XL International School of Hydraulics
23 - 26 May 2023 • Kąty Rybackie • Poland

Information entropy theory-based optimizing of gauge networks for hydrological modelling - A case study in the Loess Plateau, China

Reporter: Yiwei Guo

Yiwei Guo[1,*], Haoyu Han[2], Michael Nones[1], Wentao Xu[3],Shuguang Liu[4]
yguo@igf.edu.pl

[1] Institute of Geophysics Polish Academy of Sciences, Warsaw, Poland
[2] Changjiang Survey, Planning, Design and Research Co., Ltd, Wuhan, China
[3] Changjiang River Scientific Research Institute, Changjiang Water Resource Commission, Wuhan, China
[4] Department of Hydraulic Engineering, Tongji University, Shanghai, China

Instytut Geofizyki Polskiej Akademii Nauk

2023.05

Content

Instytut Geofizyki Polskiej Akademii Nauk

01

Background and motive

Instytut Geofizyki Polskiej Akademii Nauk

1 Background and motive

- The input data will influence the accuracy of hydrological output results.
- Spatial and temporal distribution of rainfall influence the hydrological behavior of the model.

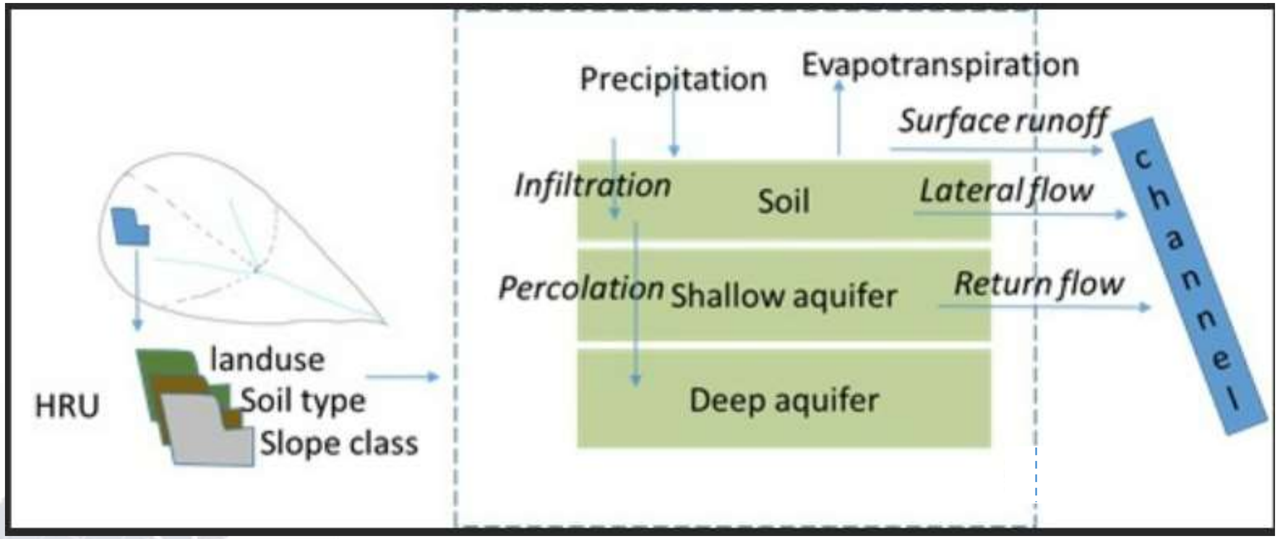


Fig. Hydrological processes simulated by SWAT model

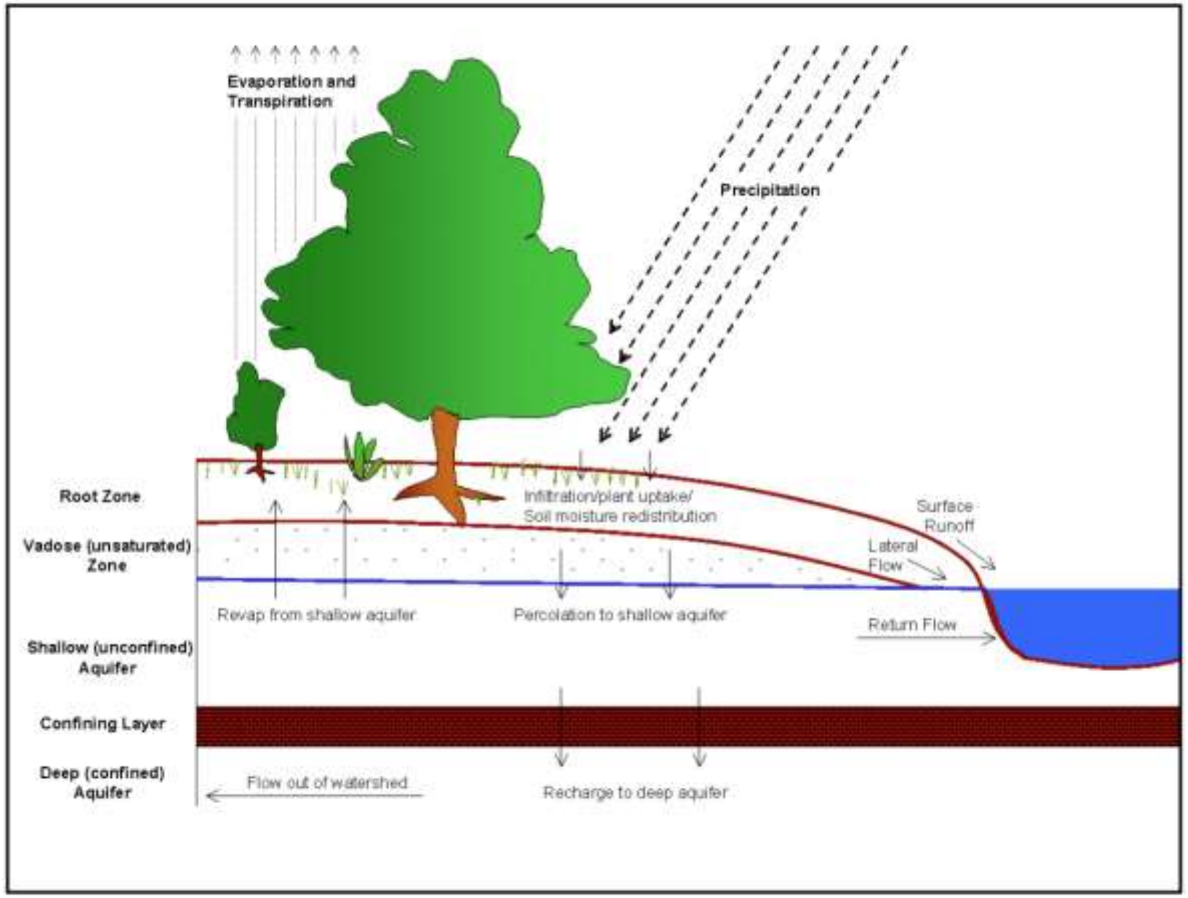


Fig. Terrestrial water circulation processes

Instytut Geofizyki Polskiej Akademii Nauk

1 Background and motive

- Spatial interpolation of rainfall at ground-based gauges are regarded as watershed areal rainfall.
- The true distribution of precipitation can't be represented well by point rainfall.
- The rain gauge network should be well designed.
- The information entropy can be used in rain gauge network optimization.

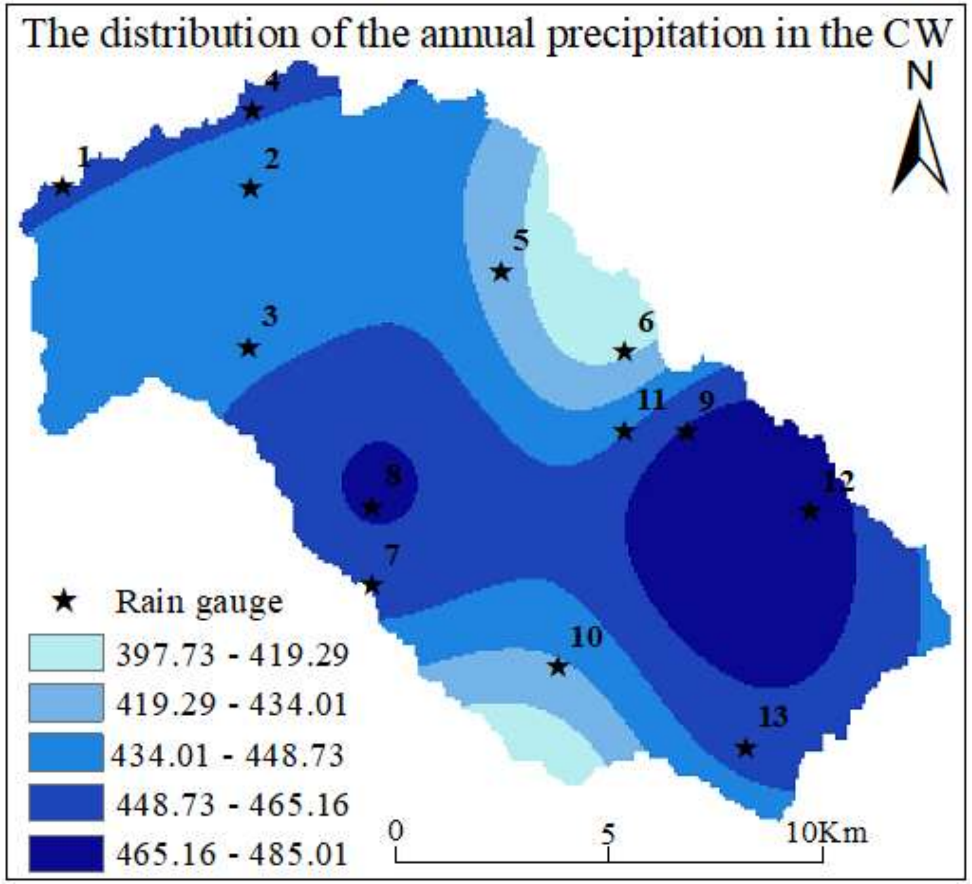


Fig. The distribution of annual precipitation, interpolated by King's method.

