A Cyberinfrastructure Approach for Big Data-driven Microtopographic Analysis

Wenwu Tang[1,2] (WenwuTang@uncc.edu)

Minrui Zheng[1,2], Zachery Slocum[1,2], Jianxin Yang[1,2], Craig Allan[1,2]

[1] Center for Applied Geographic Information Science

[2] Department of Geography and Earth Sciences
University of North Carolina at Charlotte

Objectives

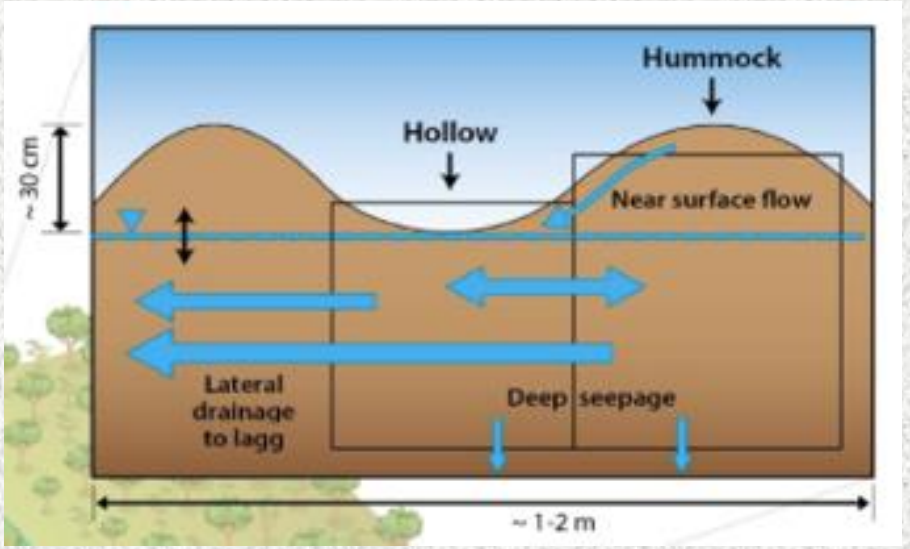
- Extract microtopographic features from very high resolution DEMs .
- Evaluate the capability of **cyberinfrastructure-enabled high-performance and parallel computing** in accelerating big data-driven GIS-based geospatial analytics.

Microtopography

- An important characteristic of the upland and wetland forests
  - Hummocks
  - Hollows (aka, depressions)

Each is typically at small scale ~ <1 – 15 m^2

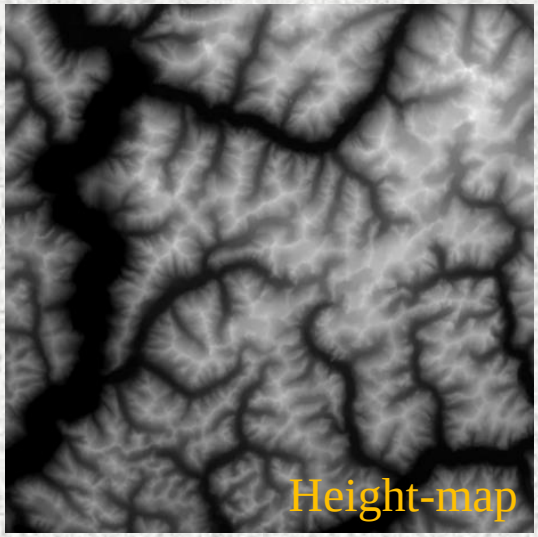
- Microtopographic features are impacted to different degrees by fluctuating water levels in the forested landscape
- Microtopography influences (Amoah et al. 2013; Trettin et al. 2016; Zheng et al. 2018)
  - Hydrologic storage properties of a watershed
  - Biogeochemical processes
  - Carbon cycling in the forest wetlands
  - Flood plain hydrologic exchanges
  - Vegetation spatial distribution

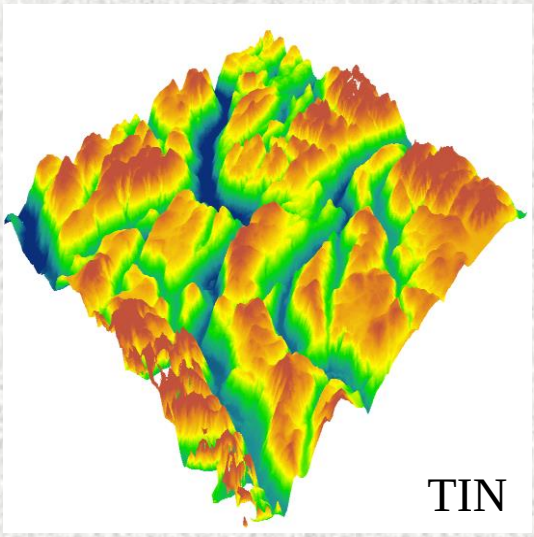


https://mnspruce.ornl.gov/node/615

Digital Elevation Models (DEMs)

- Representing the topographic surface of the Earth
- Provide the information of **spatial location** and **elevation**
- Acquired from (Li et al. 2005):
  - Photogrammetry | LiDAR | Land surveying |…
- Types:
  - Raster
  - Vector-based triangular irregular network (TIN)





https://answers.unity.com/questions/1375363/infinite-mountain-height-map.html
https://gisgeography.com/free-global-dem-data-sources/

