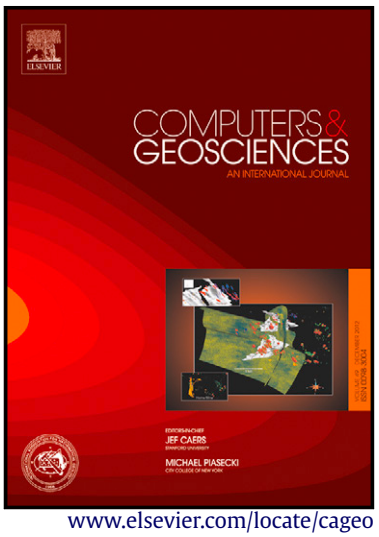
Author's Accepted Manuscript



www.elsevier.com/locate/cageo

Estimation of theoretical maximum speedup ratio for parallel computing of grid-based distributed hydrological models

Junzhi Liu, A-Xing Zhu, Cheng-Zhi Qin

PII: S0098-3004(13)00156-8
DOI: http://dx.doi.org/10.1016/j.cageo.2013.04.030
Reference: CAGEO3194

To appear in: *Computers & Geosciences*

Received date: 4 March 2013
Revised date: 26 April 2013
Accepted date: 30 April 2013

Cite this article as: Junzhi Liu, A-Xing Zhu, Cheng-Zhi Qin, Estimation of theoretical maximum speedup ratio for parallel computing of grid-based distributed hydrological models, *Computers & Geosciences,* http://dx.doi.org/10.1016/j.cageo.2013.04.030

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting galley proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

ACCEPTED MANUSCRIPT

Estimation of theoretical maximum speedup ratio for parallel computing of grid-based distributed hydrological models

Junzhi Liu[a,b] , A-Xing Zhu[a, c*] , Cheng-Zhi Qin[a]

[a] State Key Lab of Resources and Environmental Information System, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing, 100101, China.

[b] University of Chinese Academy of Sciences, Beijing, 100049, China

[c] Department of Geography, University of Wisconsin-Madison, Madison, WI 53706, USA

*Corresponding author. Email: axing@lreis.ac.cn; Fax: 86-10-64889630; Telephone: 86-13810787680; Adrress: 11A, Datun Road, Chaoyang district, Beijing, 100101, China.

Abstract: Theoretical Maximum Speedup Ratio (TMSR) can be used as a goal for improving parallel computing methods for distributed hydrological models. Different types of distributed hydrological models need different TMSR estimation methods because of the different computing characteristics of models. Existing TMSR estimation methods, such as those for sub-basin based distributed hydrological models, are inappropriate for grid-based distributed hydrological models. In this paper, we proposed a TMSR estimation method suitable for grid-based distributed hydrological models. With this method, TMSRs for hillslope processes and channel routing processes are calculated separately and then combined to obtain the overall TMSR. A

Accepted manuscript

ACCEPTED MANUSCRIPT

branch-and-bound algorithm and a critical path heuristic algorithm are used to estimate TMSRs for parallel computing of hillslope processes and channel routing processes, respectively. The overall TMSR is calculated according to the proportions of computing these two types of processes. A preliminary application showed that the more the number of sub-basins, the larger the TMSRs and that the compact watersheds had larger TMSRs than the long narrow watersheds.

Keywords: Theoretical maximum speedup ratio; Parallel computing; Grid-based distributed hydrological model; Critical path heuristic algorithm

1. Introduction

Theoretical Maximum Speedup Ratio (TMSR) in parallel computing of distributed hydrological models is defined as the potential maximum speedup ratio value that can be obtained under a given number of processors (Wang et al., 2012). It can be used as a goal for improving parallel computing methods for distributed hydrological models. Different types of distributed hydrological models need different TMSR estimation methods because of the different computing characteristics of models.