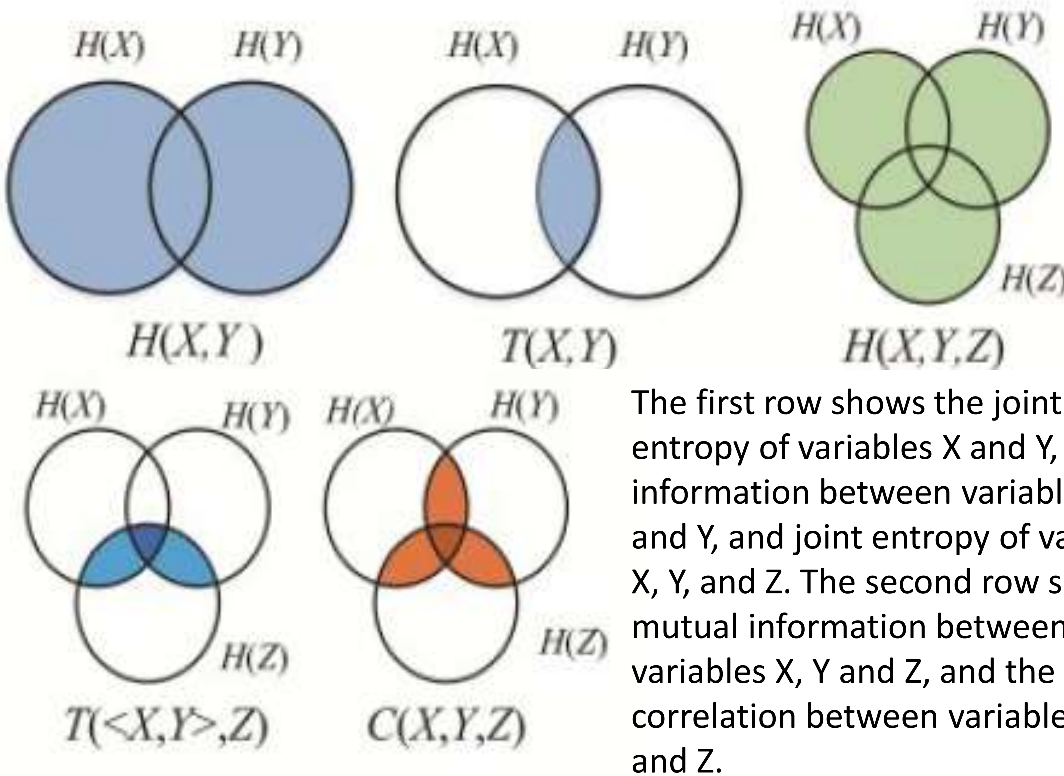
1 Background and motive

- Entropy: the mathematical foundation for measuring information or uncertainty.
- Entropy-based methods can:

1) directly define the optimization deployment information of the rain gauge network
2) quantify the uncertainty.

- The key to the design and optimization station network:

1) how much information is contained in one or several stations;
2) how much information can be transmitted from one or more stations to other stations;
3) how much information is shared among several stations.



The first row shows the joint entropy of variables X and Y, mutual information between variables X and Y, and joint entropy of variables X, Y, and Z. The second row shows mutual information between variables X, Y and Z, and the total correlation between variables X, Y, and Z.

Fig. The relationship between binary and multivariate joint entropy H, mutual information T, and total correlation C.

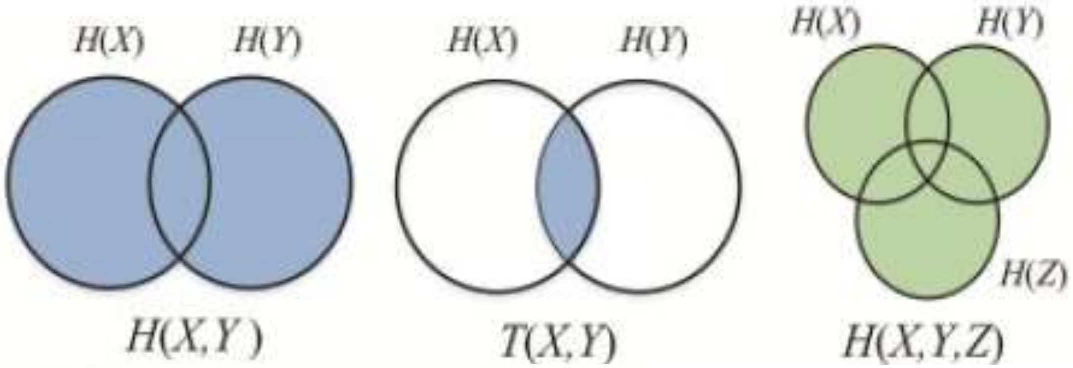
1 Background and motive

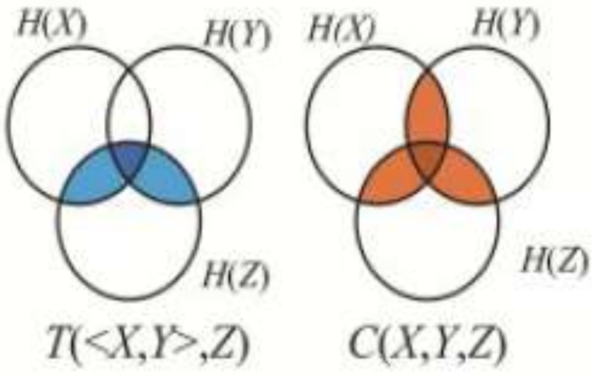
- Entropy: the mathematical foundation for measuring information or uncertainty.
- Entropy-based methods can:

1) directly define the optimization deployment information of the rain gauge network
2) quantify the uncertainty.

- The key to the design and optimization station network:

1) how much information is contained in one or several stations;
2) how much information can be transmitted from one or more stations to other stations;
3) how much information is shared among several stations.





The first row shows the joint entropy of variables X and Y, mutual information between variables X and Y, and joint entropy of variables X, Y, and Z. The second row shows mutual information between variables X, Y and Z, and the total correlation between variables X, Y, and Z.

Fig. The relationship between binary and multivariate joint entropy H, mutual information T, and total correlation C.

The step to optimize gauge network.

- measure the spatial information between rain gauges
- evaluate whether the information is sufficient
- subsequently optimize the rain gauge network

Instytut Geofizyki Polskiej Akademii Nauk

1 Background and motive

The study area (Chabagou Watershed) is in the Loess Plateau of China, where:

- ➢water resources are scarce and rainfall is concentrated and unevenly distributed
- ➢flash floods are very common
- ➢the topography of the watershed is complex
- ➢the economy is not well developed