Challenges

- Computational challenge
  - Generation of very high resolution DEMs
  - Extraction of microtopographic features

Photo source: http://www.ssocollection.com/wp-content/uploads/2014/01/picture-of-desktop-computer-95.jpg

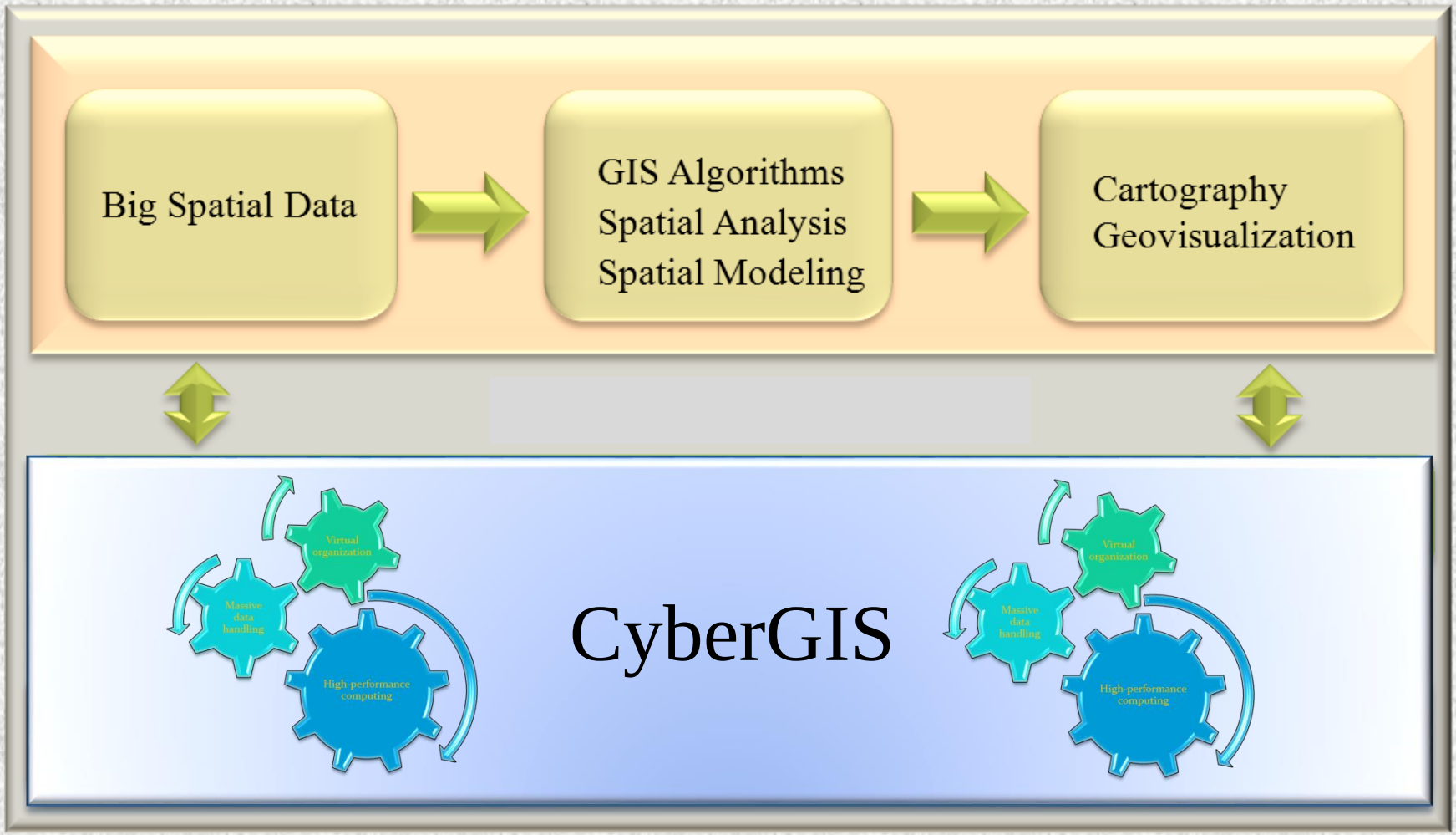
Computational Issues for Spatial Analysis

- Data- and compute-intensive
- Computer memory, compute time, and I/O

Stampede supercomputer at TACC (102,400 cores)

