Satellite and gauge rainfall merging using geographically weighted regression

QINGFANG HU[1], HANBO YANG[2], XIANMENG MENG[3], YINTANG WANG[1] & PENGXIN DENG[1]

1 State Key Labratory of Water Resources and Hydraulic Engineering & Science, Nanjing Hydraulic Research Institute, Nanjing, 210029, China
hqf_work@163.com

2 Department of Hydraulic Engineering, Tsinghua University, Beijing, 100084, China

3 School of Environmental Studies, China University of Geosciences, Wuhan, 430074, China

Abstract A residual-based rainfall merging scheme using geographically weighted regression (GWR) has been proposed. This method is capable of simultaneously blending various satellite rainfall data with gauge measurements and could describe the non-stationary influences of geographical and terrain factors on rainfall spatial distribution. Using this new method, an experimental study on merging daily rainfall from the Climate Prediction Center Morphing dataset (CMOROH) and gauge measurements was conducted for the Ganjiang River basin, in Southeast China. We investigated the capability of the merging scheme for daily rainfall estimation under different gauge density. Results showed that under the condition of sparse gauge density the merging rainfall scheme is remarkably superior to the interpolation using just gauge data.

Key words satellite rainfall; rainfall merging; geographically weighted regression; CMORPH

INTRODUCTION

Satellites usually have near global coverage for remote rainfall monitoring and they are especially valuable for regions that lack adequate surface-based measuring techniques. At the same time, satellite rainfall datasets are usually free of charge and their availability is not limited by administration factors. Due to these advantages, in recent years significant developments have been achieved in the field of satellite rainfall estimation. However, satellite rainfall estimates have been produced at rather coarse spatial resolutions (0.04° × 0.04° to 0.25° × 0.25°). Moreover, satellite rainfall is usually very inaccurate compared with gauge measurements. Thus, the full utilization of satellite rainfall in hydrologic and water resources management applications has been hindered.

To overcome this dilemma and rationally utilize satellite rainfall information, recently great efforts have been dedicated to merging satellite and gauge rainfall data. Through blending the spatially continuous but coarse satellite rainfall with discrete but accurate gauge measurements, a new kind of rainfall with finer resolution can be generated. Because merging would offset the measurement errors of the two rainfall estimates, the quality of the combined rainfall may be improved to some degree. At present, various rainfall merging schemes have been developed for experimental or operational use, such as conditional merging (Sinclair and Pegram, 2005), Bayesian merging (Todini, 2001), statistical objective analysis (Pereira Filho, 2004).

Although various schemes have been developed, rainfall merging is still a complex and important issue. The results of rainfall merging are influenced by the kind of merging scheme, the quality of satellite rainfall data, the density of raingauges and so on. Motivated by this, the objective of this paper is to develop a residual-based method for merging satellite and raingauge rainfall using geographically weighted regression (GWR). Theoretically, this novel method is capable of simultaneously blending various satellite rainfall data with gauge measurements and could describe the non-stationary influences of geographical and terrain factors on rainfall spatial distribution. Using the proposed method, an experimental study on merging the rainfall from CMOROH (Joyce *et al*., 2004) and gauge measurements was conducted for the Ganjiang River basin, in southeast China. The capability of our merging scheme for constructing daily rainfall fields under different gauge densities is investigated and discussed. The accuracy gain achieved by rainfall merging relative to traditional interpolation merely only raingauge measurements is analysed.

STUDY AREA AND DATASET

Study area

The Ganjiang River Basin is located between 113°30′E−116°40′E and 24°29′N−29°11′N in Southeast China. With a drainage area of 83 374 km^2, it is a major sub-catchment of Poyang Lake, the largest freshwater lake in China (Fig. 1). The study area is one of the typical rainstorm regions in China. Mean annual precipitation is about 1580 mm in this region.