Currently, there has been little published work on estimating TMSR for parallel computing of grid-based distributed hydrological models, which are an important type of hydrological model (Borah and Bera, 2004). Existing related researches mainly focus on TMSR estimation for sub-basin based distributed hydrological models

Accepted manuscript

ACCEPTED MANUSCRIPT

(Apostolopoulos and Georgakakos, 1997; Li et al., 2011; Wang et al., 2012). For example, Apostolopoulos and Georgakakos (1997) proposed a method for estimating TMSR by representing sub-basins in a watershed using Directed Acyclic Graph (DAG). Li et al. (2011) and Wang et al. (2012) estimated TMSR for parallel computing of sub-basin based distributed hydrological models by analyzing the basin width function.

These TMSR estimation methods for sub-basin based distributed hydrological models all assume that the amounts of computation for each sub-basin are the same (Wang et al., 2012), which is not true for grid-based distributed hydrological models as the numbers of grid cells in each sub-basin vary widely. So existing TMSR estimation methods for sub-basin based distributed hydrological models are inappropriate for grid-based distributed hydrological models. The objective of this paper is to propose a new TMSR estimation method suitable for grid-based distributed hydrological models considering their computing characteristics.

2. Basic idea

This paper also uses sub-basin as the basic unit for parallel computing because communication among sub-basins is low and sub-basin based parallel computing has been proved to be effective (Vivoni et al., 2011; Wang et al., 2011). According to the computation characteristics processes in sub-basins can be divided into two types: hillslope processes (such as infiltration and evaporation) and channel routing processes (such as channel flow and sediment routing). For hillslope processes calculations are independent among sub-basins while for channel routing processes

Accepted manuscript

ACCEPTED MANUSCRIPT

calculations for one sub-basin depends on the calculation results of its upstream sub-basins. Thus, different strategies should be adopted for parallel computing of these two types of processes. Different TMSR estimation methods should also be used accordingly. Our idea is first to estimate the TMSRs for these two types of processes, respectively and then combine them to obtain an overall TMSR.

3. TMSR estimation method

3.1 Assumptions

The assumptions for the TMSR estimation method in this paper are as follows:

(1) Groundwater follows the direction of overland flow and there is only interaction between upstream and downstream sub-basins (Li et al., 2011).

(2) Communication overhead among sub-basins is low, so it can be ignored for simplicity (Li et al., 2011; Wang et al., 2012).

3.2 Estimation process

Step 1: Estimating TMSR for parallel computing of hillslope processes

For parallel computing of hillslope processes, due to the fact that there are no dependencies among sub-basins, the goal of task scheduling is to assign n independent tasks (one task for each sub-basin) to m processors with the objective of minimizing execution time. The execution time of hillslope processes for a sub-basin can be represented by the number of hillslope cells in this sub-basin because generally the more the number of cells, the more the execution time there will be. Thus, the serial computing time can be represented by the total number of hillslope cells in the watershed.

Accepted manuscript

ACCEPTED MANUSCRIPT

Estimating TMSR in this situation is a classical Multiprocessor Scheduling Problem. In this paper, a branch-and-bound algorithm for exact solution of the problem (Dell'Amico and Martello, 1995) is used to get the minimum execution time of parallelizing hillslope processes under *m* processors. Then TMSR for hillslope processes is calculated by dividing the serial computing time by the minimum parallel computing time.

Step 2: Estimating TMSR for parallel computing of channel routing processes

According to assumption (1) the dependence among sub-basins for calculations of channel routing processes can be represented by a tree-structured DAG (Apostolopoulos and Georgakakos, 1997; Li et al., 2010) (Fig. 1). Each node in the DAG represents the calculation task of channel routing processes for a sub-basin. The goal of task scheduling for channel routing processes is to distribute tasks in the DAG among a given number of processors to achieve minimum execution time. The execution time of channel routing processes for a sub-basin can be represented by the number of channel cells in this sub-basin and the serial computing time can be represented by the total number of channel cells in the watershed.

(Fig. 1 is about here)

The scheduling problem for this case belongs to the static task scheduling problems and the critical path heuristic algorithm has been proved to be effective on solving such problems (Shirazi et al., 1990). This paper uses the critical path heuristic algorithm to obtain the minimum execution time of channel routing processes parallelized on a given number of processors. To illustrate this algorithm, some terms

Accepted manuscript

ACCEPTED MANUSCRIPT

are defined as below.

