribution can be detected from the combinations that yield maximum and minimum correlation values: (a) a strong effect of the geographical location: raingauges which yield a maximum correlation value are all located at the centre of the area, whereas the ones that yield the minimum correlation are located in the most remote places; and (b) the addition of new raingauges to the existing raingauges (which give maximum and minimum correlation coefficients) show a successive property. The old ones remain; the new raingauges added are generally based on the existing network for combinations which give the maximum correlation values. So for the combinations that yield minimum and maximum correlation values, the expansion of the network is based on the old raingauges. However, for the combination of raingauges which give the mean correlation value, this is not the case. They are more evenly distributed than the other two cases.

DISCUSSION

After superimposing Fig. 8(c) on the DEM map of Fig. 1, it can be seen that most of the raingauges which yield maximum correlation values between areally averaged rainfall and discharge (and also exhibit the best HBV modelling performance as seen from Fig. 6) are located in the mountains around the Enshi basin, at the banks of the tributaries: the Mashui and Zhongjian rivers. Most rainfall occurs in summer (76%), and during summer seasons, this area is mainly influenced by two climate systems: the subtropical anticyclone in the western Pacific Ocean and the monsoon. If the former climate system dominates the area, rainfall events will move from east to west, bringing heavy orographic rain to the east of Yuxiakou. This meteorological factor is not considered here because the area affected is out of interest of this study. If the monsoon is prevailing on the local climate system, wind blows from the southwest to the northeast and brings humid air from the South China Sea or the Bay of Bengal. Because the Enshi basin happens to have an open end towards the southwest direction,

Fig. 8 Geographical locations of three combinations of 3–7 raingauges: (a) minimum correlation value, (b) mean correlation value, and (c) maximum correlation value.

heavy orographic rainfall frequently occurs around the uphill slope of the basin. This leads to the rich contributions (in the sense of annual mean discharge) from the Marshui and Zhongjian rivers, as seen in Table 2. On average, 34% of the discharge at the outlet of the study area (Yuxiakou) comes from these two rivers. Therefore, as expected, in order to provide good flow simulation results, the majority of the raingauges (if the number is limited) should be concentrated in this rain-rich area (see Fig. 8(c)).

Table 2 Annual mean discharge (AMD) of the five tributaries with area >500 km^2 and area upstream of Enshi, compared to the AMD at the outlet of the study area, Yuxiakou (HSCSC, 1991).

	Upstream Yuxiakou	Upstream Enshi	Zhongjian River	Mashui River	Yesan River	Longwang River	Zhaolai River
Area (km^2)	12 209	1 900	1 881	1 693	1 092	624	787
Fraction of total area	0.65	0.16	0.15	0.14	0.09	0.05	0.06
AMD (m^3 s^{-1})	312	70.3	48.8	55.2	28.4	16	16.7
Fraction of total AMD	0.75	0.23	0.16	0.18	0.09	0.05	0.05

Figure 8(a) shows an even more remarkable geographical preference: all the combinations of raingauges which give a minimum correlation between areally averaged rainfall and discharge (and also the worst performance of HBV modelling, in the sense of R^2 value, see Fig. 6) are located at the west of the basin. Because the western area (upstream of Enshi) contributes quite a proportion of the total runoff (23%), this cannot be ascribed to a shortage of rain. Instead, this reveals that the method used so far is strongly dependent on the spatial resolution of the model, i.e. on the fact that the whole area is treated as one sub-basin. Because of this, the cross-correlation coefficients are calculated with a time lag of three time units (18 h), which represents the overall characteristics of the area. If the area is sub-divided into more sub-basins and different travelling times are considered for each sub-basin, the raingauges will be distributed more evenly, and the total number of raingauges that gives the maximum correlation value is expected to increase.

The found appropriate number of raingauges (5) is valid solely for the river basin under study. Also, such a number of raingauges depends on the criteria used. The choice of the threshold value of the first derivative of the cross-correlation in equation (9) is subjective. A different threshold value will lead to a different number of raingauges. However, the methodology can be adopted for any area.