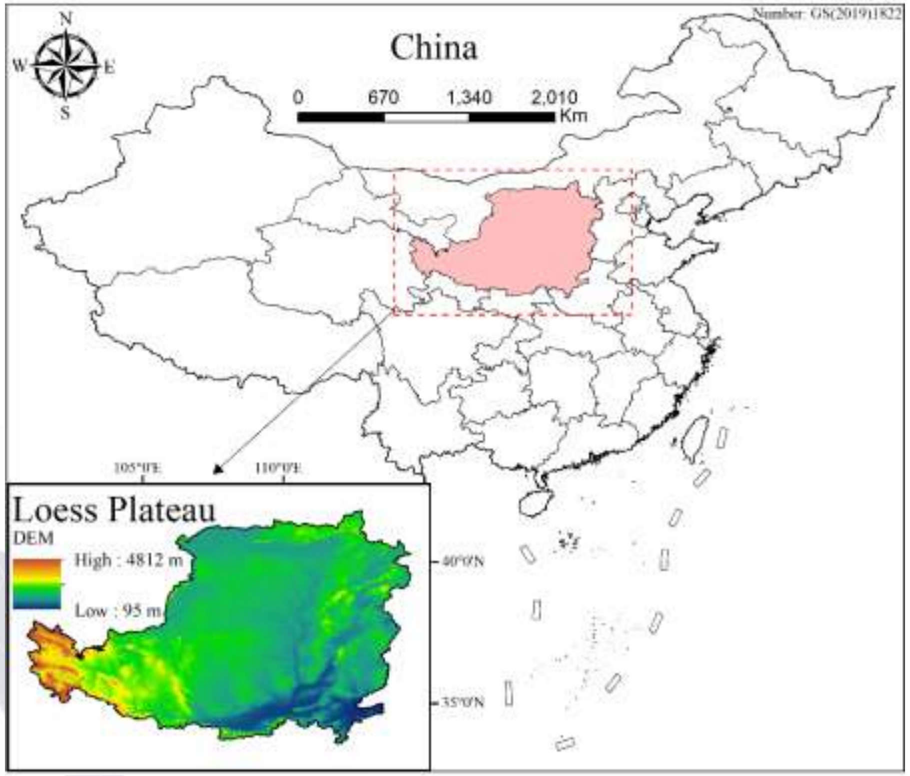


Fig. The location of the study area

1 Background and motive

The study area (Chabagou Watershed) is in the Loess Plateau of China, where:

- water resources are scarce and rainfall is concentrated and unevenly distributed
- flash floods are very common
- the topography of the watershed is complex
- the economy is not well developed

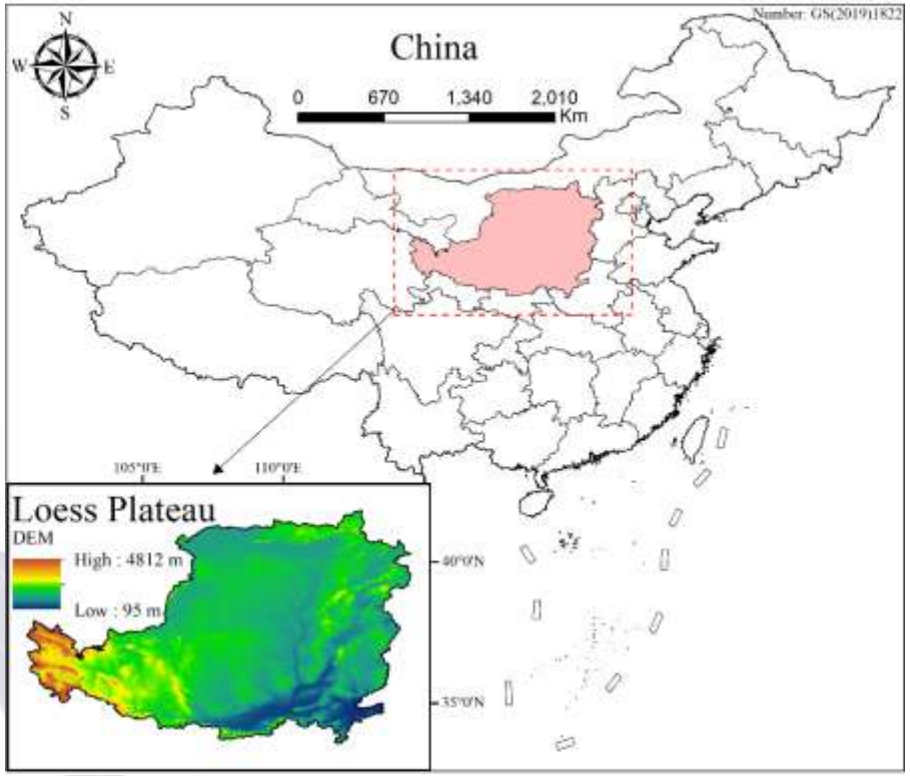


Fig. The location of the study area

Main objectives

(1) to understand the distribution of the annual precipitation and information entropy of the CW;

(2) to optimize the rain gauge network by using the MIMR based on the information entropy and evaluate the optimized rain gauge network;

(3) evaluate the impact of the different rain gauge networks on simulating the watershed hydrology via the SWAT model.

02

Study area and method

Instytut Geofizyki
Polskiej Akademii Nauk

2.1 Study area

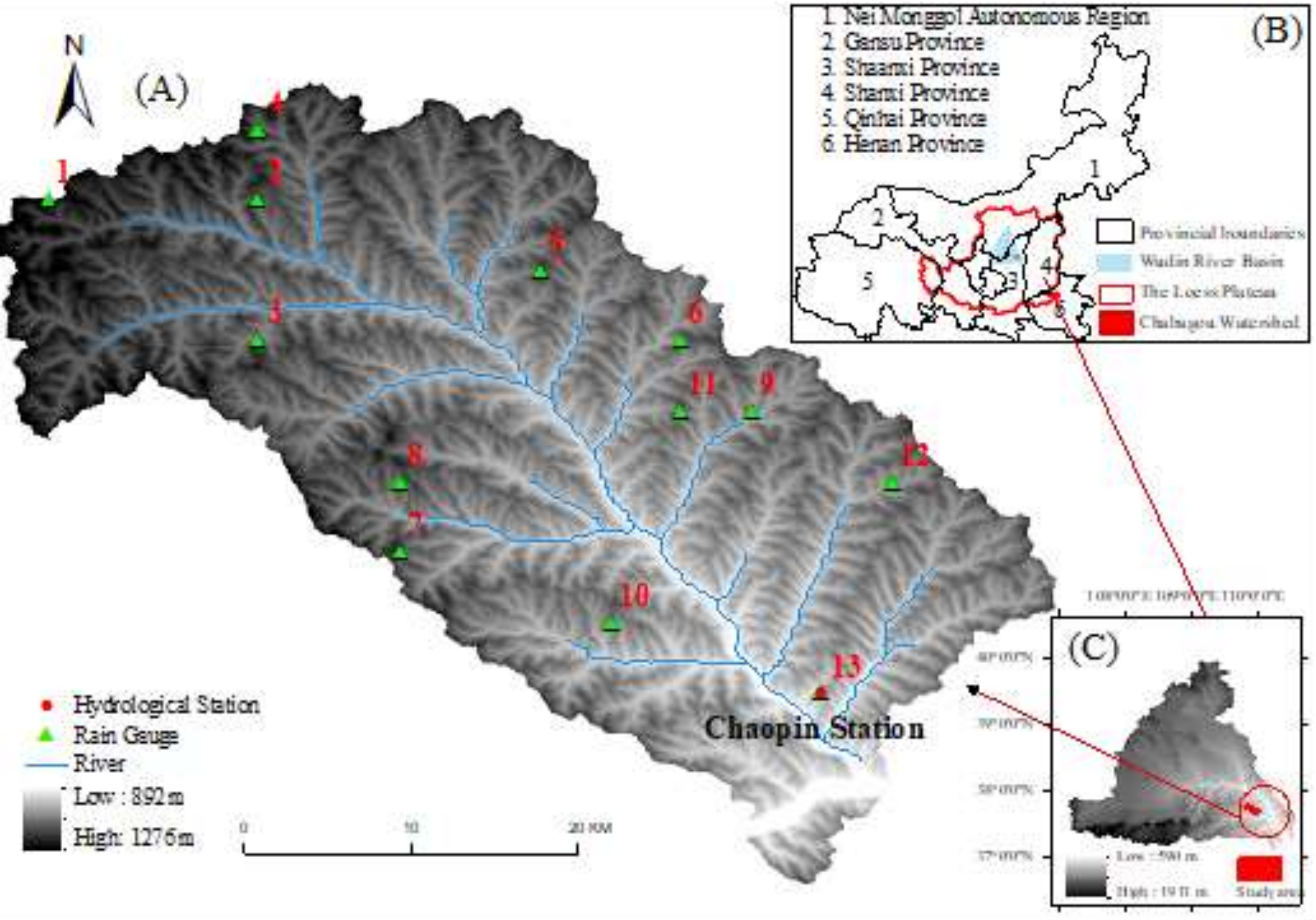


Fig. The study area

- Located in Shaanxi Province in Northwest China.
- The drainage area of the CW is 205 km^2.