(1) The "accumulation time" of a processor P_i, $AT(P_i)$, is the total time needed for a processor to finish all the tasks assigned to it.

(2) A node is said to be "mature" if it is ready to be assigned to a processor, that is, this node has no precedent nodes or all its precedent nodes have been already completed.

(3) The "exit path" of a node means the path from this node to the exit node (representing the outlet sub-basin) and the length of exit path is the sum of execution time of every node in this path..

The longest exit path (the critical path, e.g. G-D-C-A in Fig.1) determines the minimum possible execution time for parallel computing. So the basic principle of this algorithm is to assign the mature node with the largest exit path length at current time. Load balance is achieved by assigning this mature node to the processor with the minimum accumulation time. The steps of the algorithm are given as follows:

(1) Calculate the length of exit path for each node in the DAG. This can be accomplished by a recursive algorithm:

   a) The length of exit path for the exit node is equal to the execution time of this node;

   b) The length of exit path for other nodes is equal to the execution time of this node plus the exit path length of its downstream node.

(2) Assign tasks following the steps below:

   a) Choose the processor with the minimum accumulation time, P_{min};

Accepted manuscript

ACCEPTED MANUSCRIPT

b) Find a mature node with the largest exit path length;

c) If this mature node can be found, assign this node to P_{min};

d) If there is no mature node available, assign a dummy node to P_{min}. The length of the dummy node is equal to the difference between $AT(P_i)$ and $AT(P_{min})$. Here $AT(P_i)$ is the smallest accumulation time that is strictly larger than $AT(P_{min})$.

e) Repeat a)-d) until all tasks are completed.

When the minimum parallel computing time is obtained using the above algorithm TMSR for channel routing processes is calculated by dividing the serial computing time by the minimum parallel computing time.

Step 3: Estimating the overall TMSR

The overall TMSR can be calculated by combing the TMSRs of hillslope processes and of channel routing processes:

$$R_{all} = 1/(p_{hs}/R_{hs} + p_{ch}/R_{ch}) \quad (1)$$

where R_{all} is the overall TMSR; R_{hs} and R_{ch} are the TMSRs for hillslope processes and channel routing processes, respectively; p_{hs} and p_{ch} are the proportions of computing hillslope processes and channel routing processes in the total amount of computations respectively. $p_{hs} + p_{ch} = 1$. The values of p_{hs} and p_{ch} are determined by the simulation methods used for hillslope and channel processes and these values can be determined by profiling model simulations using a performance profiler (for example, Intel® VTune™).

Accepted manuscript

ACCEPTED MANUSCRIPT

4 Application

The proposed TMSR estimation method was implemented using the Python 2.6 programming language. Six watersheds in China (the Qingshuihe watershed in Hebei province and its two sub-watersheds, the Fenkeng watershed in Jiangxi province and its two sub-watersheds) were selected as the study area (Fig. 2). The area, shape and sub-basin division information of these watersheds is shown in Table 1. The basin elongation ratio was adopted to represent the shape of watershed and it was defined as the ratio of the diameter of a circle with the same area as the basin to the maximum basin length (Schumm, 1956). The basin elongation ratio is large for compact watersheds (e.g. Fenkeng1) and small for long narrow watersheds (e.g. Qingshuihe2 and Fenkeng2). The numbers of sub-basins were selected in order to represent the range from 40 to 500, a range determined empirically. The 90m resolution SRTM (Shuttle Radar Topography Mission) DEM data were used for the TMSR calculation of these six watersheds with different numbers of sub-basins.

(Fig. 2 is about here)

(Table 1 is about here)

TMSRs for different numbers of processors form a TMSR curve (as shown in Fig. 3). It should be pointed out that TMSRs in these curves are theoretical values calculated by Equation (1) rather than experimental values by running a real distributed hydrological model. Fig. 3 shows the estimated TMSR curves for parallel computing of both sub-processes (i.e. hillslope and channel routing processes) and the model in the Qingshuihe watershed when the number of sub-basins is 101. All TMSR

Accepted manuscript

ACCEPTED MANUSCRIPT

curves reached plateaus after the numbers of processes exceeded certain thresholds. These thresholds represent the maximum numbers of processors that parallel computing of distributed hydrological models can make use of. Parallelizing channel routing processes has much smaller TMSRs than parallelizing hillslope processes, so reducing p_{ch} improves overall TMSRs dramatically.