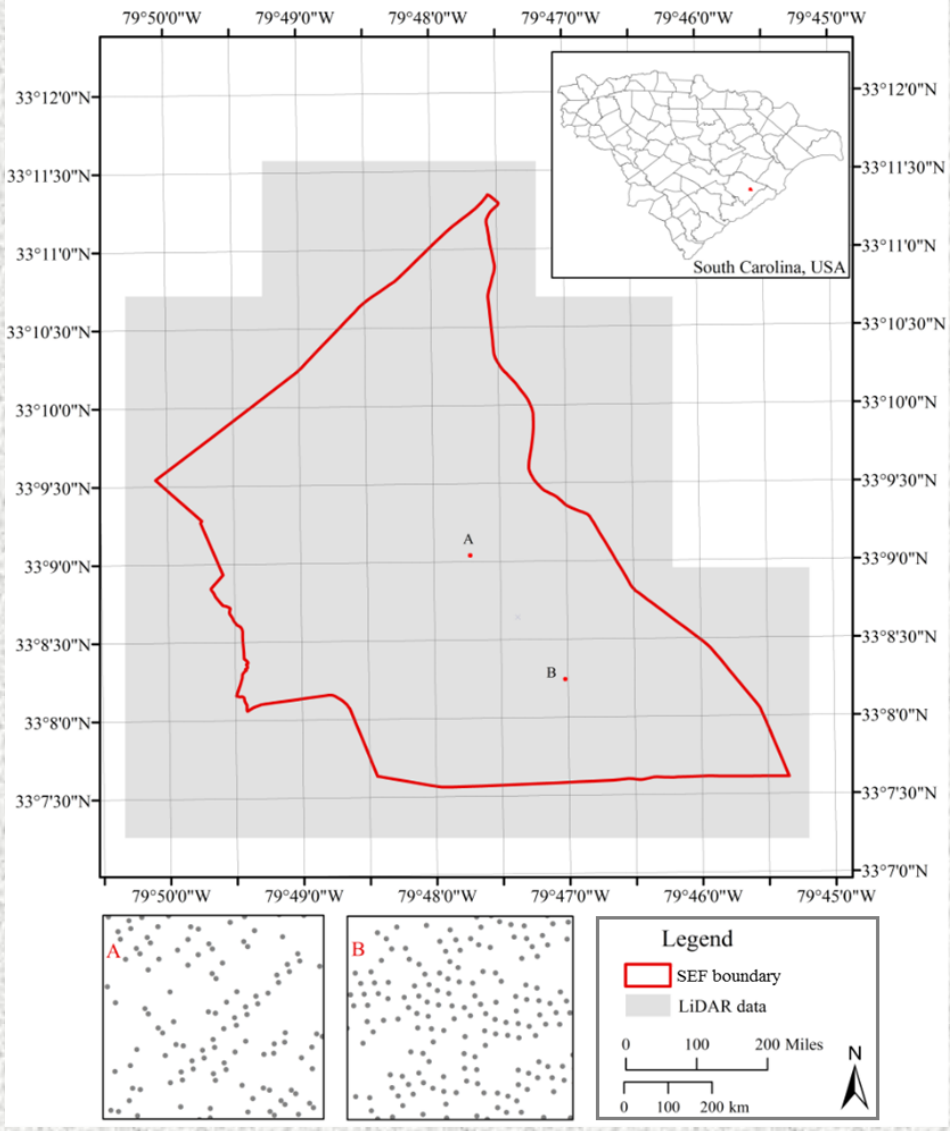
http://www.ssocollection.com/wp-content/uploads/2014/01/picture-of-desktop-computer-95.jpg
https://portal.tacc.utexas.edu/user-guides/stampede

CyberGIS for Big Spatial Data Analytics



Study Area: Santee Experimental Forest (USDA Forest Service)

- Long-term monitoring of field-scale data
  - Ecological,
  - Hydrologic
  - Climatic
  - Land resources
- Total Area
  - 2,468 ha