Instytut Geofizyki Polskiej Akademii Nauk

2.1 Study area

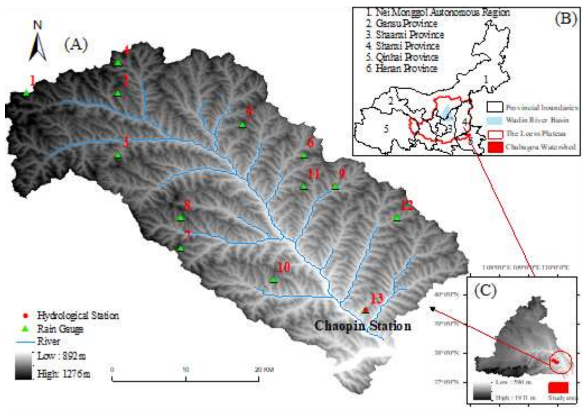


Fig. The study area

Fig. The CW

Fig. Install rain gauges at the CW

Fig. The CW

- Located in Shaanxi Province in Northwest China.
- The drainage area of the CW is 205 km^2.
- Climate type: arid and semi-arid continental monsoon climate.
- Features less rainfall and more sunshine.
- Rainfall period is June to September, accounting for about 70% of the annual precipitation .
- Annual average rainfall is 450 mm
- Annual average runoff is 3684900 m^3

Instytut Geofizyki Polskiej Akademii Nauk

2.1 Study area

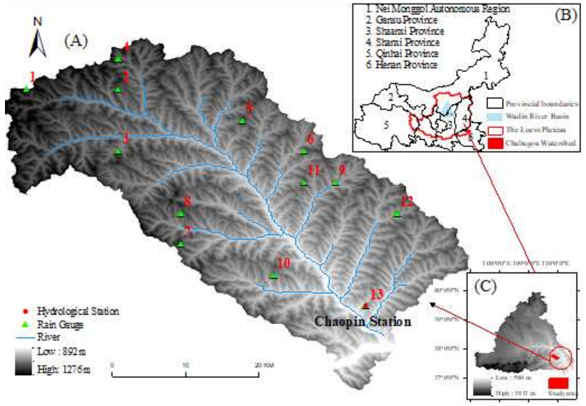


Fig. The study area


Fig. The installed rain gauge


Fig. The CW

- Located in Shaanxi Province in Northwest China.
- The drainage area of the CW is 205 km^2.
- Climate type: arid and semi-arid continental monsoon climate.
- Features less rainfall and more sunshine.
- Rainfall period is June to September, accounting for about 70% of the annual precipitation .
- Annual average rainfall is 450 mm
- Annual average runoff is 3684900 m^3
- Mainly soil type is loess soil, which has a soft structure and is easy to be eroded
- Cropland is the dominant land-use type

2.2 method

Calculate the joint entropy of every gauges

Instytut Geofizyki
Polskiej Akademii Nauk

2.21 method—— Entropy

Marginal entropy

Describe the degree of discreteness and uncertainty of a random variable X, where higher discreteness corresponds to greater uncertainty.

$$H(X) = -k\sum_{i=1}^{n} p(x_i)\log_b p(x_i)$$

- k is an arbitrary positive constant, and in this study, we take k=1.
- The dimension of entropy varies with the base b used, with bit (Binary Digit) being the dimension when b=2, nat (Natural Digit) being the dimension when b=e (natural logarithm base), and dit (Decimal Digit) being the dimension when b=10. In this study, we use b=2.

2.21 method—— Entropy

Marginal entropy

Describe the degree of discreteness and uncertainty of a random variable X, where higher discreteness corresponds to greater uncertainty.

$$H(X)=-k\sum_{i=1}^{n} p(x_i)\log_b p(x_i)$$

Joint entropy

- For a multidimensional random variable, joint entropy is defined as a measure of the total information retained by the variables.
- By extending the concept to two random variables, the total information retained by a multidimensional random variable can be obtained.

$$H(X,Y)=-\sum_{i=1}^{n} p(x_i,y_j)\log_2 p(x_i,y_j)$$

$$H(X_1,X_2,...,X_n)=$$
$$-\sum_{x_1}\sum_{x_2}...\sum_{x_n} p(x_1,...,x_n)\log_2 p(x_1,...,x_n)$$

Instytut Geofizyki Polskiej Akademii Nauk

- k is an arbitrary positive constant, and in this study, we take k=1.
- The dimension of entropy varies with the base b used, with bit (Binary Digit) being the dimension when b=2, nat (Natural Digit) being the dimension when b=e (natural logarithm base), and dit (Decimal Digit) being the dimension when b=10. In this study, we use b=2.

2.21 method—— Entropy

Marginal entropy

Describe the degree of discreteness and uncertainty of a random variable X, where higher discreteness corresponds to greater uncertainty.

$$H(X)=-k\sum_{i=1}^{n} p(x_i)\log_b p(x_i)$$

Joint entropy

- For a multidimensional random variable, joint entropy is defined as a measure of the total information retained by the variables.
- By extending the concept to two random variables, the total information retained by a multidimensional random variable can be obtained.

$$H(X,Y)=-\sum_{i=1}^{n} p(x_i,y_j)\log_2 p(x_i,y_j)$$

$$H(X_1,X_2,\ldots,X_n)=$$
$$-\sum_{x_1}\sum_{x_2}\cdots\sum_{x_n} p(x_1,\ldots,x_n)\log_2 p(x_1,\ldots,x_n)$$

Mutual information

- Mutual information describes the amount of shared information between two random variables, and its magnitude reflects the degree of correlation between the variables.
- It is superior to Pearson correlation coefficient.
- It captures both linear and nonlinear dependencies.

$$T(X,Y)=$$
$$-\sum_{x}\sum_{y} p(x_i,y_j)\log_2 \frac{p(x_i,y_j)}{p(x_i)p(y_j)}$$

$$T(X,Y)=H(X)+H(Y)-H(X,Y)$$

- k is an arbitrary positive constant, and in this study, we take k=1.
- The dimension of entropy varies with the base b used, with bit (Binary Digit) being the dimension when b=2, nat (Natural Digit) being the dimension when b=e (natural logarithm base), and dit (Decimal Digit) being the dimension when b=10. In this study, we use b=2.

2.21 method—— Entropy

Marginal entropy

Describe the degree of discreteness and uncertainty of a random variable X, where higher discreteness corresponds to greater uncertainty.

$$H(X)=-k\sum_{i=1}^{n} p(x_i)\log_b p(x_i)$$

Joint entropy

- For a multidimensional random variable, joint entropy is defined as a measure of the total information retained by the variables.
- By extending the concept to two random variables, the total information retained by a multidimensional random variable can be obtained.

$$H(X,Y)=-\sum_{i=1}^{n} p(x_i,y_j)\log_2 p(x_i,y_j)$$

$$H(X_1,X_2,\ldots,X_n)=$$

$$-\sum_{x_1}\sum_{x_2}\cdots\sum_{x_n} p(x_1,\ldots,x_n)\log_2 p(x_1,\ldots,x_n)$$

Mutual information

- Mutual information describes the amount of shared information between two random variables, and its magnitude reflects the degree of correlation between the variables.
- It is superior to Pearson correlation coefficient.
- It captures both linear and nonlinear dependencies.

$$T(X,Y)=$$

$$-\sum_{x}\sum_{y} p(x_i,y_j)\log_2 \frac{p(x_i,y_j)}{p(x_i)p(y_j)}$$

$$T(X,Y)=H(X)+H(Y)-H(X,Y)$$

Total correlation

$$C(X_1,X_1,\ldots,X_n)=$$

$$-\sum_{i=1}^{n} H(X_i)-H(X_1,X_2,\ldots X_n)$$

Total correlation describes the information redundancy between multidimensional random variables Namely, the measure of the amount of repeated information between the variables.

- k is an arbitrary positive constant, and in this study, we take k=1.
- The dimension of entropy varies with the base b used, with bit (Binary Digit) being the dimension when b=2, nat (Natural Digit) being the dimension when b=e (natural logarithm base), and dit (Decimal Digit) being the dimension when b=10. In this study, we use b=2.

2.22 method——MIMR

Calculate the joint entropy of every gauges

↓

Use the Maximum Information Minimum Redundancy (MIMR) to optimize the gauge network.

MIMR

Rank the importance of hydrological network sites to select a set of sites that maximize overall information, maximize information transfer capacity, and minimize redundant information.

Instytut Geofizyki Polskiej Akademii Nauk

2.22 method——MIMR

MIMR

Rank the importance of hydrological network sites to select a set of sites that maximize overall information, maximize information transfer capacity, and minimize redundant information.

Calculate the joint entropy of every gauges

⇩

Use the Maximum Information Minimum Redundancy (MIMR) to optimize the gauge network.

STEP

1. Set a threshold for joint entropy.
2. Calculated the marginal entropy of each site.
3. Select the station with the highest H as the central station.
4.
5.

Instytut Geofizyki Polskiej Akademii Nauk

2.22 method——MIMR

MIMR

Rank the importance of hydrological network sites to select a set of sites that maximize overall information, maximize information transfer capacity, and minimize redundant information.

Calculate the joint entropy of every gauges

↓

Use the Maximum Information Minimum Redundancy (MIMR) to optimize the gauge network.

STEP

1. Set a threshold for joint entropy.
2. Calculated the marginal entropy of each site.
3. Select the station with the highest H as the central station.
4. Use MIMR to rank the remaining sites.
5.

$$\begin{cases} Max.H(X_{S_1}, X_{S_2}, \ldots, X_{S_k}) \\ Max.T(< X_{S_1}, X_{S_2}, \ldots, X_{S_k} >, \\ \qquad < X_{F_1}, X_{F_2}, \ldots, X_{F_m} >) \\ Min.C(X_{S_1}, X_{S_2}, \ldots, X_{S_k}) \end{cases}$$

Assuming there are N sites in total, S represents the k sites already selected in the optimal network, and F represents the remaining m sites to be selected, where k+m=N.