(Fig. 3 is about here)

Fig. 4 shows the estimated TMSRs under enough number of processors for different numbers of sub-basins in the six watersheds assuming p_{ch} is 10%. It is obvious that the more the number of sub-basins, the larger the TMSRs were. For the hillslope processes, the relationships between TMSR and number of sub-basins were quantitatively similar for watersheds with different areas and shapes. This reflected the similarities among all watersheds to some extent. Meanwhile for the channel routing processes and the whole model, the relationships between TMSR and number of sub-basins were affected by the shapes of watersheds. Generally, TMSRs are larger for compact watersheds (e.g. Fenkeng2) than for long narrow watersheds (e.g. Qingshuihe1 and Fenkeng1).

(Fig. 4 is about here)

5 Conclusion

This paper proposed a TMSR estimation method for parallel computing of grid-based distributed hydrological models. In this method, TMSRs for hillslope processes and channel routing processes are calculated separately. Then they are combined to obtain overall TMSR. A preliminary application showed that the more

Accepted manuscript

ACCEPTED MANUSCRIPT

the number of sub-basins, the larger the TMSRs and that the compact watersheds had larger TMSRs than the long narrow watersheds.

Acknowledgements

This study was funded by the National High-Tech Research and Development Program of China (No. 2011AA120305) and the National Natural Science Foundation of China (No. 41023010). This study was also partly funded by the Program of International S&T Cooperation, MOST of China (No. 2010DFB24140). We thank Prof. Mauro Dell'Amico for his kindly providing the programming library for solving the multiprocessor scheduling problem.

References

Apostolopoulos, T.K., Georgakakos, K.P., 1997. Parallel computation for streamflow prediction with distributed hydrologic models. Journal of Hydrology 197(1-4), 1-24.

Borah, D.K., Bera, M., 2004. Watershed-scale hydrologic and nonpoint-source pollution models: Review of applications. Transactions of the ASAE 47(3), 789-803.

Dell'Amico, M., Martello, S., 1995. Optimal scheduling of tasks on identical parallel processors. ORSA Journal on Computing 7(2), 191-200.

Li, T.J., Wang, G.Q., Chen, J., 2010. A modified binary tree codification of drainage networks to support complex hydrological models. Computers & Geosciences

Accepted manuscript

ACCEPTED MANUSCRIPT

36(11), 1427-1435.

Li, T.J., Wang, G.Q., Chen, J., Wang, H., 2011. Dynamic parallelization of hydrological model simulations. Environmental Modelling & Software 26(12), 1736-1746.

Schumm, S. A., 1956. Evolution of drainage systems and slopes in badlands at Perth Amboy, New Jersey. Geological Society of American Bulletin, 67, 597-646.

Shirazi, B., Wang, M., Pathak, G., 1990. Analysis and evaluation of heuristic methods for static task scheduling. Journal of Parallel and Distributed Computing 10(3), 222-232.

Vivoni, E.R., Mascaro, G., Mniszewski, S., Fasel, P., Springer, E.P., Ivanov, V.Y., Bras, R.L., 2011. Real-world hydrologic assessment of a fully-distributed hydrological model in a parallel computing environment. Journal of Hydrology 409(1-2), 483-496.

Wang, H., Wang, G.Q., Gao, J., Fu, X.D., 2010. Parallel characteristics of river basin based on temporal-spatial-discrete approach. Sciencepaper Online 5(7), 494-498 [in Chinese].

Wang, H., Fu, X.D., Wang, G.Q., Li, T.J., Gao, J., 2011. A common parallel computing framework for modeling hydrological processes of river basins. Parallel Computing 37(6–7), 302-315.

Wang, H., Zhou, Y., Fu, X.D., Gao, J., Wang, G.Q., 2012. Maximum speedup ratio curve (MSC) in parallel computing of the binary-tree-based drainage network. Computers & Geosciences 38(1), 127-135.