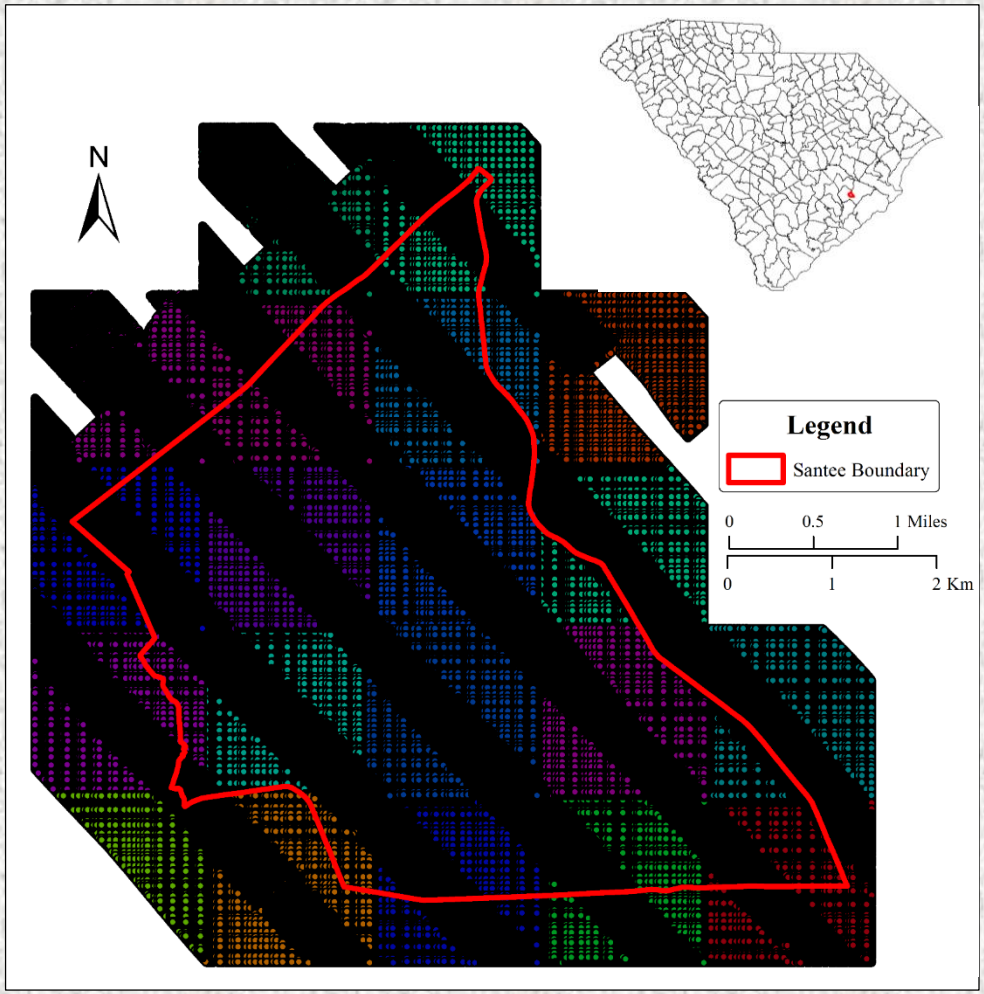
Tidal wetlands

- Tidal wetland
  - Periodically flooded in response to tidal level
  - Importance ecosystem service, such as carbon sequestration
  - Complex wetland hydrological process
- Categories
  - Salt marshes,
  - Mud flats
  - Mangrove Swamps
  - Tidal freshwater marshes
  - Tidal bottomland forests
- Tidal wetlands in SEF
  - Tidal freshwater marshes
  - Tidal bottomland forests




Photo source: https://www.dec.ny.gov/lands/4940.html
https://www.dec.ny.gov/lands/5120.html https://en.wikipedia.org/wiki/Mudflat
http://www.dcr.virginia.gov/natural-heritage/natural-communities/ncea1

Data - LiDAR



Published year	2007
Total tiles of dataset	20
Total size of dataset	3.5G
Geometry type	Point cloud
Unit	Meter
Point density	1 point/m^2

Downloadable from Santee Web GIS Portal: http://cybergis.uncc.edu/santee/

But …

- Total computing time using a single CPU
  - Total: 20.85 hours!

Solutions?

- High-performance and parallel computing!

https://i0.wp.com/hanusoftware.com/wp-content/uploads/event_218867862.png?w=360&ssl=1

Parallel Spatial Analysis Framework

- Very high resolution DEM
  - Scientific workflow of parallel generation of very high resolution DEM (Zheng et al. 2018)
- Depression Filling
  - Parallel priority-flood depression filling algorithm (Barnes 2016)
- Extract and quantify depressions
  - Area, number,…

