2018

Influence of rainfall data scarcity on non-point source pollution prediction

Reporter: Cheng Sun
Superviser: Zhenyao Shen, Lei Chen
Institution: Beijing Normal University, China

CONTENT

Background

Background

Non-point source (NPS) pollution has been a key threat to water quality

Dispersiveness and Stealthiness

Randomness and Uncertainty

Universality and Undetectability

Soil and Water Assessment Tool (SWAT) models are the main tool used to quantify NPS pollution

Background

Meteorological data

DEM

Land use data

Soil type data

Act as the driving force of runoff generation and pollutant transportation

Rainfall station

Rador product

Data scarcity

Time series

Spatial distribution

Background

Typical data scarcity are divided into three categories

Missing completely at random | Missing at random | Missing not at random

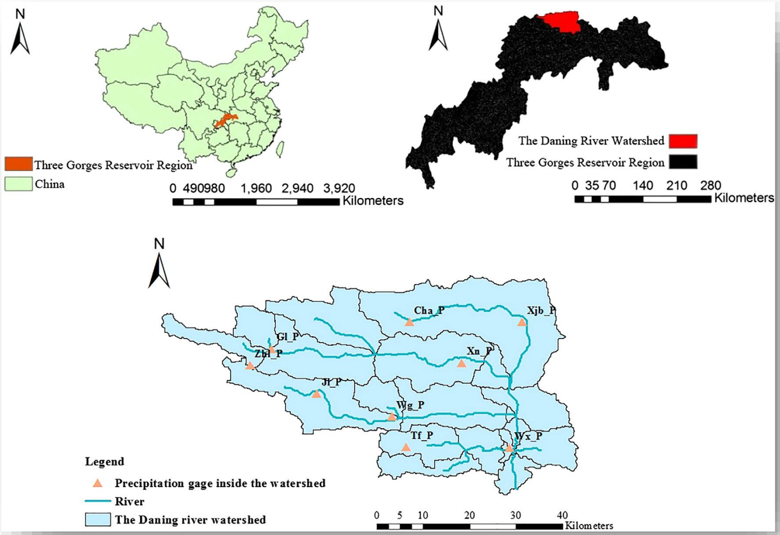
Single imputation

Multiple imputation

- Expectation maximization with bootstrap (EMB) algorithm
- Data augmentation (DA) algorithm

An effective method to address scarce data

Background



- Daning River watershed is a significant tributary of the Three Gorges Reservoir area and is located in Wushan and Wuxi Counties in the municipality of Chongqing, China.
- It suffers from severe NPS pollution, and phosphorus is the limiting nutrient causing eutrophication.

Data types	Resolution	Acquisition path
Digital Elevation Model (DEM)	1:250000	National Fundamental Geographic Information Center of China.
land use map	1:100000	Resources and Environment science data Center of the Chinese Sciences Academy
Soil type map	1:1000000	Agricultural Science Committee of Wuxi city
Hydrologic data	Daily and monthly	Meteorological Bureau of Wuxi County and China National Meteorological Administration
Meteorological data	Daily and monthly	Meteorological Bureau of Wuxi County

Method

Methods

Database establishment

- DEM
- Land use data
- Soil type data
- Meteorological data

Setup of SWAT model

- Correlation coefficient R^2
- Nash-Sutcliffe coefficient (Ens)
- Information entropy method

Calibration and verification

- TP
- Flow

Analysis of scarce scenario

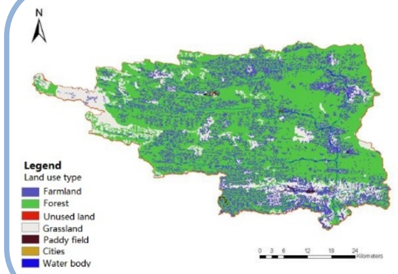
- Temporal data scarcity
- Spatial data scarcity

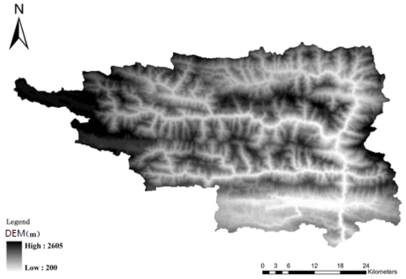
Multiple imputation

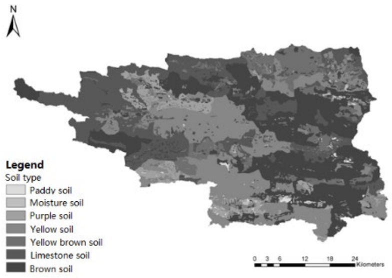
- Data augmentation (DA) algorithm
- Expectation maximization with bootstrap (EMB) algorithms

Methods

Model description







\+ Meteorological data for each site

simulate six P pools in the soil

$$P_{surf} = 0.001 \times C_{orgP} \cdot \frac{Q_{sed}}{A_{hru}} \cdot \varepsilon_{P:sed}$$

$$Q_{sed} = 11.8 (Q_{surf} \cdot q_{peak} \cdot A_{hru})^{0.56} \cdot K_{ulse} \cdot C_{ulse} \cdot P_{ulse} \cdot L_{ulse} \cdot F_{CFRG}$$

Markov Chain - skewed distribution model to simulate daily rainfall data

$$R_{day} = \mu_{mon} + 2\sigma_{mon} \frac{\left[\left(SND_{day} - \frac{g_{mon}}{6} \right) \left(\frac{g_{mon}}{6} \right) + 1 \right]^3 - 1}{g_{mon}}$$

Methods

Calibration and verification

Sequential Uncertainty Fitting Version-2 (SUFI-2)