Accepted manuscript

ACCEPTED MANUSCRIPT

Figure List

Fig. 1: Example of representing the dependence relationships among sub-basins (a) using a tree-structured Directed Acyclic Graph (b). Node A is the exit node, representing the outlet sub-basin. The first number in a DAG node represents the execution time of corresponding channel routing processes and the second number in a DAG node represents the exit path length correspondingly. Fig. 1(a) was modified from Wang et al. (2010).

Fig.2: Location, DEM, and the drainage network of the study areas, including the Qingshuihe watershed and its two sub-watersheds (Qingshuihe1 and Qingshuihe2), the Fenkeng watershed and it two sub-watersheds (Fenkeng1 and Fenkeng2).

Fig. 3: Estimated TMSR curves for parallel computing of both sub-processes and the model in Qingshuihe watershed when the number of sub-basins is 101. “Hillslope processes” and “Channel routing processes” in the legend means the TMSR curves for parallel computing of hillslope processes and channel routing processes, respectively. “Overall” means the overall TMSR curve of the model. “p_{ch}” means the proportion of computing channel routing processes in the total amount of computations.

Fig. 4: Estimated TMSRs under enough number of processors for different numbers of sub-basins in the six watersheds assuming p_{ch} (the proportion of computing channel routing processes in the total amount of computations) is 10%.

Table List

Table 1: Area, shape and sub-basin division information of the six watersheds.

Accepted manuscript

ACCEPTED MANUSCRIPT

Table 1

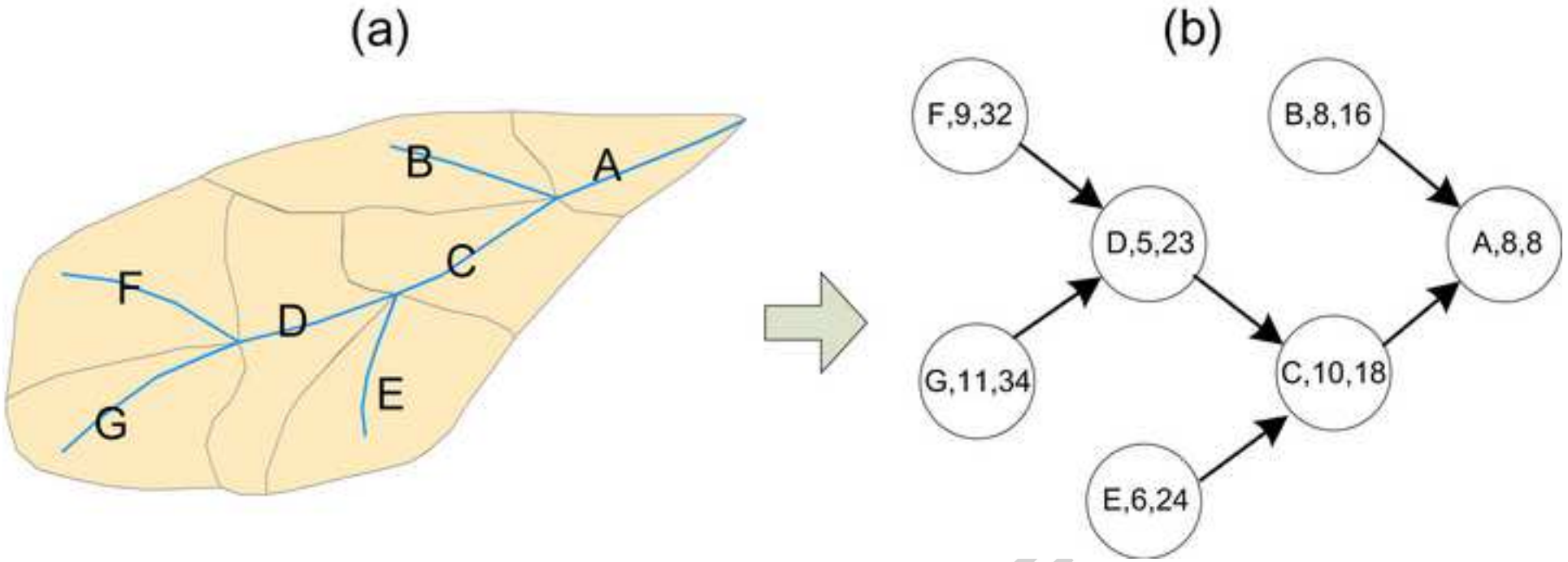
Area, shape and sub-basin division information of the six watersheds.