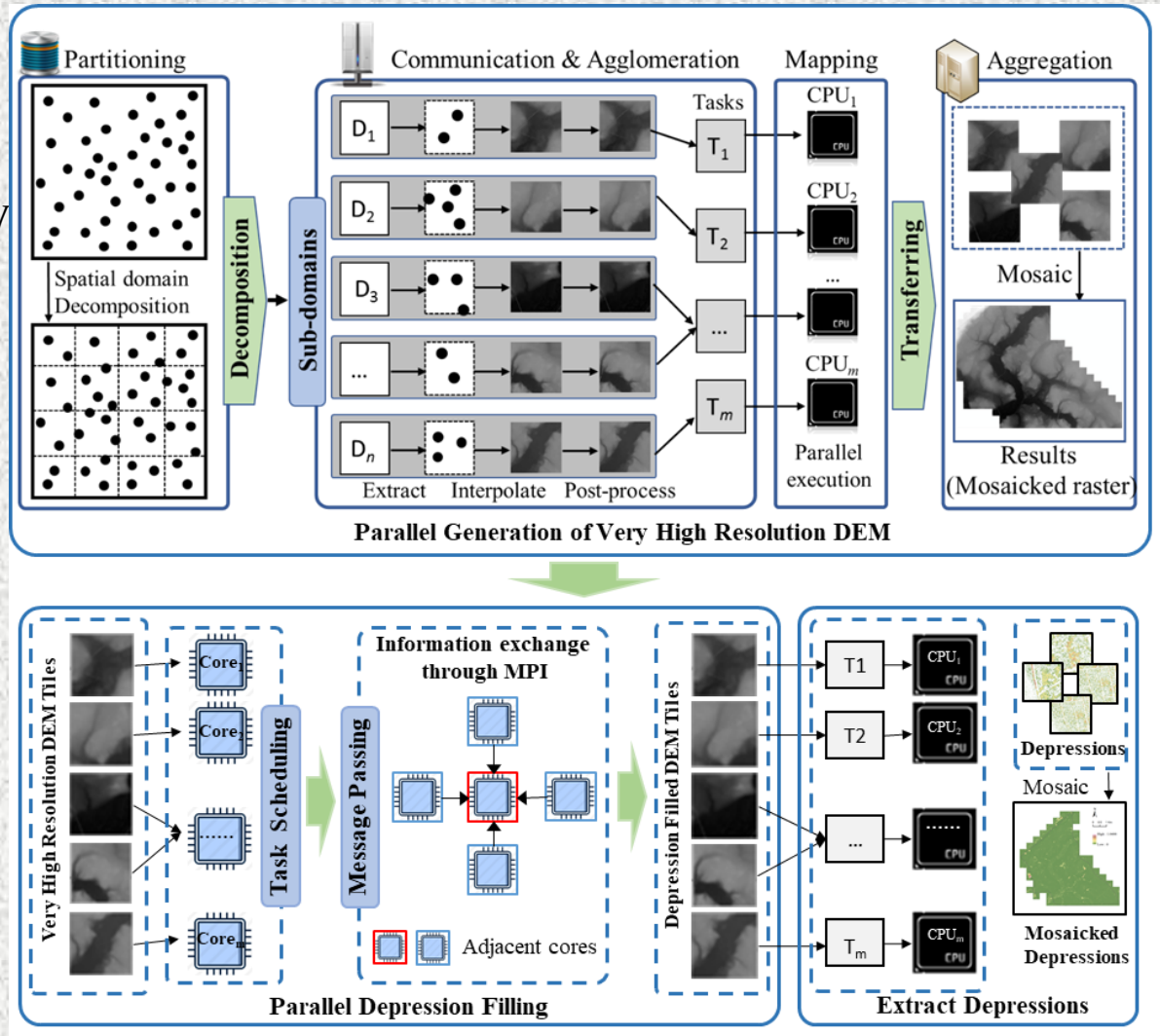


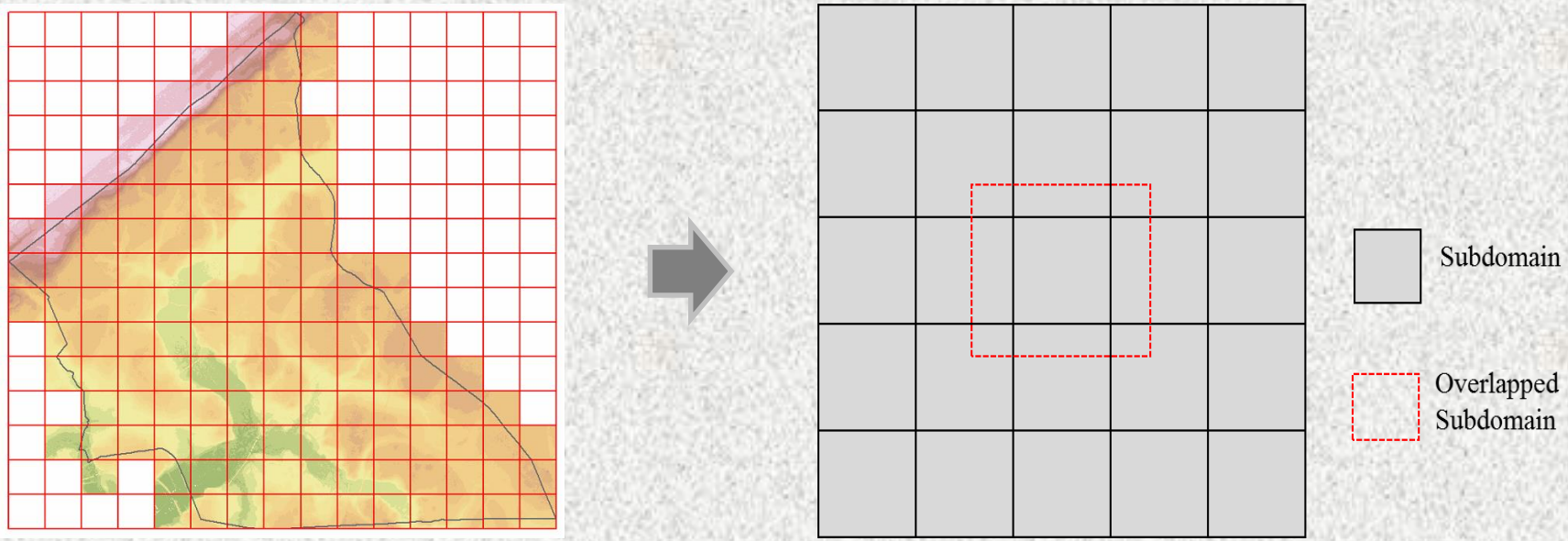
Figure adapted from Zheng et al. 2018

Generation of very high resolution DEMs

- Partitioning is the first step of the parallel spatial interpolation for the generation of DEM (see Zheng et al. 2018).
- Handling of communication among tasks associated with subdomains
- Spatial interpolation
  - IDW (Inverse Distance Weighting)

Parallel Computing Solutions

- Spatial Domain Decomposition



2D Spatial Domain Decomposition (15*15 tiles)

Depression filling

- Priority-Flood depression filling algorithm (Barnes 2016)