↓ **Simplify the calculation.**

$\lambda 1$ and $\lambda 2$ are weights, $\lambda 1+\lambda 2=1$

$$Max.\ \lambda_1 (H(X_{S_1}, X_{S_2}, \ldots, X_{S_k}) + T(< X_{S_1}, X_{S_2}, \ldots, X_{S_k} >, < X_{F_1}, X_{F_2}, \ldots, X_{F_m} >)) - \lambda_2 C(X_{S_1}, X_{S_2}, \ldots, X_{S_k})$$

2.22 method——MIMR

MIMR

Rank the importance of hydrological network sites to select a set of sites that maximize overall information, maximize information transfer capacity, and minimize redundant information.

Calculate the joint entropy of every gauges

Use the Maximum Information Minimum Redundancy (MIMR) to optimize the gauge network.

$$\eta \geq \frac{H(X_{S_1}, X_{S_2}, ..., X_{S_k})}{H(X_{S_1}, X_{S_2}, ..., X_{S_k}, X_{F_1}, X_{F_2}, ..., X_{F_m})}$$

When η≥95%, the optimal final decision for the hydrological network site is obtained.

Instytut Geofizyki Polskiej Akademii Nauk

STEP

1. Set a threshold for joint entropy.
2. Calculated the marginal entropy of each site.
3. Select the station with the highest H as the central station.
4. Use MIMR to rank the remaining sites.
5. The final decision made when the η reached.

$$\begin{cases} Max.H(X_{S_1}, X_{S_2}, ..., X_{S_k}) \\ Max.T(< X_{S_1}, X_{S_2}, ..., X_{S_k} >, \\ \quad < X_{F_1}, X_{F_2}, ..., X_{F_m} >) \\ Min.C(X_{S_1}, X_{S_2}, ..., X_{S_k}) \end{cases}$$

Assuming there are N sites in total, S represents the k sites already selected in the optimal network, and F represents the remaining m sites to be selected, where k+m=N.

Simplify the calculation.

λ1 and λ2 are weights, λ1+λ2=1

$$Max.\ \lambda_1 (H(X_{S_1}, X_{S_2}, ..., X_{S_k}) + T(< X_{S_1}, X_{S_2}, ..., X_{S_k} >, < X_{F_1}, X_{F_2}, ..., X_{F_m} >)) - \lambda_2 C(X_{S_1}, X_{S_2}, ..., X_{S_k})$$

2.2 method

Calculate the joint entropy of every gauges

↓

Use the Maximum Information Minimum Redundancy (MIMR) to optimize the gauge network.

↓

Compare the areal precipitation before and after optimization.

Use

- r : correlation coefficient,
- PBIAS: percent bias,
- NSE: Nash-Sutcliffe efficiency coefficient

to evaluate rainfall.

2.2 method

Calculate the joint entropy of every gauges

Use the Maximum Information Minimum Redundancy (MIMR) to optimize the gauge network.

Compare the areal precipitation before and after optimization.

Use
- r : correlation coefficient,
- PBIAS: percent bias,
- NSE: Nash-Sutcliffe efficiency coefficient

to evaluate rainfall.

Comparing the ability of rainfall-driven SWAT models to simulate runoff before and after rainfall station network optimization
Use SWAT-CUP to calibrate the result.

Use
- r^2 : coefficient of determination,
- PBIAS: percent bias,
- NSE: Nash-Sutcliffe efficiency coefficient

to evaluate rainfall.

Instytut Geofizyki Polskiej Akademii Nauk

2.23 method——Data source

Table. Used data and their source

Data name	Revelation	Period	Source
Daily precipitation		2007-2016	Yellow River Water Conservancy Commission of the Ministry of water resources of China
Daily runoff		2007-2016	Yellow River Water Conservancy Commission of the Ministry of water resources of China
Shuttle Radar Topography Mission (SRTM)	30m		https://earthexplorer.usgs.gov/
30m-resolution Global Land Cover (GLC30)	30m	2015	http://www.globallandcover.com/
World Soil Database (HWSD)	1 km	1971-1981	http://www.fao.org/soils-portal/soil-survey/soil-maps-and-databases/harmonized-world-soil-database-v12/en/

Instytut Geofizyki Polskiej Akademii Nauk

03

Results

Instytut Geofizyki
Polskiej Akademii Nauk

3 Result

- The information entropy around Stations 7 and 8 was higher than that around other stations.
- The information entropy was the lowest around Stations 5, 6, and 13.
- Reason why station has more information: the transpiration of vapour carried by the monsoon being blocked by the terrain.

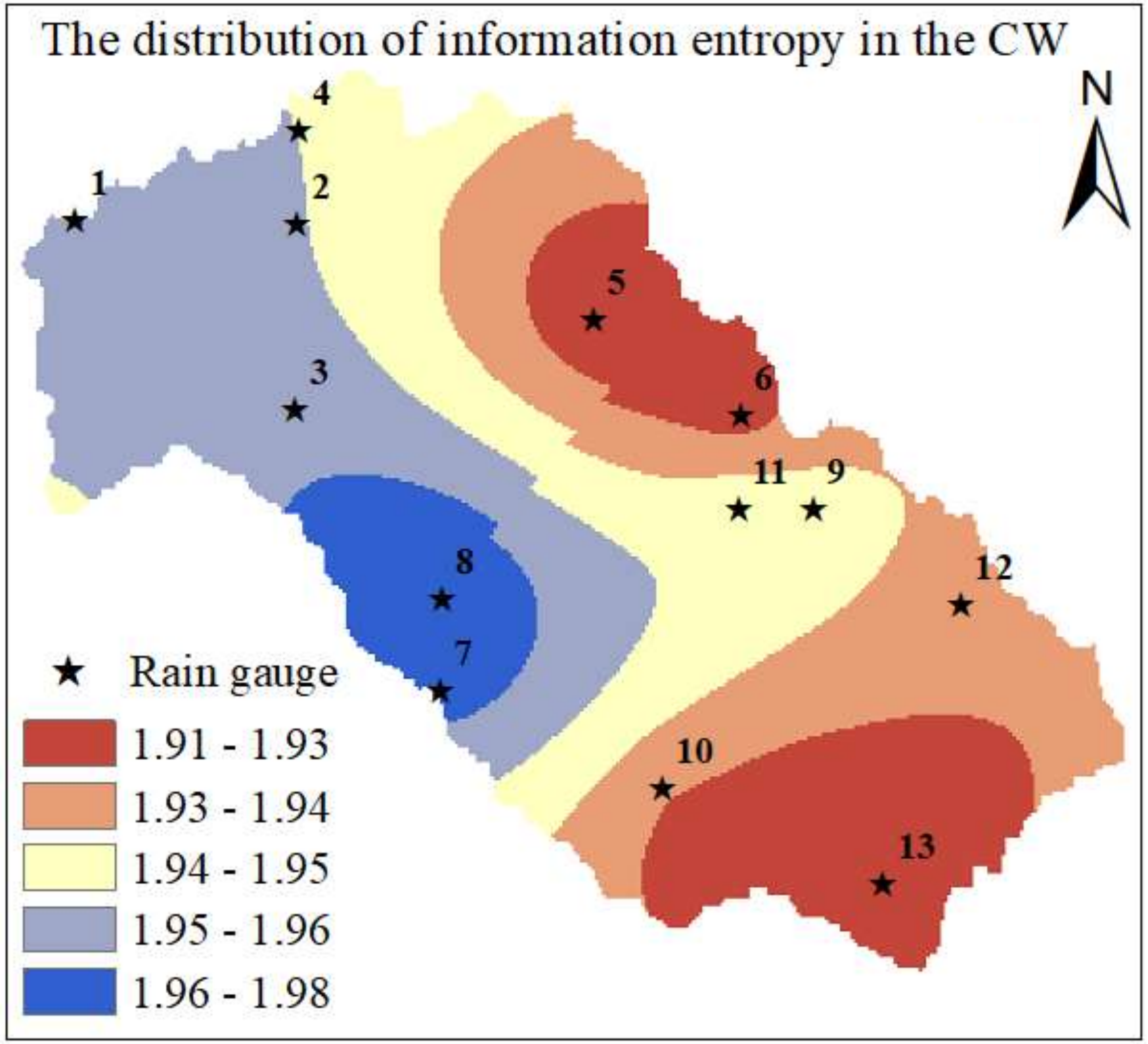


Fig. The distribution of information entropy in the CW interpolated by Ordinary Kriging method .

Instytut Geofizyki Polskiej Akademii Nauk

3 Result

- ➢The joint entropy and total correlation variations with more stations being considered.
- ➢The mutual information keeps increasing until the joint entropy reaches stable and then decreases.
- ➢The redundant information between stations increases with more stations being added and the repetitive information becomes more.