Name	Area (km^2)	Basin elongation ratio	List of sub-basin division numbers
Qingshuihe	2178	0.5	[55, 101, 203, 277, 491]
Qingshuihe1	231	0.56	[53, 134, 253, 346, 498]
Qingshuihe2	354	0.38	[45, 95, 227, 324, 507]
Fenkeng	6323	0.47	[57, 103, 193, 341, 501]
Fenkeng1	373	0.62	[45, 95, 235, 324, 487]
Fenkeng2	475	0.36	[45, 79, 193, 257, 476]

- A new method for estimating theoretical maximum speedup ratio was proposed.
- TMSRs for hillslope processes and channel processes were estimated respectively.
- Channel processes have much smaller TMSRs than hillslope processes.
- For one watershed, the more the number of sub-basins, the larger the TMSRs.
- The compact watersheds had larger TMSRs than the long narrow watersheds.

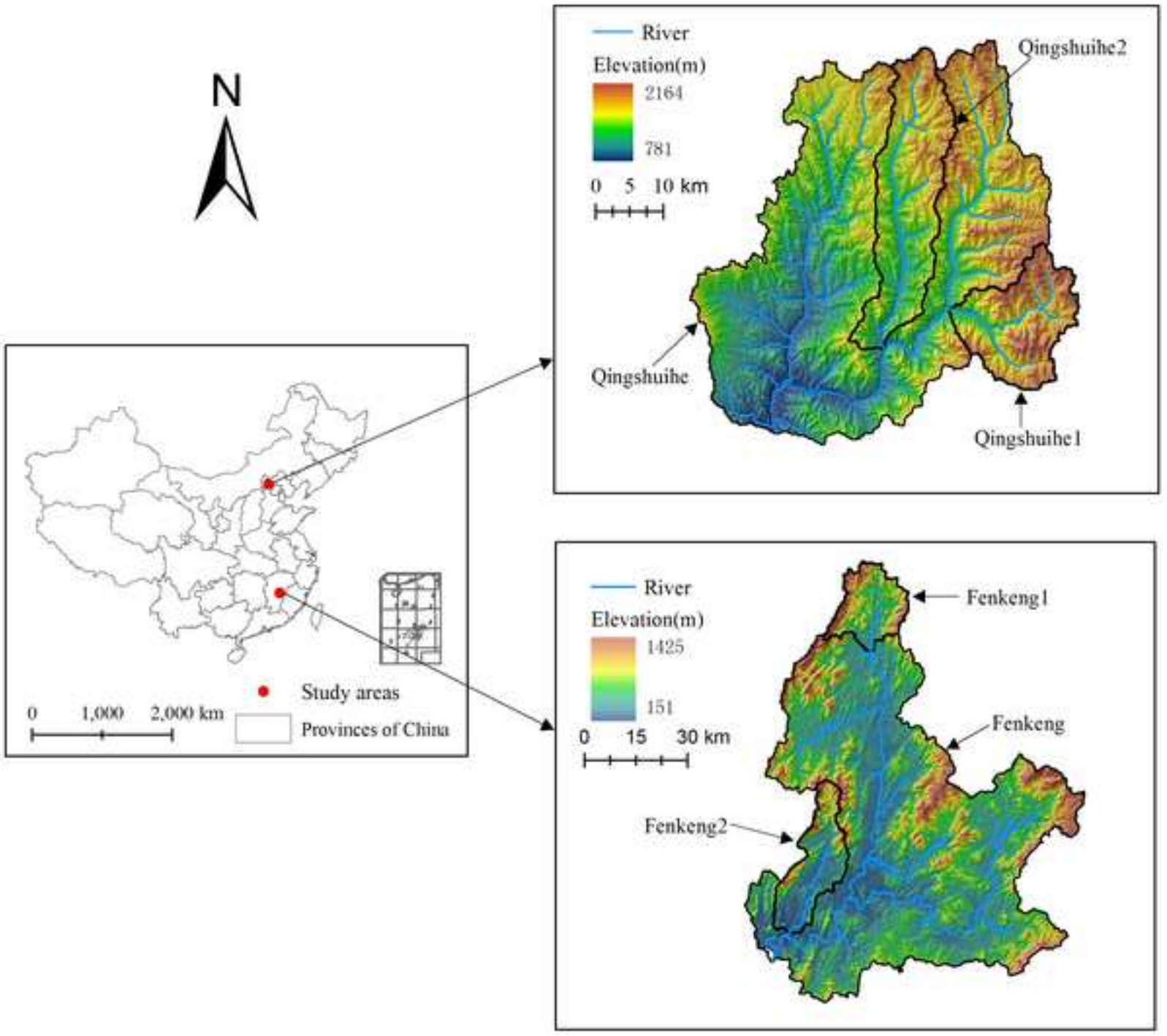
Accepted manuscript

Fig. 1



ACCEPTED MANUSCRIPT

Fig. 2



ACCEPTED MANUSCRIPT

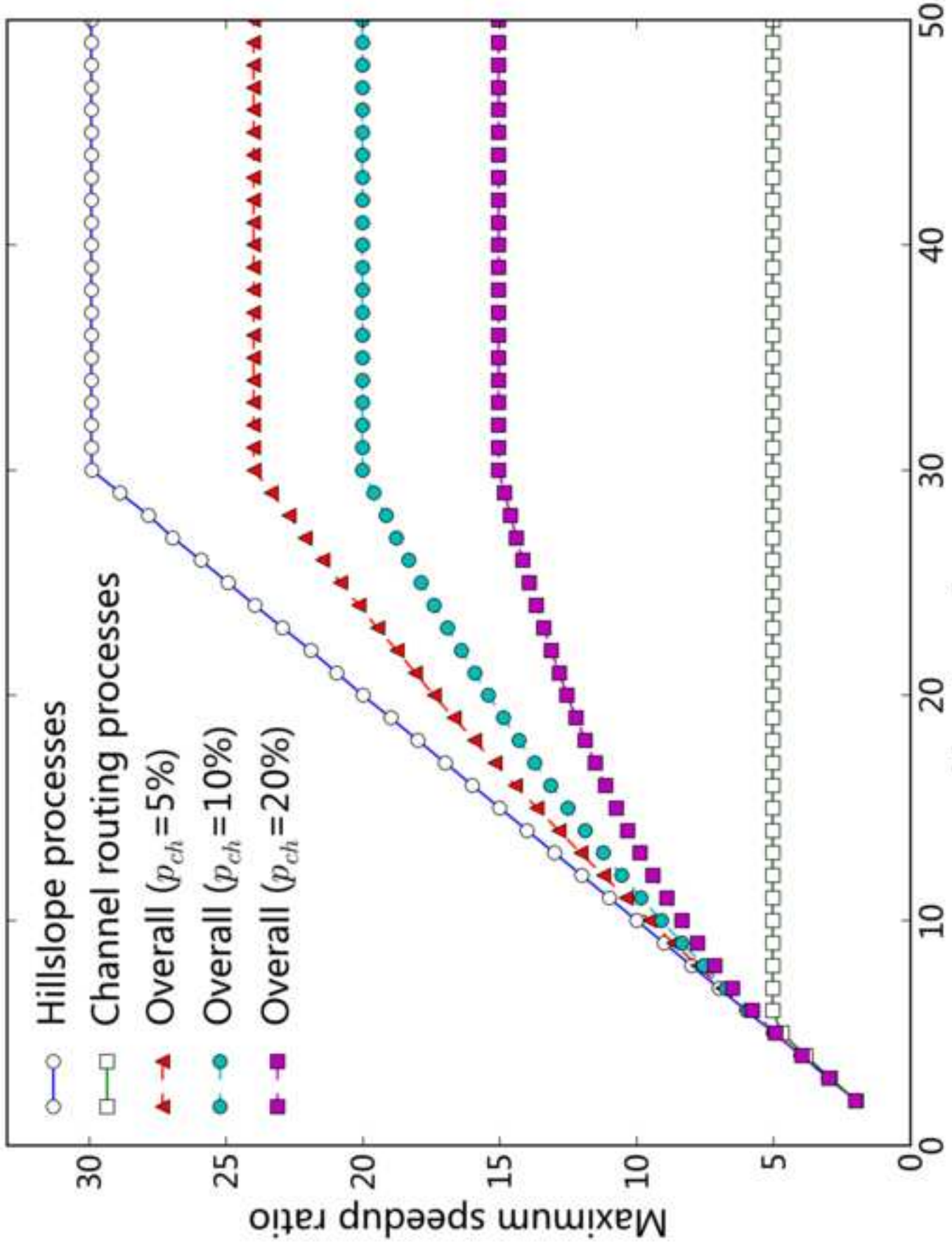