Monthly time step, one-year warm-up period

Calibration period:2004-2008
Validation period: 2000-2003

Verification point: 13 subbasin outlet

ITEMS	NUM	PARAMETER	INFERIOR LIMIT	UPPER LIMIT
FLOW	1	Sol_Awc	0	1
	2	Sol_K	-20	30000
	3	Esco	0	1
	4	Gwqmn	0	5000
	5	Cn2	-25	19
	6	Canmx	0	10
	7	Sol_Z	-25	25
	8	Blai	0	1
	9	Ch_K2	0	150
	10	Surlag	0	10
	11	Gw_Delay	1	45
	12	Ch_N2	0	1
	13	Epco	0	1
	14	Revapmn	0	500
	15	Biomix	0	1
SEDIMENT	1	Spcon	0.0001	0.05
	2	Ch_Cov	0	1
	3	Ch_Erod	0	1
	4	Usle_P	0	1
	5	Spexp	1	1.5
TP	1	Sol_Orgp	0	400
	2	Pperco	10	18
	3	Phoskd	100	200
	4	Rchrg_Dp	0	1
	5	Sol_Labp	-25	25

Methods

Evaluation indicators

① The correlation coefficient R^2:

$$R^2=\left[\left(\sum_{i=1}^{n}(O_i-\bar{O})(P_i-\bar{P})\right)\Big/\left(\sqrt{\sum_{i=1}^{n}(O_i-\bar{O})}\sqrt{\sum_{i=1}^{n}(P_i-\bar{P})^2}\right)\right]^2$$

② The Nash-Sutcliffe coefficient (Ens):

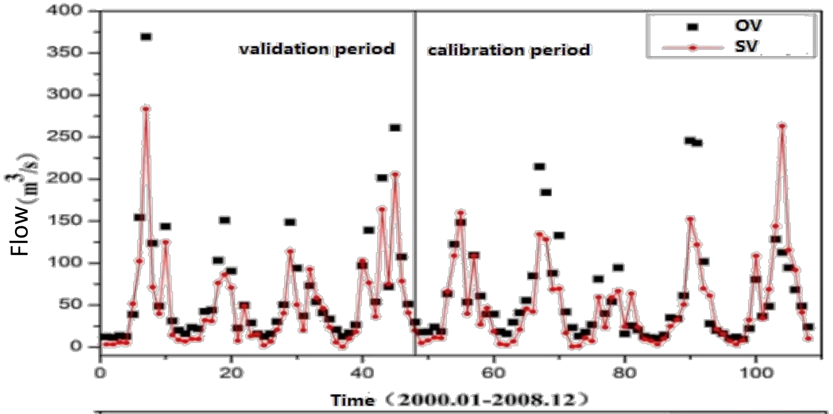
$$Ens=1-\sum_{i=1}^{n}(O_i-P_i)^2/\sum_{i=1}^{n}(O_i-\bar{O})^2$$

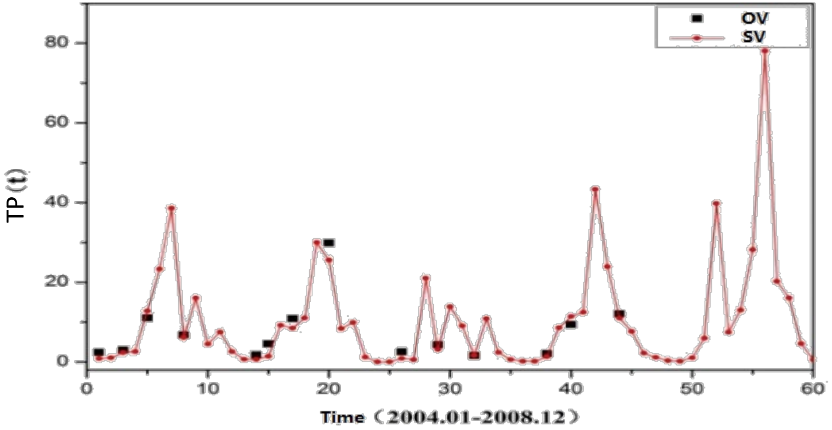
③ Information entropy method:

$$H_{spatial}(x_j)=-\sum_{i=1}^{n}P(x_{ij})logP(x_{ij})$$

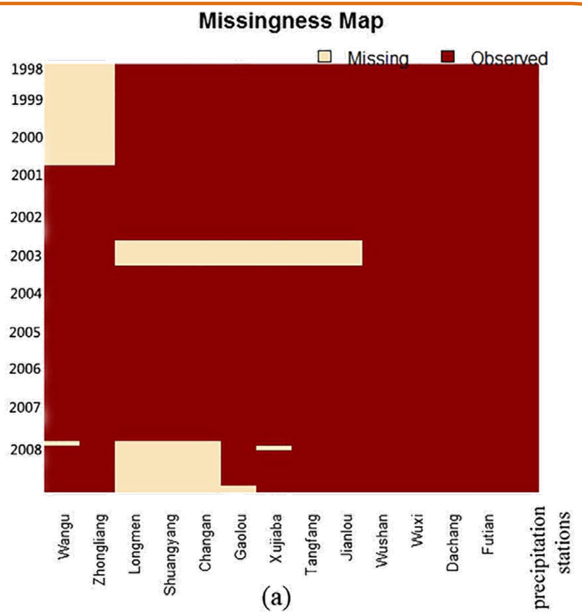
$$H_{temporal}(x_i)=-\sum_{j=1}^{m}P(x_{ij})logP(x_{ij})$$

Simulation period	Data type	R^2	E_{NS}
2000.01-2008.1	Flow	0.79	0.74
	TP	0.95	0.93