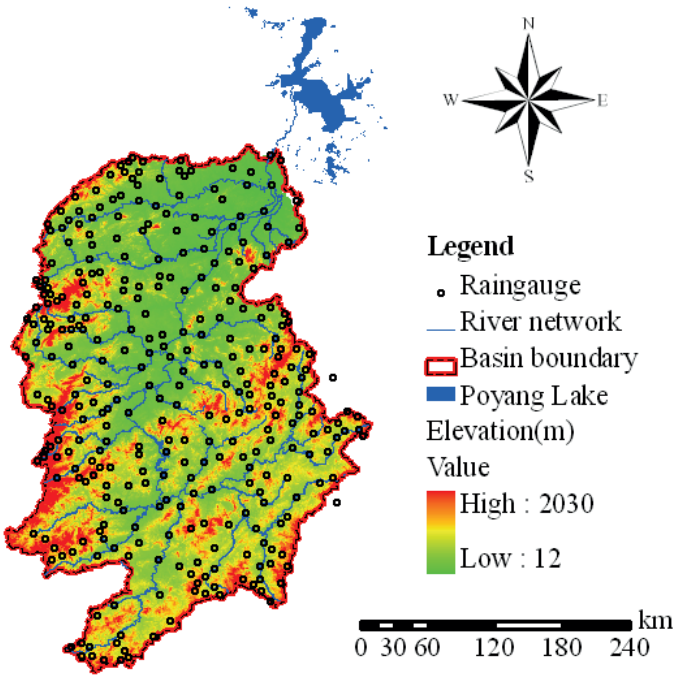

Fig. 1 Sketch map of the location of the study area and the raingauges' distribution.

Dataset

This study area has a dense raingauge network consisting of 325 stations (Fig. 1). These gauges are well-distributed spatially and their density is about one per 256 km^2. The quality of all the rainfall measurements has been proven by strict quality checks and control. Using these observations, point-wise daily rainfall series were obtained for the period of 2003–2009.

Satellite rainfall from CMORPH during the period of 2003–2009 was also collected. The spatial and temporal resolutions of CMORPH are0.25°×0.25° and half-hourly, respectively. The daily rainfall series is obtained by accumulating the rainfall of 48 half-hour episodes within a day.

METHODOLOGY

GWR background

GWR is a type of regression model with spatially varying coefficients (Fotheringham *et al*., 2003). It enables a non-stationary relationship between the variables in the regression model. By calculating local statistics, spatial relationships can be identified and utilized for prediction. GWR also disaggregates spatial patterns in the model residuals and reduces the spatial autocorrelation. The basic formula of GWR is expressed as:

$$Y_i = \beta_{i0} + \sum_{k=1}^{p} \beta_{ik}(u_i, v_i) X_{ik} + \varepsilon_i \qquad i = 1,2,\ldots,n \tag{1}$$

where Y_i and X_{ik} are respectively the dependent and k-th independent variable at location i; u_i and v_i are the coordinates; $\beta_{i0}(u_i, v_i)$ is the intercept, $\beta_{ik}(u_i, v_i)$ is the local regression parameter for X_{ik} and ε_i is the residual, p is the number of independent variables and n is the number of observations.

Equation (1) could be rewritten using a matrix method:

$$\boldsymbol{Y} = \boldsymbol{X} \otimes \boldsymbol{\beta}' + \boldsymbol{\varepsilon} \tag{2}$$

where $\otimes$ is the sign for logical multiplication; $\boldsymbol{\varepsilon}$ is the error vector; $\boldsymbol{X}$ and $\boldsymbol{\beta}$ are two matrices consisting of independent variables and local regression coefficients, respectively:

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{X_1} \\ \boldsymbol{X_2} \\ \vdots \\ \boldsymbol{X_n} \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1p} \\ 1 & X_{21} & X_{22} & \cdots & X_{2p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{np} \end{bmatrix} \tag{3}$$

$$\boldsymbol{\beta} = [\boldsymbol{\beta}_1 \quad \boldsymbol{\beta}_2 \quad \cdots \quad \boldsymbol{\beta}_n] = \begin{bmatrix} \beta_{10} & \beta_{20} & \cdots & \beta_{n0} \\ \beta_{11} & \beta_{21} & \cdots & \beta_{n1} \\ \vdots & \vdots & \vdots & \vdots \\ \beta_{1p} & \beta_{2p} & \cdots & \beta_{np} \end{bmatrix} \tag{4}$$

The number of unsolved parameters in equation (2) is $n \times (p + 1)$, which exceeds the number of observations. To solve this equation, GWR estimates the coefficients using local weighted least-squares regression:

$$\widehat{\boldsymbol{\beta}}_i = (\boldsymbol{X}'\boldsymbol{W}_i\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W}_i\boldsymbol{Y} \tag{5}$$

where $\widehat{\boldsymbol{\beta}}_i$ is the coefficient vector for location i and $\boldsymbol{W}_i$ is the spatial weight matrix:

$$\boldsymbol{W}_i = \begin{bmatrix} w_{i1} & 0 & \cdots & 0 \\ 0 & w_{i1} & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & w_{in} \end{bmatrix} \tag{6}$$

The estimates calculated by $\widehat{\boldsymbol{\beta}}_i$ for the observation at location i are as follows:

$$\hat{y}_i = \boldsymbol{X}_i(\boldsymbol{X}'\boldsymbol{W}_i\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W}_i\boldsymbol{Y} \tag{7}$$

GWR assumes that observations closer together will have more impact on each other than on observations further apart. Hence, a distance decay kernel function is employed for the spatial weight matrix. When the distance between observations is greater than the kernel bandwidth, the weight rapidly approaches zero. In summary, the kernel function could be grouped into two types, namely the fixed and adaptive bandwidths. The former calculates a bandwidth that is held constant over space, whereas the latter can adapt bandwidth distance in relation to variable density; bandwidths are smaller where data are dense and larger when data are sparse.

In this study, the adaptive kernel bandwidth was used as sample densities varied spatially. The weight using the exponential distance decay function is calculated as:

$$w_{ij} = \begin{cases} [1 - (d_{ij}/d_{ik})^2]^2 & d_{ij} \le d_{ik} \\ 0 & d_{ij} > d_{ik} \end{cases} \tag{8}$$

where w_{ij} is the weight of observation j for observation i; d_{ij} is the distance between observation i and j; d_{ik} is the distance between observation i and its k-th nearest neighbour.

The key step of calibrating GWR is to determining the optimal bandwidth distance (i.e. d_{ik}). In this paper, it was determined automatically using the corrected Akaike information criterion (AICC) (Fotheringham *et al*., 2003).

GWR based merging

A residual-based analysis is proposed for merging satellite and raingauge rainfall. It estimates a preliminary rainfall field, known as the background field, using satellite rainfall and estimates the residual field using residuals at observed points considering the influences of some related variables. The merged field is then given by the combination of the predicted error field and background field.

Taking $\boldsymbol{P}^B$ and $\boldsymbol{P}^O$ as the notation for the background field and observed field respectively, the relationship between them and the true field $\boldsymbol{P}^T$ is expressed as:

$$\boldsymbol{P}^T = \boldsymbol{P}^B + \boldsymbol{e}^B \tag{9}$$

$$\boldsymbol{P}^T = \boldsymbol{P}^O + \boldsymbol{e}^O \tag{10}$$

where $\boldsymbol{e}^B$ and $\boldsymbol{e}^O$ represent the background and observation errors. Here, the expection of $\boldsymbol{e}^B$ and $\boldsymbol{e}^O$ is denoted using μ_B and μ_O, and the variation is denoted by σ_B^2 and σ_O^2, respectively.