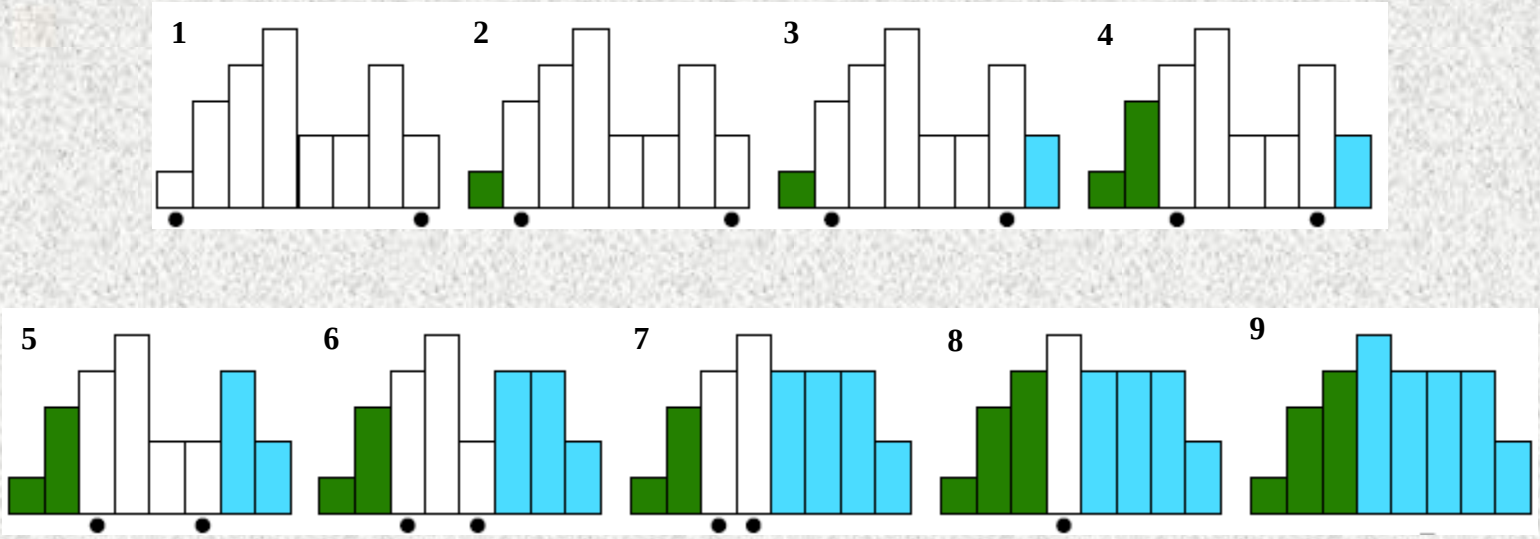


Figure source: Barnes, Richard. "Parallel priority-flood depression filling for trillion cell digital elevation models on desktops or clusters." Computers & Geosciences 96 (2016): 56-68.

Parallel Computing Resources

- Windows Cluster
  - **Sapphire**: **Windows** based HPC Cluster
    - 32 compute nodes/ 128 computing cores
    - Total Memory: 120GB
    - CPU: Intel(R) Core(TM) I7-2600 @ 3.40GHz
    - Software installed: ArcGIS, Python, R, TeamViewer, …
    - Job scheduling software: HPC Cluster Manager
    - Located at CAGIS center at University of North Carolina at Charlotte (http://gis.uncc.edu)

Parallel Computing Resources

- Linux Cluster
  - **Copperhead**: **Linux**-based HPC Cluster
    - 91 compute nodes/ 1,908 computing cores
    - Total Memory: 18,004 GB
    - CPU:
      - Dual Intel 3.2 GHz 8-core processors -Xeon E5-2667 v3 or v4
      - Dual Intel 2.6 GHz 16-core processors -Xeon E5-2697A v4
      - Dual Intel 3.0 GHz 18-core processors -Xeon Gold 6154
    - Located at University Research Computing at the University of North Carolina at Charlotte (https://urc.uncc.edu/)

Computing Performance Metrics

- Computing Time
- Speedup (S) and Efficiency (E)