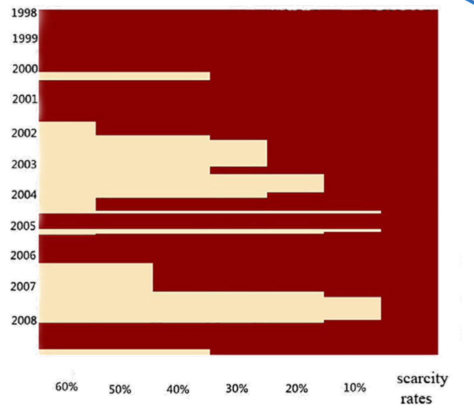
Methods Temporal data scarcity (in the Xining station)



(a)

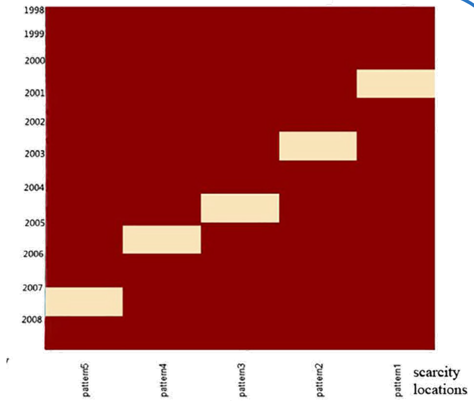
Rainfall data is available at random times, the most parts are discontinuous, and some short time series are missing.

missing rates



(b)

missing positions



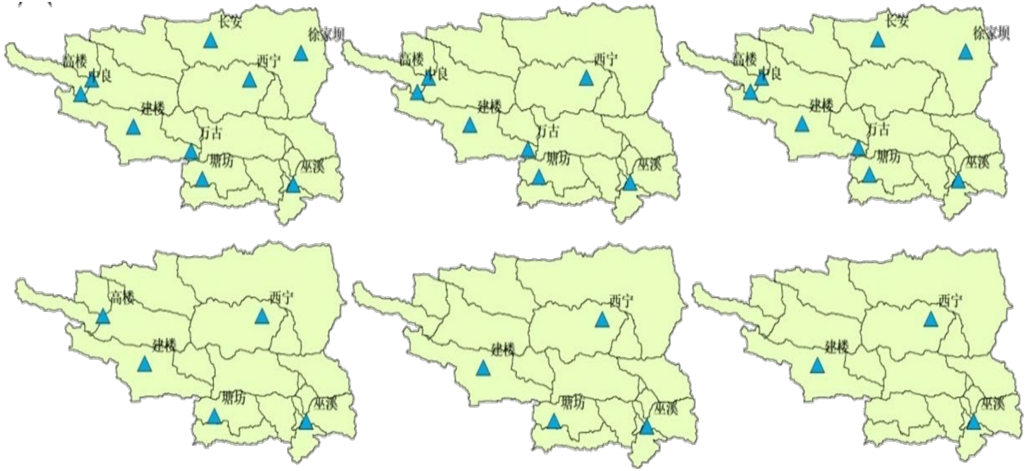
(c)

Methods

Spatial data scarcity

The location of a removed rainfall stations

Precipitation station	Information entropy
Gaolou	0.2002
Zhongliang	0.0906
Jianlou	0.2756
Xining	0.3638
Wuxi	0.2410
Wangu	0.0757
Tagnfang	0.1656
Xujiaba	0.3111
Changan	0.1909



Number of gauges	Scarce gauges	Information etropy
9 gauges	-	1.9145
8 gauges	Xining station	1.5507
7 gauges	Changan and Xujiaba stations	1.4125
5 gauges	Changan, Xujiaba, Wangu and Zhongliang stations	1.2462
4 gauges	Changan, Xujiaba, Wangu, Zhongliang and Gaolou stations	1.0460
3 gauges	Changan, Xujiaba, Wangu, Zhongliang, Gaolou and Tangfang stations	0.8804

Methods Design of rainfall data scarcity

The experimental design of rainfall data scarcity scenario.

Category		Name	Detailed description
Baseline		S0	No scarce data scenarios (no lack of time series and also retained a complete nine stations)
Spatial data scarcity	Decreasing number of rainfall stations	S1	An 8-gauge scenario that lacked the Xining station
		S2	A 7-gauge scenario that lacked the Changan and Xujiaba stations
		S3	A 5-gauge scenario that lacked the Changan, Xujiaba, Wangu and Zhongliang stations
		S4	A 4-gauge scenario that lacked the Changan, Xujiaba, Wangu, Zhongliang and Gaolou stations
		S5	A 3-gauge scenario that lacked the Changan, Xujiaba, Wangu, Zhongliang, Gaolou and Tangfang stations
	Effect of the location of a removed rainfall stations	S6	A scenario with 6 stations that lacked the Gaolou station
		S7	A scenario with 6 stations that lacked the Jianlou station
		S8	A scenario with 6 stations that lacked the Tangfang station
		S9	A scenario with 6 stations that lacked the Wangu station
		S10	A scenario with 6 stations that lacked the Wuxi station
		S11	A scenario with 6 stations that lacked the Xining station
		S12	A scenario with 6 stations that lacked the Zhongliang station
Temporal data scarcity (in the Xining station)	Rainfall time series degradation with increasing missing period	S13	A scenario with 10% data scarcity
		S14	A scenario with 20% data scarcity
		S15	A scenario with 30% data scarcity
		S16	A scenario with 40% data scarcity
		S17	A scenario with 50% data scarcity
		S18	A scenario with 60% data scarcity
	Rainfall time series degradation with variable timing of the missing period	S19	Pattern 1 with data scarcity in the high flow year of 2000
		S20	Pattern 2 with data scarcity in the normal flow year of 2002
		S21	Pattern 3 with data scarcity in the low flow year of 2004
		S22	Pattern 4 with data scarcity in the high flow year of 2005
		S23	Pattern 5 with data scarcity in the high flow year of 2007