Under the assumption of μ_O equal to zero and σ_B^2 much larger than σ_O^2, the following equation can be derived:

$$\boldsymbol{e}^B \approx \boldsymbol{P}^O - \boldsymbol{P}^B \tag{11}$$

Equation (11) implies that the residual field could be approximated by the difference between the observation and background fields. However, considering P^O is just known at limited locations, it is required to estimate $\boldsymbol{e}^B$ at those locations without gauge observations. Assuming that background errors are generally correlated in space, this issue can be resolved through local interpolation using some nearby values with observations.

Based on the framework of the residual-based merging, this paper proposed the merging scheme based on GWR. This method has three main steps. First, to construct a background using GWR, we describe the relationship between the background rainfall and the satellite estimates at any place using local regression:

$$P_i^B = b_{i0} + \sum_{k=1}^{p} b_{ik} P_{ik}^S + \varepsilon_i \tag{12}$$

where P_{ik}^S is the estimate corresponding to the k-th kind of satellite rainfall at location i; b_{i0} is the intercept, b_{ik} is the local regression parameter. b_{i0} and b_{ik} are both probably non-stationary in space.

Secondly, also using GWR, at those locations without observations $\boldsymbol{e}^B$ is calculated. We assumed that the relationship between $\boldsymbol{e}^B$ and geographic factors including coordinates u, v and elevation z could be described using a locally non-stationary regression equation:

$$e_i^B = \beta_{i0} + \beta_{i1}\mu_i + \beta_{i2}v_i + \beta_{i3}z_i + \varepsilon_i \tag{13}$$

Thirdly, the merged field $\boldsymbol{P}^M$ can be obtained by combining the background field and the estimated residual:

$$P_i^M = b_{i0} + \sum_{k=1}^{p} b_{ik}P_{ik}^S + \varepsilon_i \ + \beta_{i0} + \beta_{i1}\mu_i + \beta_{i2}v_i + \beta_{i3}z_i + \varepsilon_i \tag{14}$$

After combination of the similar items, equation (14) can be rewritten as:

$$P_i^M = \beta_{i0} + \beta_{i1}\mu_i + \beta_{i2}v_i + \beta_{i3}z_i + \sum_{k=4}^{p} b_{ik}P_{ik}^S + \varepsilon_i \tag{15}$$

Equation (15) is a general form for the rainfall merging scheme based on GWR. As a regression model, the number of satellite rainfall in the merging method is theoretically limitless. Thus, this proposed method is capable of simultaneously blending multiple kinds of satellite rainfall data with gauge measurements. At the same time, equation (15) describes the non-stationary influences of geographical and terrain factors on the rainfall spatial distribution. Although, the gauge rainfall observations are not seen directly in the regression model, their effect on the merged results is indirectly reflected via the spatially varying regression coefficients derived by equation (5).

Performance assessment

After the coefficients in equation (15) are optimized using AICC, the merged rainfall at any location within the study area can be estimated. We divided all the rainfall data from 325 gauges in the Ganjiang River basin into two parts. One part was selected as the calibration data for the GWR merging model and the remainder was used for validation. Then, the merged rainfall was compared with the validation data and two performance indices: the mean absolute error (MAE) and spatial correlation coefficient (CC) were calculated. For one day, the two indices are calculated as follows:

$$\mathrm{MAE} = \sum_{i=1}^{n_v} \left| P_i^M - P_i^O \right| / n_v \tag{16}$$

$$\mathrm{CC} = \sum_{i=1}^{n_v} \left(P_i^M - \bar{P}_i^M \right) \left(P_i^O - \bar{P}_i^O \right) / \sqrt{\sum_{i=1}^{n_v} (P_i^M - \bar{P}_i^M)^2 \, (P_i^O - \bar{P}_i^O)^2} \tag{17}$$

where n_v is the number of observation for validation, and $\bar{P}_i^O$ and $\bar{P}_i^M$ stand for the average values at the validation locations.