The method used in this study is valid in river basins where a large number of raingauges already exist, and the problem of rational network reduction comes about. If there are very few raingauges, and the river basin manager wants to find out the appropriate number of raingauges, this method will be difficult to implement, because it will be difficult to find out the final value of s_∞^2 (Osborn & Hulme, 1997). The methodology used here shows that by looking only at the rainfall–runoff data, one can re-evaluate the efficiency of the network and determine a subset of the most important raingauges, so that the rainfall information provided by these gauges is sufficient for obtaining good flow simulation results.

A lumped HBV model was used, which treats the whole study area as one sub-basin, to check the results from the statistical analyses. Therefore, the resulting appropriate number of five raingauges, and their locations will not necessarily stay the same if this HBV model is further developed into a distributed model, i.e. more sub-basins are involved. More rainfall gauges are expected to be necessary in this case. However, the same method can be implemented in each sub-basin to determine the appropriate number of raingauges. The subdivision of the whole basin should be in accordance with the number of discharge gauges in the basin; the outlet of each sub-

basin should have at least one flow gauge which observes discharge records. If the parameters in one sub-basin cannot be calibrated according to the observed discharge data, regionalization methods can be used (Seibert, 1999).

The number of raingauges found by this study provides a lower limit of the number of raingauges to be chosen. However, a certain number of additional raingauges should be considered to cope with the possible malfunctioning of the network, and to provide a certain degree of redundant rainfall information for flow simulation. The choice of additional raingauges depends on how the river management authority handles the malfunctioning of the raingauge network.

CONCLUSIONS

The effect of an increase in the number of raingauges on the variance reduction of mean areal precipitation and on the increase of the cross-correlation coefficient between mean areal precipitation and discharge was investigated. The aim was to identify the appropriate number of raingauges and their geographical locations for improved flow simulation using a spatially lumped HBV model. The results reveal that the correlation coefficient increases hyperbolically with an increase in the number of raingauges but levels off after a critical number of raingauges, which for this study area turns out to be five. The performance of the lumped HBV model (in terms of coefficient of efficiency, R^2) increases in a similar way with the increase in the number of raingauges. Therefore, five was identified to be the sufficient number of raingauges in the study area (for the satisfactory performance of the lumped HBV model). The geographical locations of the raingauge combinations which give the maximum value of correlation coefficient and R^2 are strongly correlated with local climatic and geographical conditions. Most of them are located in an area where heavy orographic rainfall is the dominant form of the local precipitation pattern. The combinations of raingauges which gave the worst performance of HBV modelling (and also gave the minimum correlation value) are located in the west of the study area (the farthest from the outlet of the area). This is understandable, because the farthest raingauges will have the least influence on the runoff at the outlet if one treats the whole area as only one sub-basin in the HBV model. However, if the study area is divided into multiple sub-basins, the raingauges which give the worst performance of the HBV model may be located elsewhere. One can conclude that the methodology introduced herein is dependent on the spatial resolution of the model.

Acknowledgements The research is jointly sponsored by Royal Dutch Academy of Science and Arts (KNAW project no. 02CDP006), University of Twente and China Three Gorges University (CTGU Science Foundation). The hydrological data were kindly provided by Qingjiang Hydropower Development Cooperation (Reservoir Regulation Centre) in China. Thanks are due to Dr Déborah Idier (now working in BRGM, Orléans, France) for translating the abstract into French. The comments of two anonymous reviewers helped to improve the manuscript substantially.

REFERENCES

Azimi-Zonooz, A., Krajewski, W. F., Bowles, D. S. & Seo, D. J. (1989) Spatial rainfall estimation by linear and non-linear co-kriging of radar-rainfall and raingauge data. *Stochast. Hydrol. Hydraul.* **3**, 51–67.

Bergström, S. (1995) The HBV model. In: *Computer Models of Watershed Hydrology* (ed. by V. P. Singh). Water Resources Publications, Littleton, Colorado, USA.