$$S = \frac{T_1}{T_n}$$

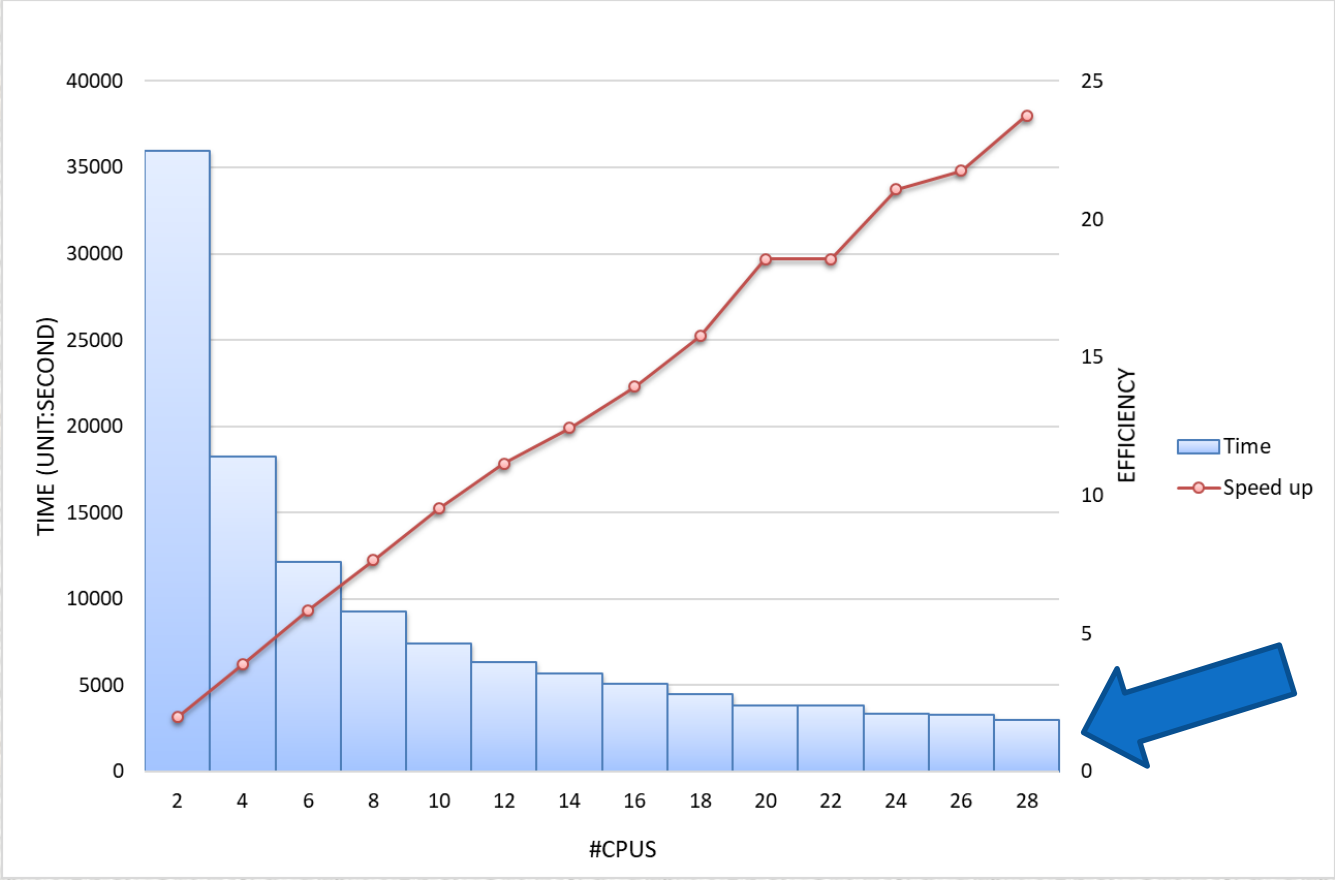
$$E = \frac{S}{n}$$

Where

T_1: computing time using 1 CPU;
T_n: computing time using n CPU;
n: number of CPUs

Experiment

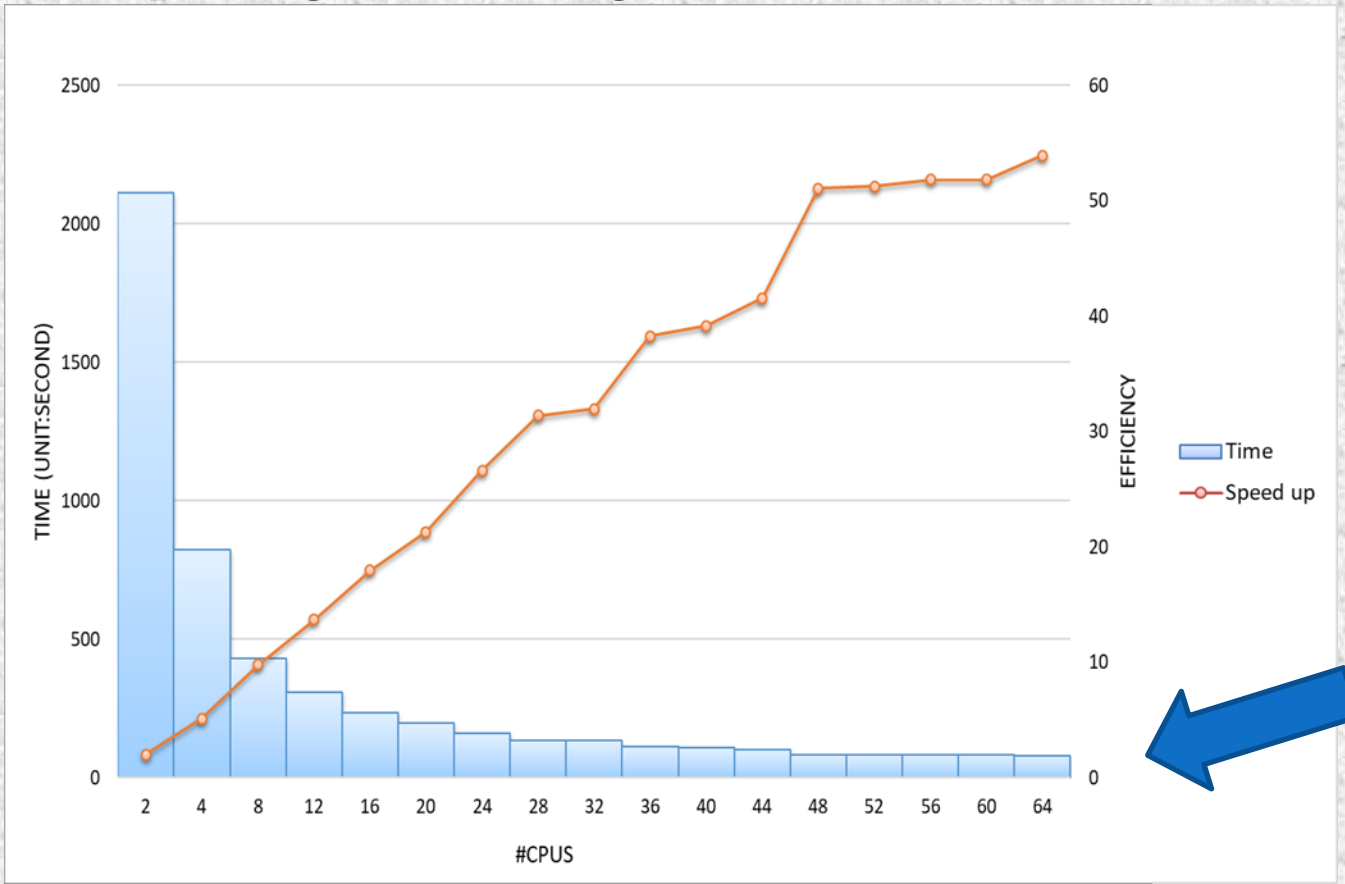
- Varying the #CPUs
  - Extraction of DEM based on spatial interpolation
    - Total computing time using 1 CPU: 70,853.724s (19.68 hours)