Hillslope processes
Channel routing processes
Overall (p_{ch}=5%)
Overall (p_{ch}=10%)
Overall (p_{ch}=20%)
Maximum speedup ratio
Number of processors
0
5
10
15
20
25
30
0
10
20
30
40
50

Fig. 3

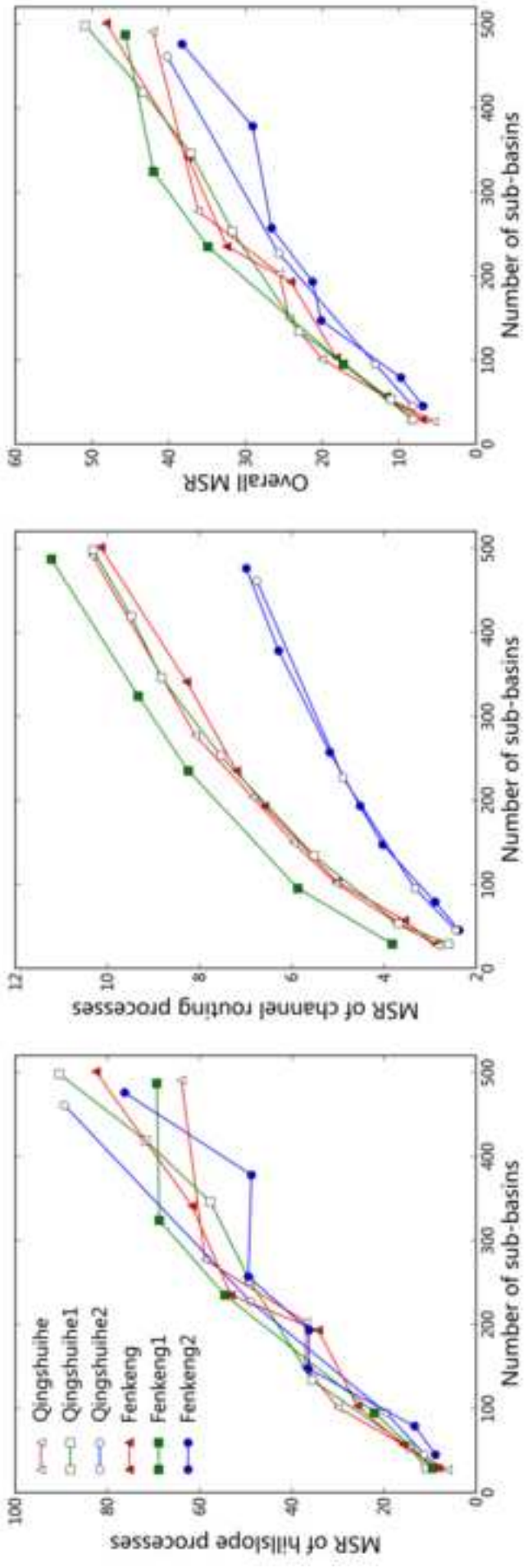

Overall MSR
Number of sub-basins
MSR of channel routing processes
Number of sub-basins
MSR of hillslope processes
Number of sub-basins
Qingshuihe
Qingshuihe1
Qingshuihe2
Fenkeng
Fenkeng1
Fenkeng2

Fig. 4