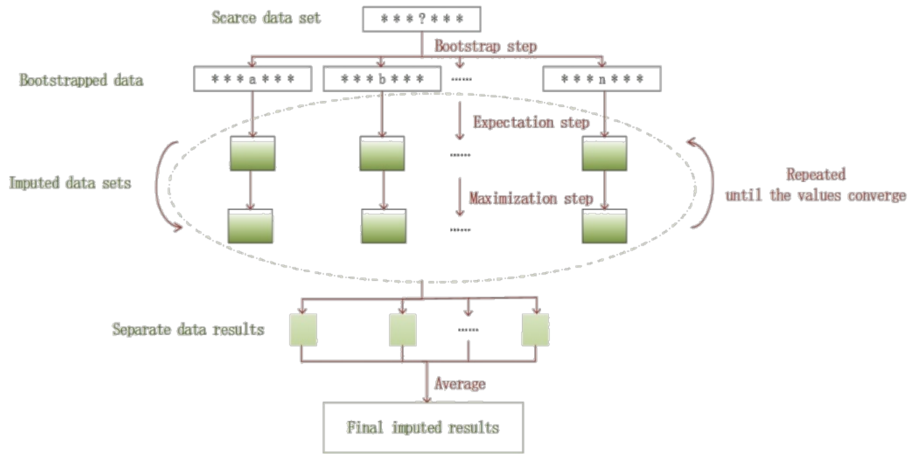
Methods **Multiple imputation (DA)**

The DA algorithm is an iterative optimization and sampling algorithm method that introduces latent variables.

First step (i.e. I-step) $Y_{miss}^{(t+1)} \sim f(Y_{miss} \mid Y_{obs}, \theta^{(t)})$

Second step (i.e. P-step) $\theta^{(t+1)} \sim f(\theta \mid Y_{obs}, Y_{miss}^{(t+1)})$

Methods Multiple interpolation (EMB)



The EMB algorithm is a combination of the EM algorithm and the bootstrap method.

Two basic assumptions should be satisfied:
(1) Obey the multivariate normal distribution $Y \sim N_p(\mu, \Sigma)$;
(2) The response mechanism of scarce data should be MAR, that is, $Y = (Y_{obs}, Y_{miss})$.

Such that $\theta = (\mu, \Sigma)$

Decomposition of likelihood functions

$$f(Y_{obs}, M | \theta, \phi) = f(Y_{obs} | \theta) f(M | Y_{obs}, \phi)$$

Uninformative prior distribution

$$f(\theta | Y_{obs}) \propto f(Y_{obs} | \theta) = \int f(Y | \theta) dY_{miss}$$

Results

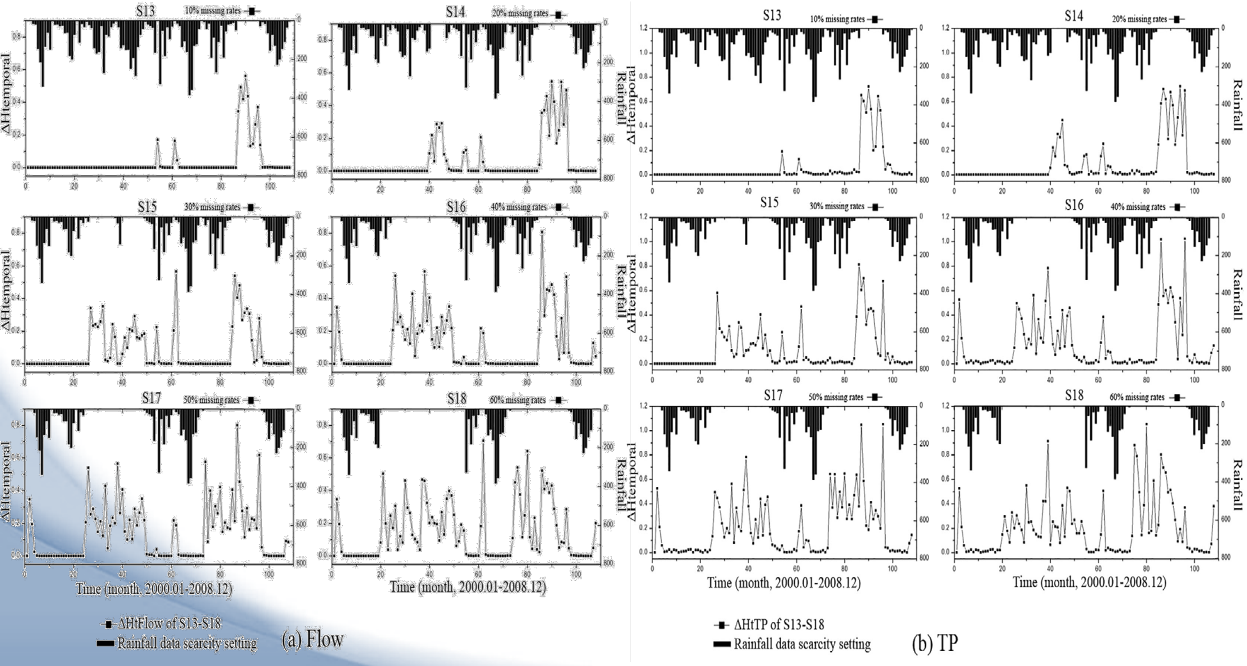
Results

Impacts of temporal data scarcity (missing rates)

Evaluation of the simulation results for the different temporal data scarcities.