Table. The MIMR calculation table of the CW at the daily scale

Station	8	1	12	13	7	4	9	6	5	3	10	2	11
H	2.00	2.89	3.32	3.52	3.65	3.75	3.79	3.83	3.85	3.89	3.92	3.95	3.98
T	1.92	2.76	3.15	3.25	3.33	3.36	3.36	3.34	3.28	3.12	2.78	1.93	
C	0.00	1.08	2.57	4.27	6.11	7.97	9.89	11.74	13.61	15.53	17.42	19.33	
H	0.50	0.73	0.83	0.88	0.92	0.94	0.95	0.96	0.97	0.98	0.99	0.99	1.00

- ➢The rainfall stations were ranked as 7, 10, 1, 3, 11, 4, 9, 5, 6, 2, and 8.
- ➢After 8 iterations, the threshold value reaches 95.8%.
- ➢The rainfall information of 8 stations can reflect 95.8% of the rainfall information of the watershed.

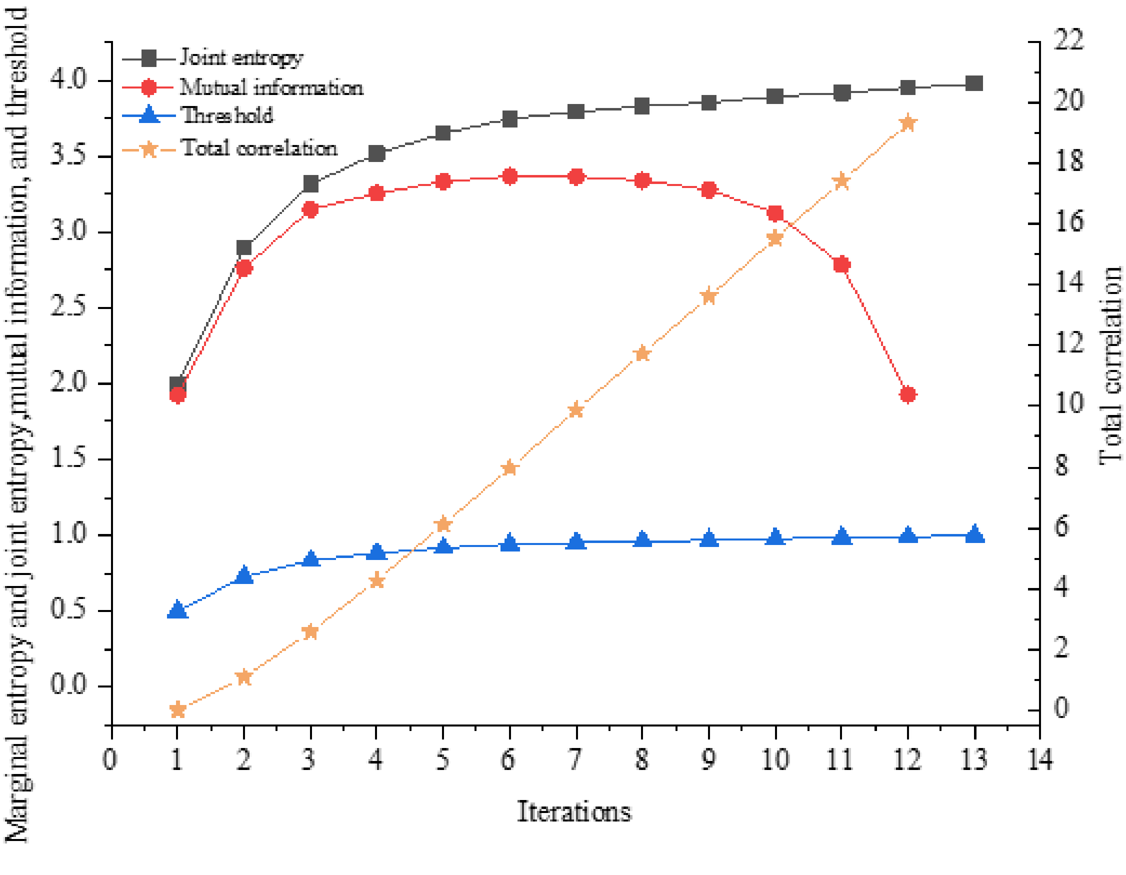


Fig. The variation of marginal entropy and joint entropy (H), mutual information (T), total correlation (C), and threshold (η) with the number of iterations.

3 Result

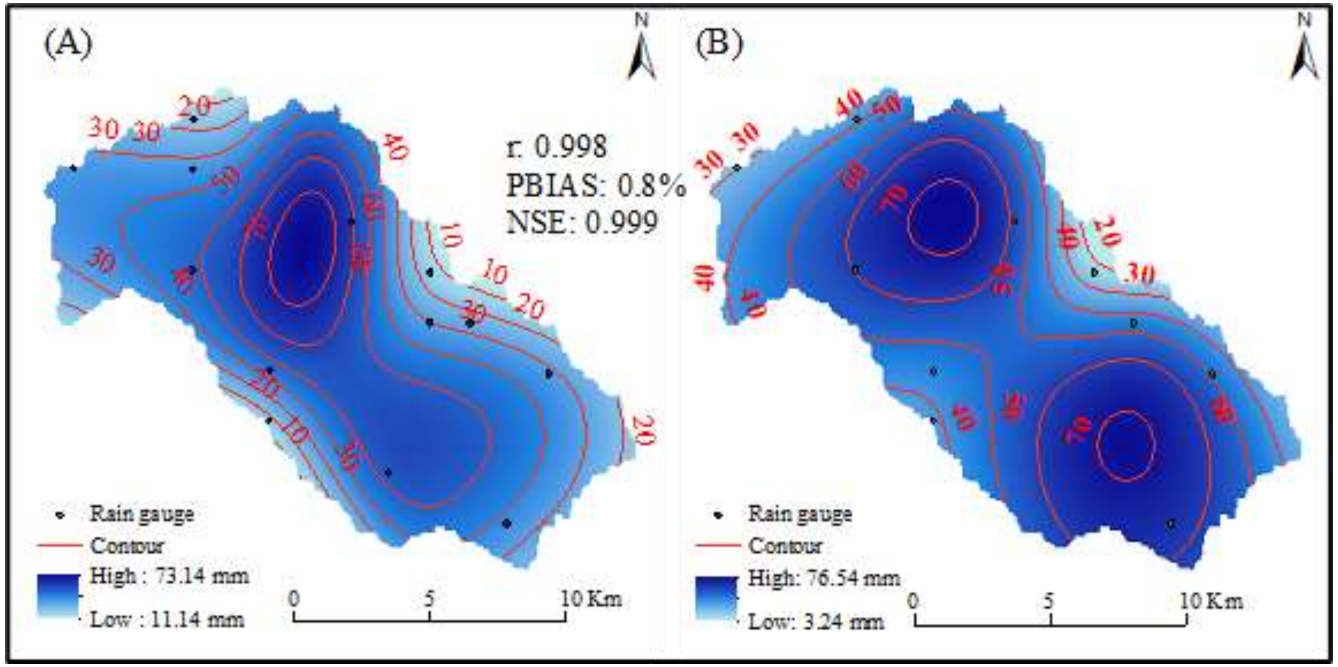


Fig. The distribution of rainfall in the CW (A) before optimization and (B) after optimization. The r, PBIAS, and NSE of them were 0.998, 0.8%, and 0.999 respectively.

- A strong spatially correlation between the rainfall distribution before and after the rainfall network was optimized.
- The precipitation after the rainfall network optimization was hardly overestimated and the accuracy of the data was high.
- The density of the new rain station network was 20.4 km^2.
- the rainstorm center can be caught by the optimized rainfall network
- The rainfall isoline of the optimized rainfall network was similar to that of the original rainfall network.