To explore the accuracy gained by merging surface measurements with CMORPH rainfall relative to traditional interpolation using only gauge measurement, two kinds of daily rainfall fields were generated. The first was generated by the GWR based merging scheme both using CMORPH and gauge rainfall as the data sources, whereas the second was generated by GWR interpolation using only the same gauge rainfall (see equation (18)). Here, we use GWR-M and GWR-I to denote the two rainfall field construction methods, respectively. Then, under different calibration gauge data, MAE and CC for GWR-M and GWR-I were calculated respectively.

$$P_i^I = \beta_{i0} + \beta_{i1}\mu_i + \beta_{i2}v_i + \beta_{i3}z_i + \varepsilon_i \tag{18}$$

To evaluate the performance improvement gained by GWR-M relative to GWR-I, we further calculate the ratio of MAE and CC between the two methods:

$$R_{\mathrm{MAE}} = 1 - \mathrm{MAE}_M / \mathrm{MAE}_I \tag{19}$$

$$R_{\mathrm{CC}} = \mathrm{CC}_M / \mathrm{CC}_I - 1 \tag{20}$$

where MAE_M and MAE_I mean the MAE for GWR-M and GWR-I, respectively, and similarly for CC_M and CC_I. When R_{MAE} and R_{CC} are positive, the error magnitude of the estimated rainfall field produced by the merging scheme is lower than with interpolation and the spatial structure is also raised.

RESULT AND DISCUSSION

For the Ganjiang River basin, an experimental study on CMORPH and gauge rainfall merging was conducted. Using GWR-M and GWR-I, two sets of daily rainfall fields were generated. The rainfall fields are all at a spatial resolution of 1 km × 1 km. To investigate the accuracy gain by GWR-M relative to GWR-I under different gauge densities, we gradually changed the number of raingauges data for model calibration and calculated the accuracy indices using the validation data.

Table 1 shows the results for GWR-M and GWR-I. In Table 1, MAE and CC for the two kinds of daily rainfall fields are mean values for the 2557 days from 2003 to 2009; the raingauge relative density (denoted using R_d) means the number of calibration raingauges divided by the total 325 gauges over the study area. For example, when R_d is 2/3, it means that daily rainfall data from 2/3 of the 325 gauges were selected for calibration while the other 1/3 were used for validation. It is seen from Table 1 that both accuracy indices for GWR-M and GWR-I are improved as R_d increases. This phenomenon is easy to recognize because the efficient information provided by surface measurements for the analysed rainfall fields is approximately proportional to the raingauge density. However, the change ratios of MAE and CC with R_d are not even. When R_d is less than 1/5, MAE and CC are rather sensitive to the increasing of R_d. However, when R_d exceeds 1/5, the changing traits for MAE and CC are reversed.

Table 1 Accuracy indices for daily rainfall fields generated by GWR-M and GWR-I during 2003–2009 in the Ganjiang River basin.

Method	Accuracy index	The relative density of raingauges for calibration (R_d)											
		2/3	1/2	1/3	1/4	1/5	1/6	1/8	1/10	1/12	1/15	1/20	1/30
GWR-I	CC	0.63	0.61	0.57	0.54	0.53	0.51	0.49	0.47	0.44	0.41	0.39	0.33
	MAE (mm/d)	2.37	2.46	2.62	2.74	2.82	2.91	3.05	3.18	3.33	3.52	3.75	4.2
GWR-M	CC	0.63	0.61	0.59	0.56	0.55	0.54	0.52	0.51	0.5	0.48	0.47	0.44
	MAE (mm/d)	2.34	2.42	2.56	2.67	2.73	2.8	2.89	2.98	3.09	3.21	3.34	3.59