		No missing	Different missing rates					
			10%	20%	30%	40%	50%	60%
Flow	Ens	0.7425	0.6054	0.5457	0.5677	0.5366	0.5693	0.5494
	R^2	0.786	0.662	0.611	0.649	0.612	0.641	0.633
	ΔH		3.7115	6.0821	8.2601	11.7208	14.3642	15.508
TP	Ens	0.9299	0.8655	0.8711	0.8242	0.8615	0.8613	0.7515
	R^2	0.952	0.933	0.932	0.905	0.928	0.914	0.794
	ΔH		5.9542	9.2635	11.5274	16.8487	20.6244	21.3764

The corresponding threshold



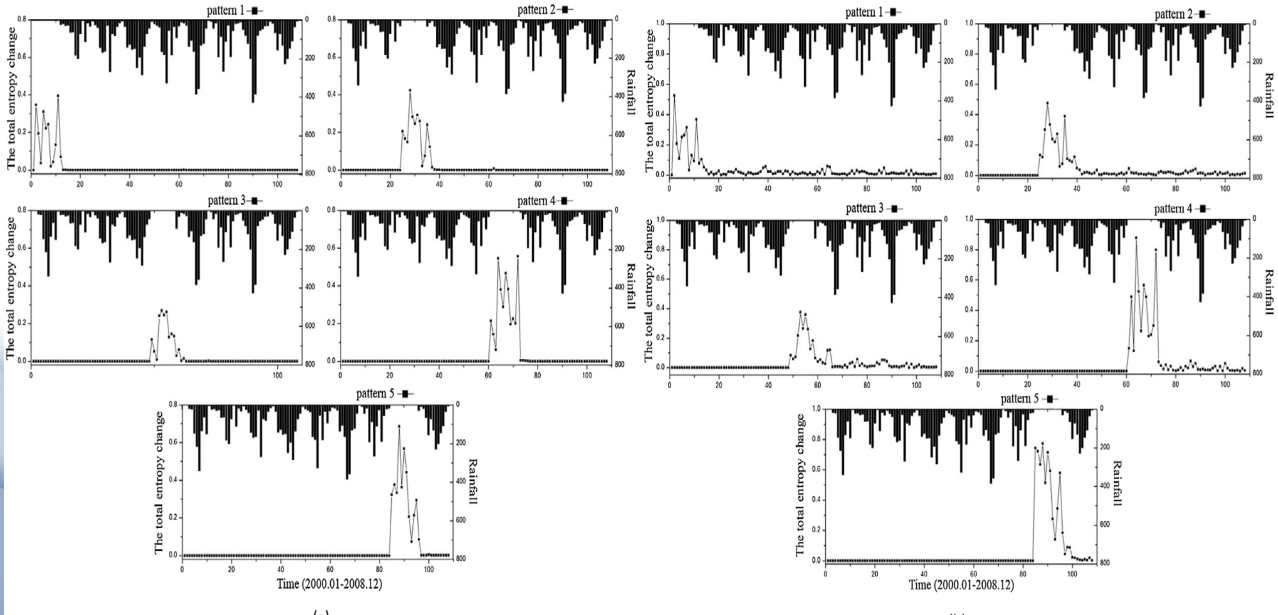
- ΔHtemporal increased with increasing missing rates;
- At the same missing rate, the ΔHtemporal for flow and TP simulations were consistent;
- The ΔHtemporal in TP was larger than the flow.

Results

Impacts of temporal data scarcity (scarcity locations)

		No missiong	Different data location scarcity				
			Pattern 1	Pattern 2	Pattern 3	Pattern 4	Pattern 5
Flow	Ens	0.7425	0.6871	0.7287	0.7342	0.6573	0.6103
	R	0.7860	0.7410	0.7780	0.7870	0.7140	0.6670
	△H		2.0351	2.5275	1.7184	3.6907	3.9754
TP	Ens	0.9299	0.9329	00.9137	0.9289	0.4339	0.8514
	R	0.9520	0.9540	0.9340	0.9610	0.5510	0.9250
	△H		4.0201	4.0038	3.0790	5.7503	6.5524

These corresponding years are 2000, 2005, and 2007, which were high flow years.



(a)

(b)

- The $\Delta H_{temporal}$ increased if data in high flow years were missing, and the model performance become poor;
- The $\Delta H_{temporal}$ for TP was obviously greater than the $\Delta H_{temporal}$ for flow.

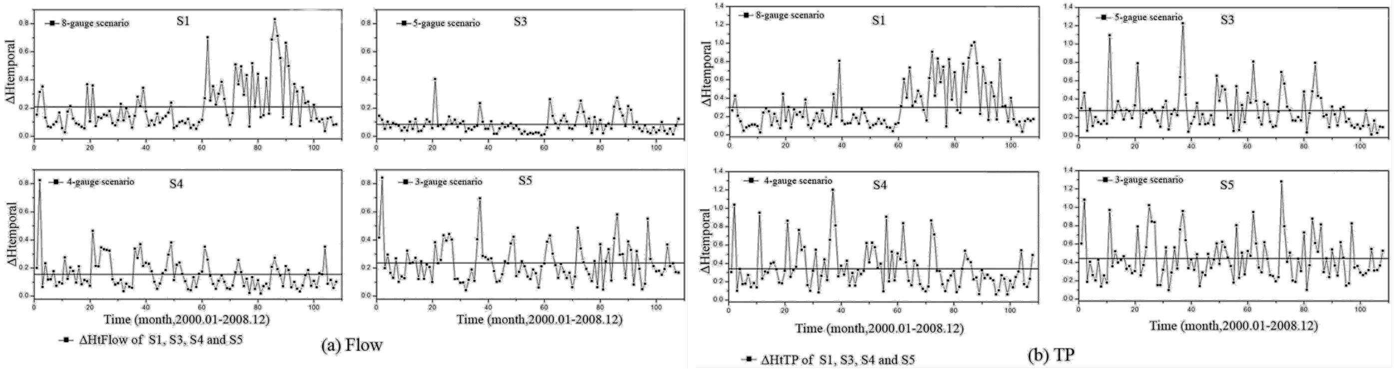
Results

Impacts of spatial data scarcity (station number and location)

Evaluation of the simulation effects for the different combinations of rainfall stations.