Instytut Geofizyki Polskiej Akademii Nauk

3 Result

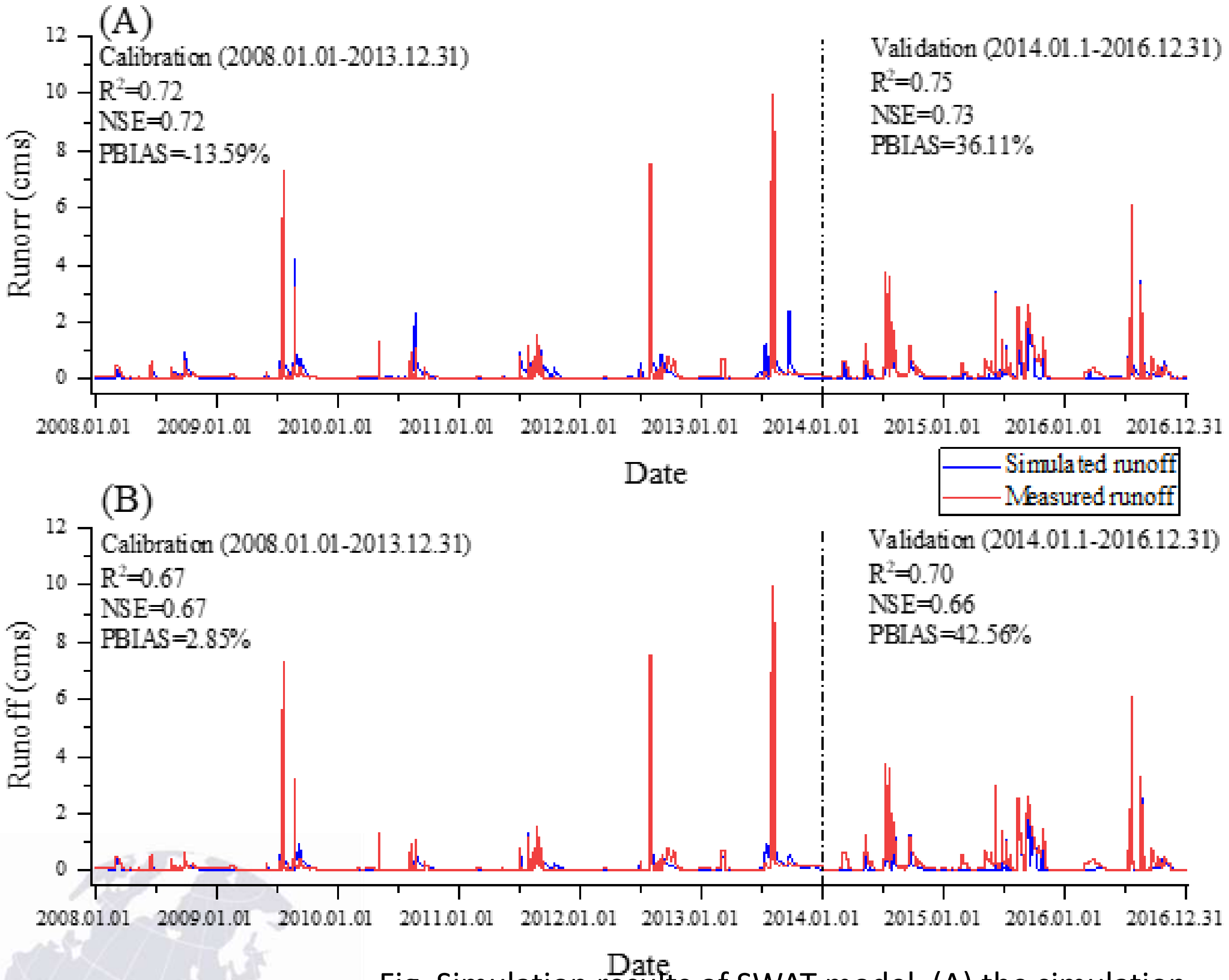


Fig. Simulation results of SWAT model. (A) the simulation results of the original gauge network; (B) the simulation results of the optimized gauge network. Noting that the calibration period was 2008.01.01 – 2013.12.31 and the validation period was 2014.01.01 -2016.12.31. And the data was given by year.month.day.

- the NSE and R^2 were higher than 0.6.
- The models driven by different gauge networks perform well both in the calibration and validation period.
- The uncertainty in simulating runoff reduced with the number of the gauge increased.

Instytut Geofizyki Polskiej Akademii Nauk

3 Result

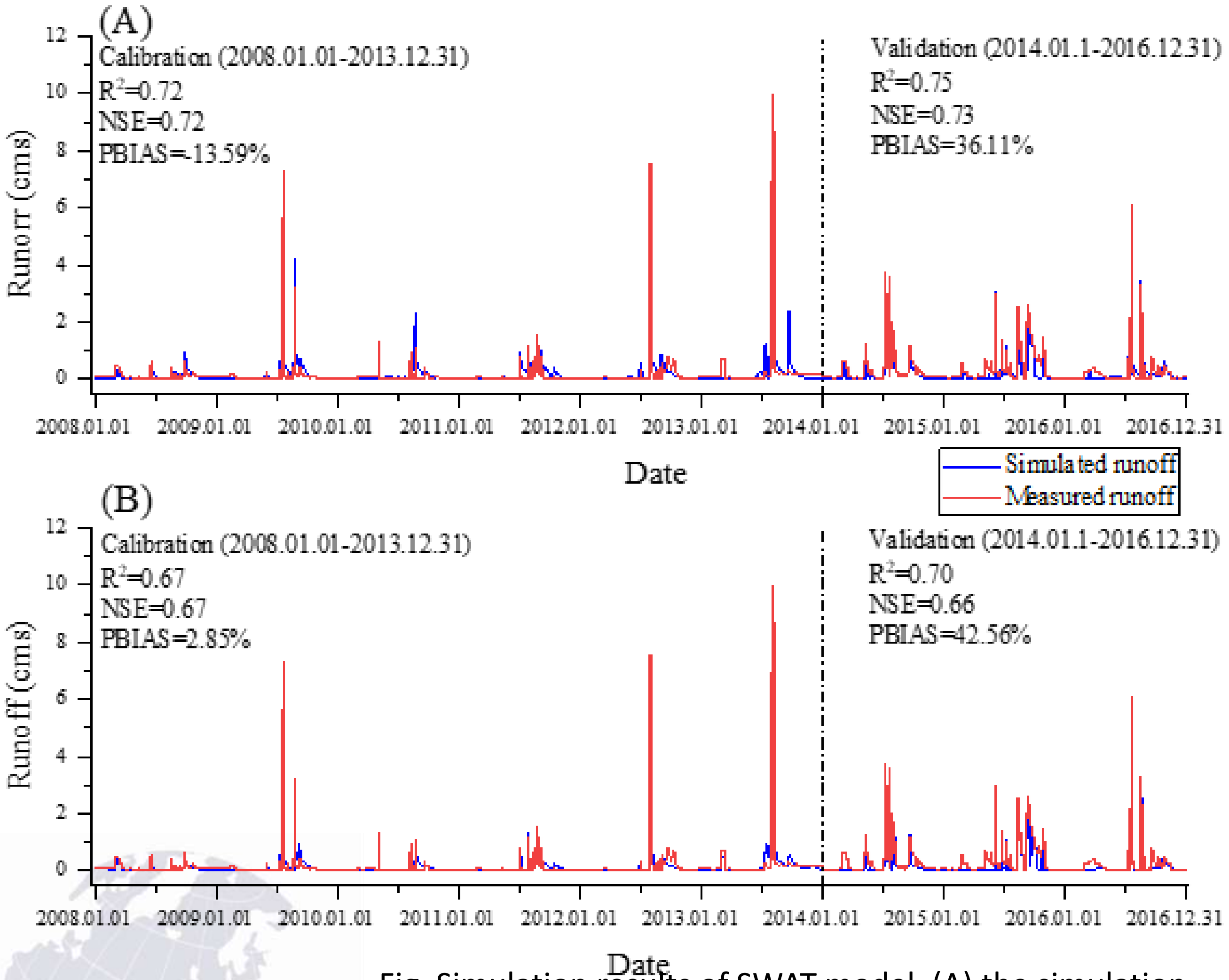


Fig. Simulation results of SWAT model. (A) the simulation results of the original gauge network; (B) the simulation results of the optimized gauge network. Noting that the calibration period was 2008.01.01 – 2013.12.31 and the validation period was 2014.01.01 -2016.12.31. And the data was given by year.month.day.

In the calibration period:

- Optimized-Model performed well
- Both the model developed on the Optimized-Model and the Original-Model performed well.
- Some of the floods were not simulated by both Optimized-Model and the Original-Model, especially the extreme floods.
- Original-Model slightly underestimated the runoff while Optimized-Model slightly overestimated the streamflow.