At the same time, Table 1 indicates that the CC as well as MAE is improved by GWR-M over GWR-I. In general, the performance of GWR-M is more or less higher than GWR-I under various scenarios of gauge density. Hence, gains are obtained by merging gauge measurements with CMORPH. This kind of satellite rainfall data provides useful information for generating daily rainfall fields. Figure 2 shows the curves of R_{MAE}–R_d and R_{cc}–R_d. According to these two curves, R_{MAE} and R_{cc} are obviously higher than zero when R_d is lower than 1/5, or else R_{MAE} and R_{cc} are close to zero. This reveals that, for the study area, the gain achieved by GWR-M is only substantial when the raingauges are rather sparse in space. For R_d at 1/10, the improvements of MAE and CC gained by GWR-M over GWR-I are 8.5% and 6.3%, respectively; for R_d at 1/20, the corresponding values are 20.5% and 10.9%, respectively.

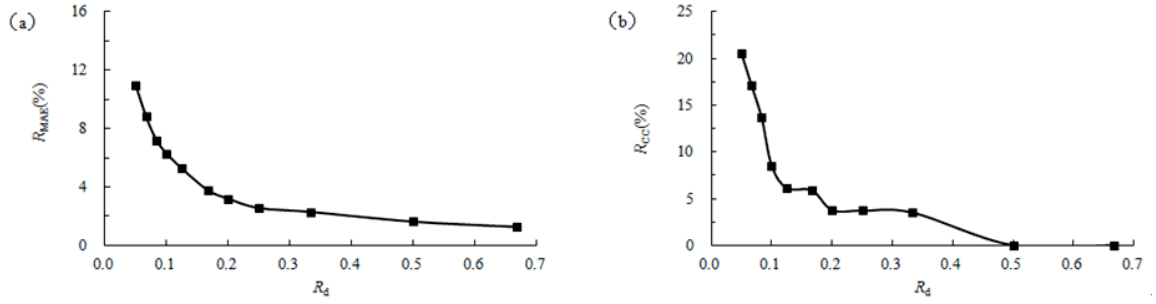

Fig. 2 R_{MAE} and R_{CC} under various scenarios of the calibration rain gauges relative to the total 325 in the Ganjiang River basin.

CONCLUSIONS

A residual-based rainfall merging method using GWR was proposed. This method is capable of simultaneously blending various satellite rainfall data with gauge measurements and could describe the non-stationary influences of geographical and terrain factors on rainfall. An experimental study for merging the satellite rainfall from CMOROH and gauge measured was conducted over the Ganjiang River basin. Main conclusions are: (1) the performance of the GWR based rainfall merging method generally improves with increasing raingauge density; in particular, when raingauges are sparse, MAE and CC will be improved rapidly with their density increase; (2) CMORPH rainfall actually provides useful information for generating daily rainfall fields. However, the gain achieved by the merging scheme relative to traditional interpolation is only substantial when the raingauges are rather sparse in space.

Acknowledgements This work was jointly supported by the National Natural Science Fund of China (Grant No 51109136; 51479118 and 51479118) and the Commonwealth Science Research Project of Ministry of Water Resources, China (Grant Nos 201301075 and 201201074).

REFERENCES

Fotheringham, A.S., Brunsdon, C., and Charlton, M. (2003) *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*. John Wiley & Sons.

Joyce, R.J., Janowiak, J. E., Arkin, P.A., *et al*. (2004) CMORPH: A method that produces global precipitation estimates from passive microwave and infrared data at high spatial and temporal resolution. *Journal of Hydrometeorology* 5(3), 487–503.

Pereira Filho, A. J. (2004) Integrating gauge, radar and satellite rainfall. In: Proceedings of the 2nd International Precipitation Working Group Workshop, Monterey, CA, USA.

Sinclair, S. and Pegram, G. (2005) Combining radar and rain gauge rainfall estimates using conditional merging. *Atmospheric Science Letters* 6(1), 19–22.

Todini, E. (1999) A Bayesian technique for conditioning radar precipitation estimates to rain-gauge measurements. *Hydrology and Earth System Sciences* 5(2), 187–199.