	Evaluation indicators	9-gauge	8-gauge	5-gauge	4-gauge	3-gauge
Flow	ENS	0.7425	0.5729	0.7308	0.6273	0.6023
	R^2	0.786	0.638	0.771	0.676	0.66
	ΔH		22.5170	9.1101	16.9243	25.4811
Total phosphorus	ENS	0.9299	0.6928	0.8299	0.7415	0.742
	R^2	0.952	0.867	0.947	0.943	0.94
	ΔH		32.4108	29.6263	37.0232	47.7475

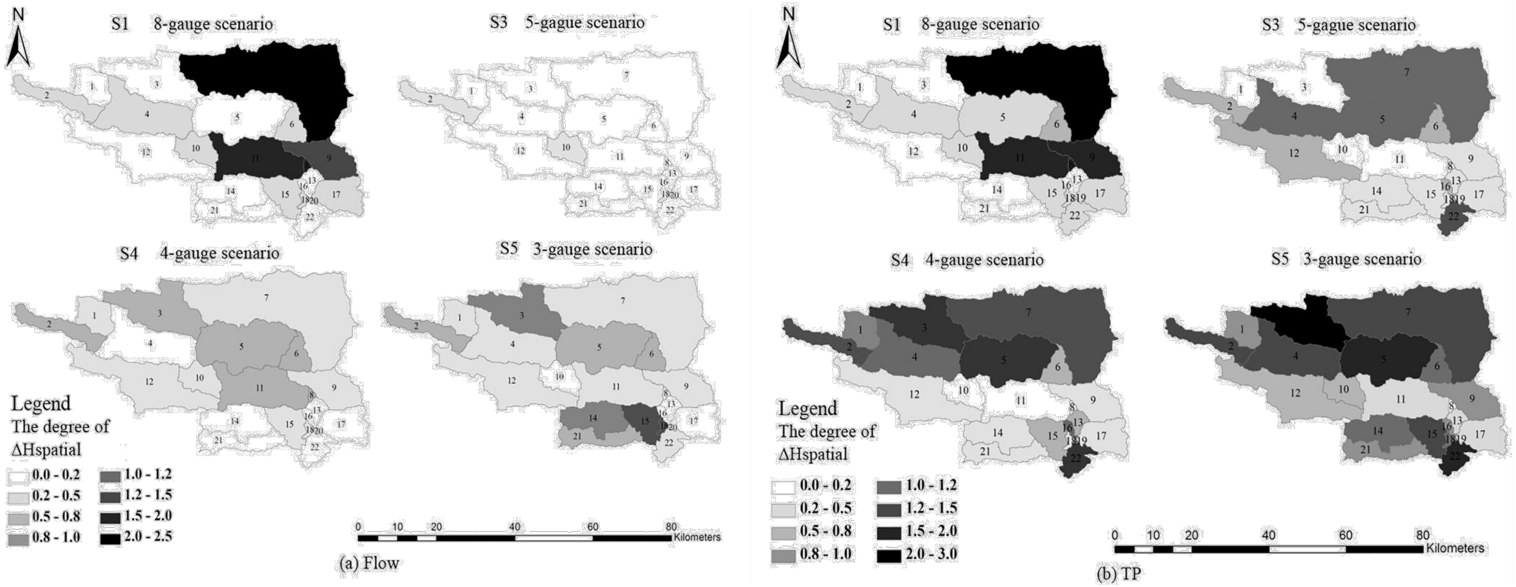
- lacked the main site reduced the total set of information by approximately 25%;
- The fewer gauges, the worse simulation results.



- For missing gauges with high information entropy, the simulated TP outputs changed dramatically .

Results

Impacts of spatial data scarcity (station number and location)



- The $\Delta H_{spatial}$ for the 7th sub-watershed and its other downstream sub-watersheds such as 9th sub-watershed increased during S1 scenario (scarcity of the Xining station).