Instytut Geofizyki Polskiej Akademii Nauk

3 Result

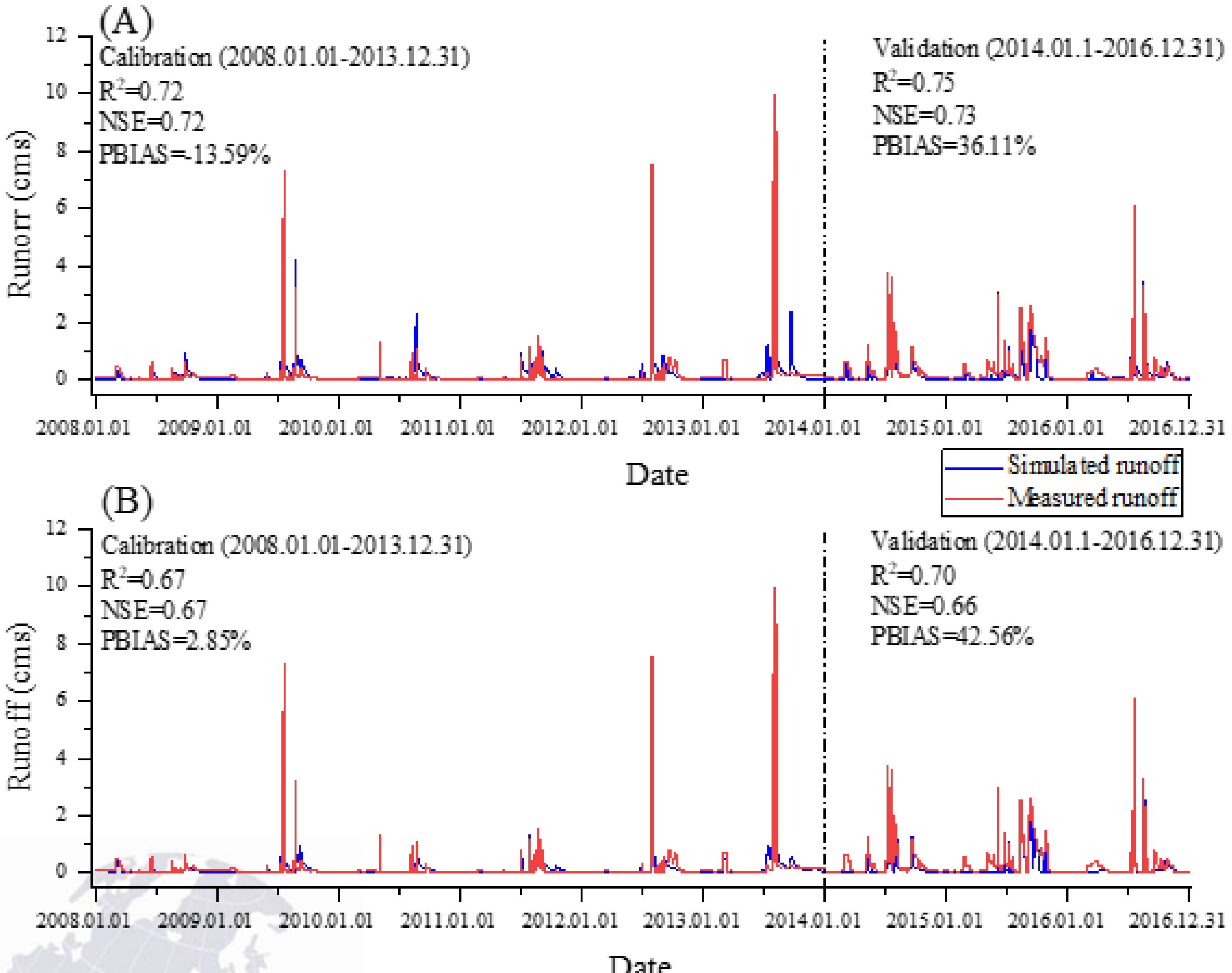


Fig. Simulation results of SWAT model. (A) the simulation results of the original gauge network; (B) the simulation results of the optimized gauge network. Noting that the calibration period was 2008.01.01 – 2013.12.31 and the validation period was 2014.01.01 -2016.12.31. And the data was given by year.month.day.

Instytut Geofizyki Polskiej Akademii Nauk

In the validation period:

- The ability of Original-Model and Optimized-Model to catch floods became better.
- Their R^2 were both higher than 0.70
- Both models achieved good performance.
- Both Original and Optimized-Model seriously overestimated the stream-flow.
- The performance of the Optimized-Model was poorer than that of the Original-Model
- The ability of the Optimized-Model to simulate extreme floods was worse than that of the Original-Model.

3 Result

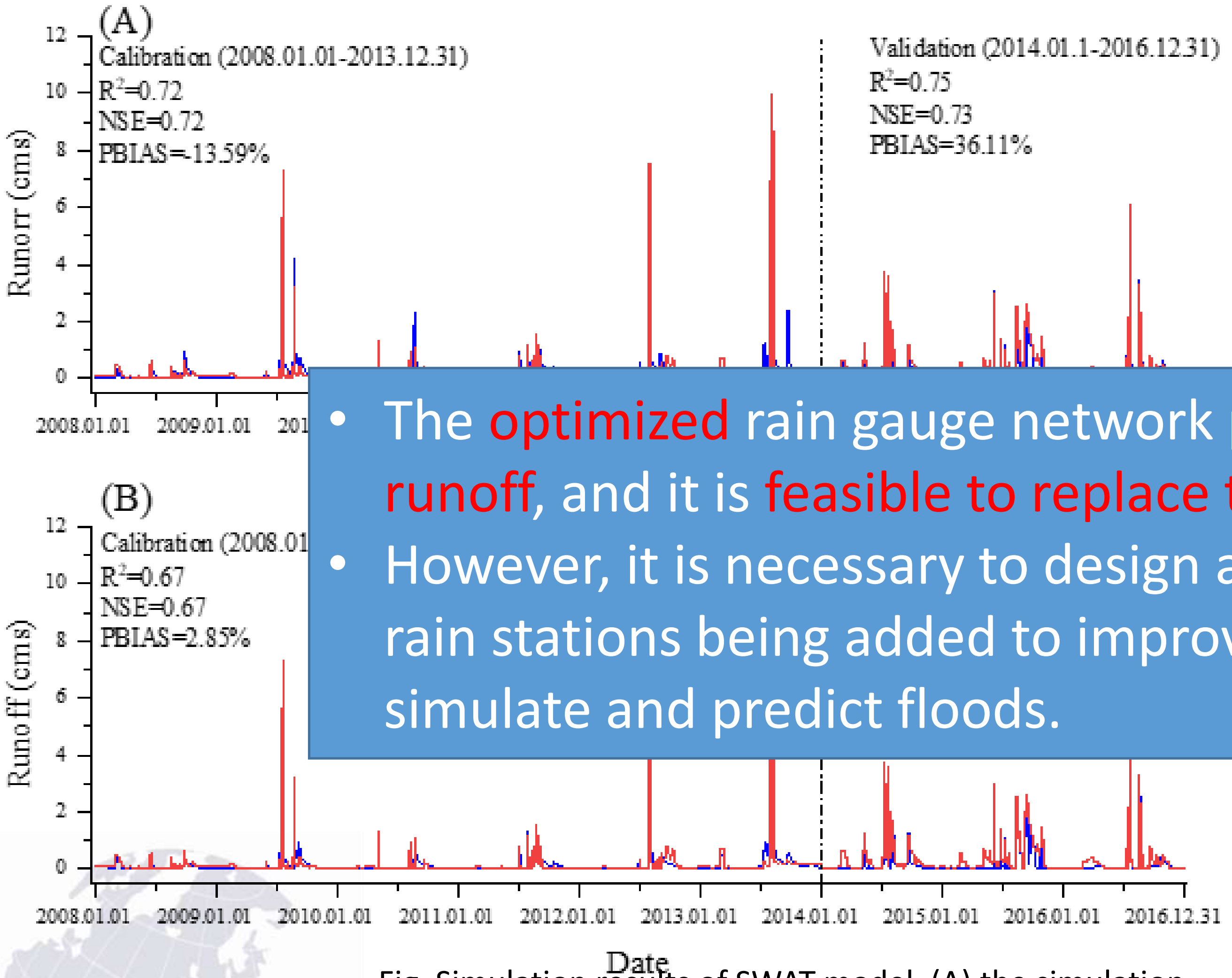


Fig. Simulation results of SWAT model. (A) the simulation results of the original gauge network; (B) the simulation results of the optimized gauge network. Noting that the calibration period was 2008.01.01 – 2013.12.31 and the validation period was 2014.01.01 -2016.12.31. And the data was given by year.month.day.

- The optimized rain gauge network performed well in simulating runoff, and it is feasible to replace the original gauge network.
- However, it is necessary to design a new gauge network with more rain stations being added to improve the model's ability to simulate and predict floods.

In the validation period:

- The ability of Original-Model and Optimized-Model to catch floods became better.
- Their R^2 were both higher than 0.70
- Both models achieved good ...
- ...d-Model ... e stream-
- ...timized-Model was poorer than that of the Original-Model
- The ability of the Optimized-Model to simulate extreme floods was worse than that of the Original-Model.

Instytut Geofizyki Polskiej Akademii Nauk

04

Conclusion

Instytut Geofizyki
Polskiej Akademii Nauk

4 Conclusion

- The distribution of precipitation and entropy exhibits the same trend, higher in the south-west and lower in the north. The heavier the rain, the more information the station contains.
- Use MIMR to optimize the rain gauge network, and 10 stations were selected according to the threshold of the joint entropy.
- The number of iterations increased, the joint entropy trended to become stable.
- The optimized rain gauge network can provide 98.5% rainfall information of the.
- In addition, according to the runoff simulation results, the optimized gauge network achieved good performance in simulating runoff and it can be used in the CW to replace the original one.

Instytut Geofizyki
Polskiej Akademii Nauk

Thank you

Instytut Geofizyki
Polskiej Akademii Nauk