Booij, M. J. (2002) Extreme daily precipitation in Western Europe with climate change at appropriate spatial scales. *Int. J. Climatol.* **22**, 69–85.

Bradley, A. A., Peters-Lidard, C., Nelson, B. R., Smith, J. A. & Young, C. B. (2002) Raingauge network design using Nexrad precipitation estimates. *J. Am. Water Resour. Assoc*. **38**(5), 1393–1407.

Bras, R. L., Tarboton, D. G. & Puente, C. (1988) Hydrologic sampling—a characterization in terms of rainfall and basin properties. *J. Hydrol.* **102**, 113–135.

Duncan, M. R., Austin, B., Fabry, F. & Austin, G. L. (1993) The effect of gauge sampling density on the accuracy of streamflow prediction for rural catchments. *J. Hydrol.* **142**, 445–476.

Georgakakos, K. P., Bae, D. H. & Cayan, D. R. (1995) Hydroclimatology of continental watersheds: 1. temporal analyses. *Water Resour. Res*. **31**(3), 655–675.

HSCSC (Hubei Science Consultant Service Centre) (1991) The Consultant Report for the Development of Qingjiang River Basin, Hubei, China (in Chinese).

Krajewski, W. F., Lakshmi, V., Georgakakos, K. P. & Jain, S. C. (1991) A Monte Carlo study of rainfall sampling effect on a distributed catchment model. *Water Resour. Res*. **27**(1), 119–128.

Lindström, G., Johansson, B., Persson, M., Gardelin, M. & Bergström, S. (1997) Development and test of the distributed HBV-96 hydrological model. *J. Hydrol.* **201**, 272–288.

Linsley, R. K., Kohler, M. A. & Paulhus, J. L. H. (1988) *Hydrology for Engineers*. McGraw-Hill, London, UK.

Nash, J. E. & Sutcliffe, J. V. (1970) River flow forecasting through conceptual models, Part 1: a discussion of principles. *J. Hydrol*. **10**, 282–290.

Osborn, T. J. & Hulme, M. (1997) Development of a relationship between station and grid-box rain day frequencies for climate model evaluation. *J. Climate* **10**, 1885–1908.

QHDC (Qingjiang Hydropower Development Cooperation-Reservoir Regulation Centre), CWRC (Changjiang Water Resources Committee-Department of Planning) (1998) Regulation rules of Geheyan Reservoir, Qingjiang. Hubei, China (in Chinese).

Rodriguez-Iturbe, I. & Mejia, J. M. (1974) On the transformation from point rainfall to areal rainfall. *Water Resour. Res*. **10**, 729–735.

SMHI (2003) Integrated hydrological modelling system (IHMS) *HBV Manual Version 4.5*. Swedish Hydrological and Hydrological Institute, Norrköping, Sweden.

St-Hilaire, A., Ouarda, T. B. M. J., Lachance, M., Bobée, B., Gaudet, J. & Gignac, C. (2003) Assessment of the impact of meteorological network density on the estimation of basin precipitation and runoff: a case study. *Hydrol. Processes* **17**(18), 3561–3580.

Tarboton, D. G., Bras, R. L. & Puente, C. E. (1987) Combined hydrologic sampling criteria for rainfall and streamflow. *J. Hydrol*. **95**, 323–339.

Tsintikidis, D., Georgakakos, K. P., Sperfslage, J. A., Smith, D. E. & Carpenter, T. M. (2002) Precipitation uncertainty and raingauge network design within Folsom Lake watershed. *J. Hydrol. Engng*. **7**(2), 175–184.

Seibert, J. (1999) Regionalisation of parameters for a conceptual rainfall–runoff model. *Agric. For. Met*. **98–99**, 279–293.

Yevjevich, V. (1972) *Probability and Statistics in Hydrology*. Water Resources Publications, Littleton, Colorado, USA.

Zhang, X. & Lindström, G. (1996) A comparative study of a Swedish and a Chinese hydrological model. *Water Resour. Bull*. **32**(5), 985–994.

Received 15 June 2004; accepted 22 December 2004