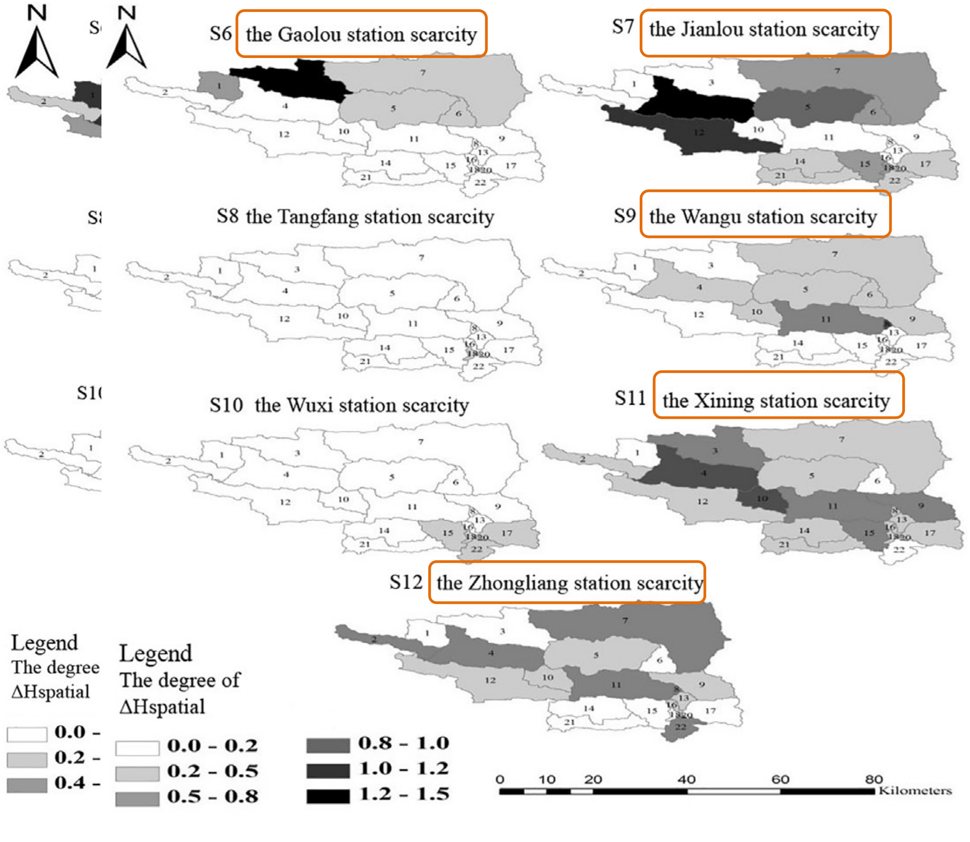
Results

Impacts of spatial data scarcity (downstream simulations)

Seven gauges with complete data

Seven scarce scenarios of gauge

The impact of single gauge



(b) TP

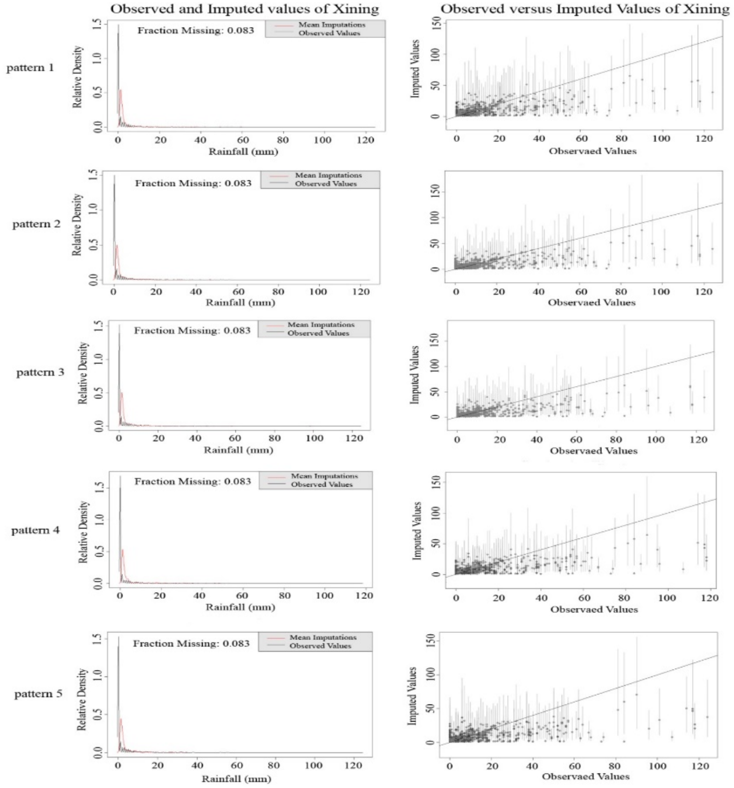
Results

Impacts of different imputation methods on rainfall data series

The imputation effects of rainfall data at six missing rates

Missing rates	DA	EMB
None missing	106.50	106.50
10%	72.19	97.28
20%	65.33	94.96
30%	61.51	93.65
40%	63.43	86.31
50%	58.10	87.92
60%	58.31	90.67

- The distribution of imputations is consistent with the distribution of observations;
- When the relative density of observations exceeded 1.5, the relative density of the imputations could be below 0.5;
- The 90% confidence interval of the imputed values is able to cover this theoretical line at different missing rates.



Pattern 1: high-flow year
Pattern 2: normal-flow year
Pattern 3: low-flow year
Pattern 4: high-flow year
Pattern 5: high-flow year

Results

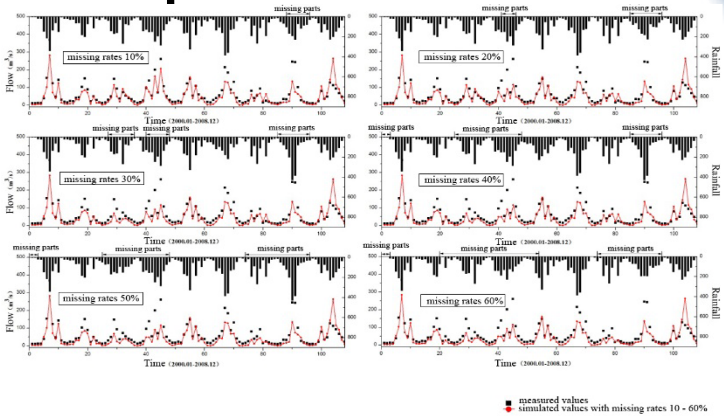
Missing rates

Impacts of different imputation data sets on NPS pollution simulations

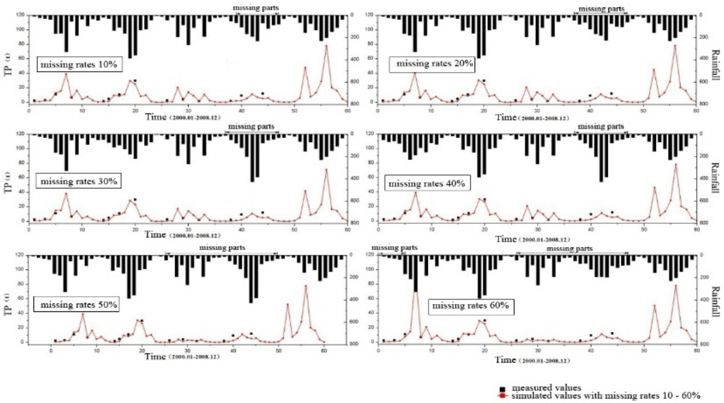
Differences in simulation results before and after imputation at six missing rates

		Evaluation indicators	None missing	10%	20%	30%	40%	50%	60%
Before imputation	Flow	NSE	0.74	0.61	0.55	0.57	0.54	0.57	0.55
		R^2	0.79	0.66	0.61	0.65	0.61	0.64	0.63
	TP	NSE	0.93	0.87	0.87	0.82	0.86	0.86	0.75
		R^2	0.95	0.93	0.93	0.90	0.93	0.91	0.79
After imputation	Flow	NSE	0.74	0.69	0.61	0.58	0.58	0.57	0.56
		R^2	0.79	0.75	0.69	0.68	0.68	0.69	0.69
	TP	NSE	0.93	0.86	0.85	0.85	0.84	0.82	0.80
		R^2	0.95	0.92	0.92	0.89	0.91	0.91	0.93

- The estimated effect of the imputed data set of the flow and the TP loads with different missing rates improved;
- The imputation effect of rainfall data is less affected by the changes in missing rates;



(a)



(b)

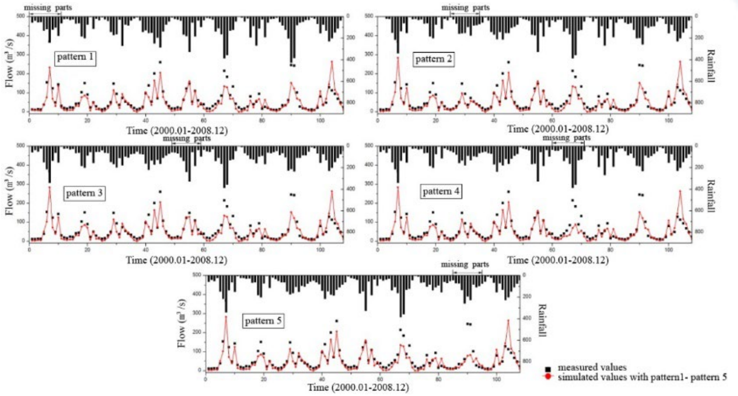
Results

Missing positions

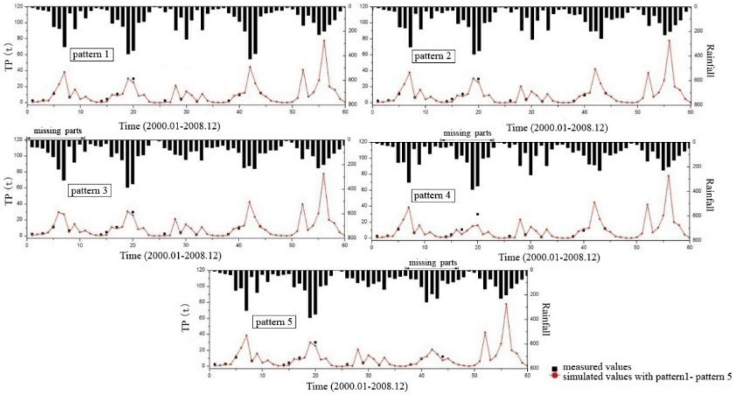
Impacts of different imputation data sets on NPS pollution simulations

Differences in simulation results before and after imputation at five scarcity locations

		Evaluation indicators	Pattern 1	Pattern 2	Pattern 3	Pattern 4	Pattern 5
Before imputation	Flow	NSE	0.69	0.73	0.73	0.66	0.61
		R^2	0.74	0.78	0.79	0.71	0.67
	TP	NSE	0.93	0.91	0.93	0.43	0.85
		R^2	0.95	0.93	0.96	0.55	0.93
After imputation	Flow	NSE	0.71	0.74	0.74	0.69	0.67
		R^2	0.76	0.79	0.79	0.74	0.72
	TP	NSE	0.93	0.93	0.94	0.67	0.91
		R^2	0.95	0.95	0.96	0.77	0.95



(a)



(b)

- Model performance of the imputed values in different missing positions are also better than the simulation results before imputation;
- The simulated values in the normal-flow years and the low-flow years are closer to the baseline values than are those in the high-flow years;

Conclusions

Conclusions

- The results highlighted the importance of critical-site rainfall stations (Xining station in this paper) on the SWAT simulations (for better rainfall spatial distribution);
- Higher missing rates above a certain threshold as well as missing locations during the wet periods resulted in poorer simulation results (for better temporal distribution).
- The repair of rainfall data and the SWAT model performance obtained by the EMB algorithm are superior to the traditional DA algorithm (weather generator).
- It is noted that even if the best algorithm is used, the imputed value is always lower than the peak observed value (multiple sources of rainfall data).

Chen et al., journal of hydrology, 2017; Chen et al., journal of hydrology, 2018;

END

THANK YOU FOR YOUR ATTENTION!