2,983 seconds!

Experiment

- Varying the #CPUs
  - Depression filling
    - Computing time using 1 CPU: 4,221,92 s (1.17 hours)



78 seconds!

Results

- DEM with very fine spatial resolution (see Zheng et al. 2018)
  - Resolution: 0.05m
  - Landscape size
    - 148,217 X 140,105

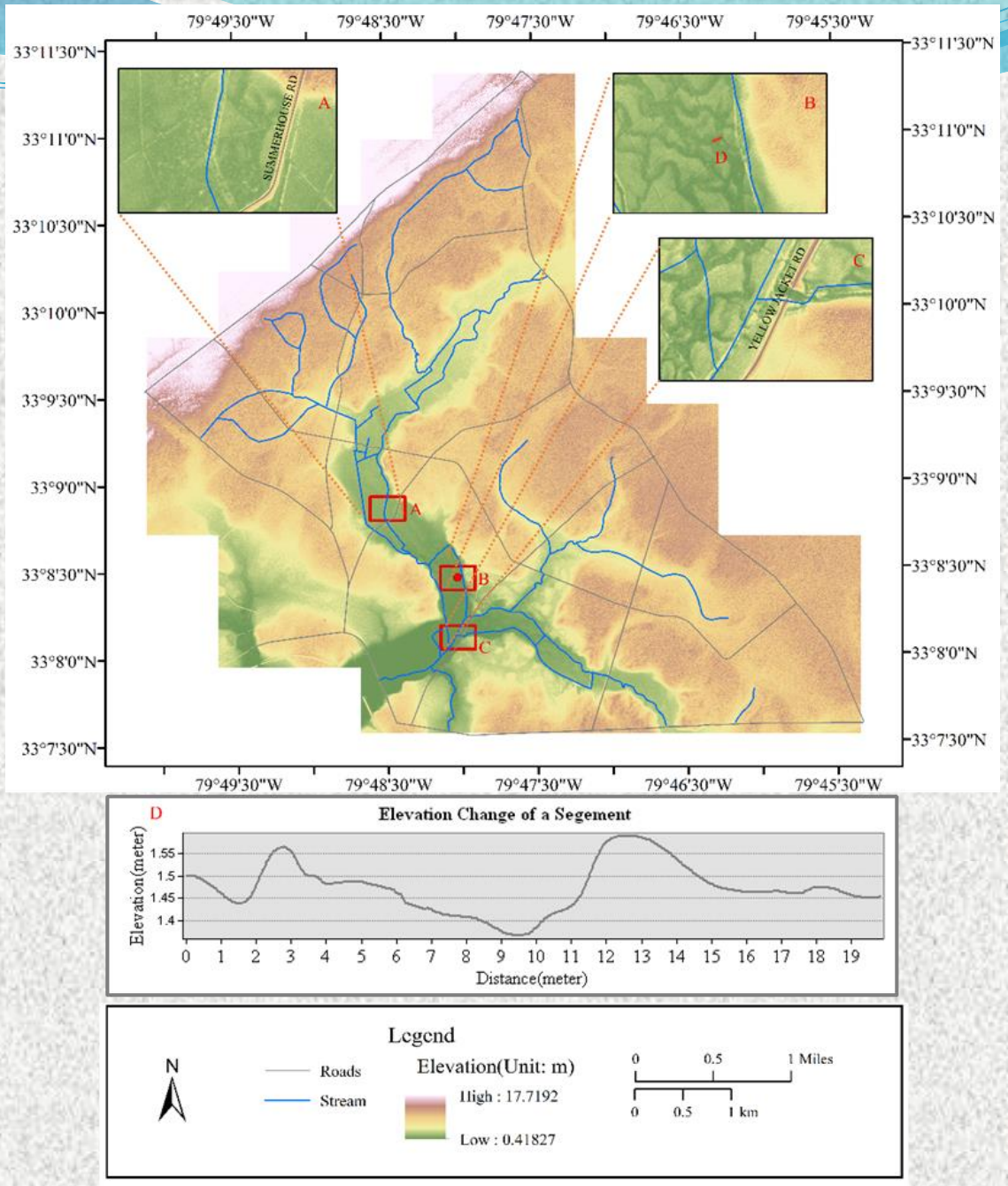
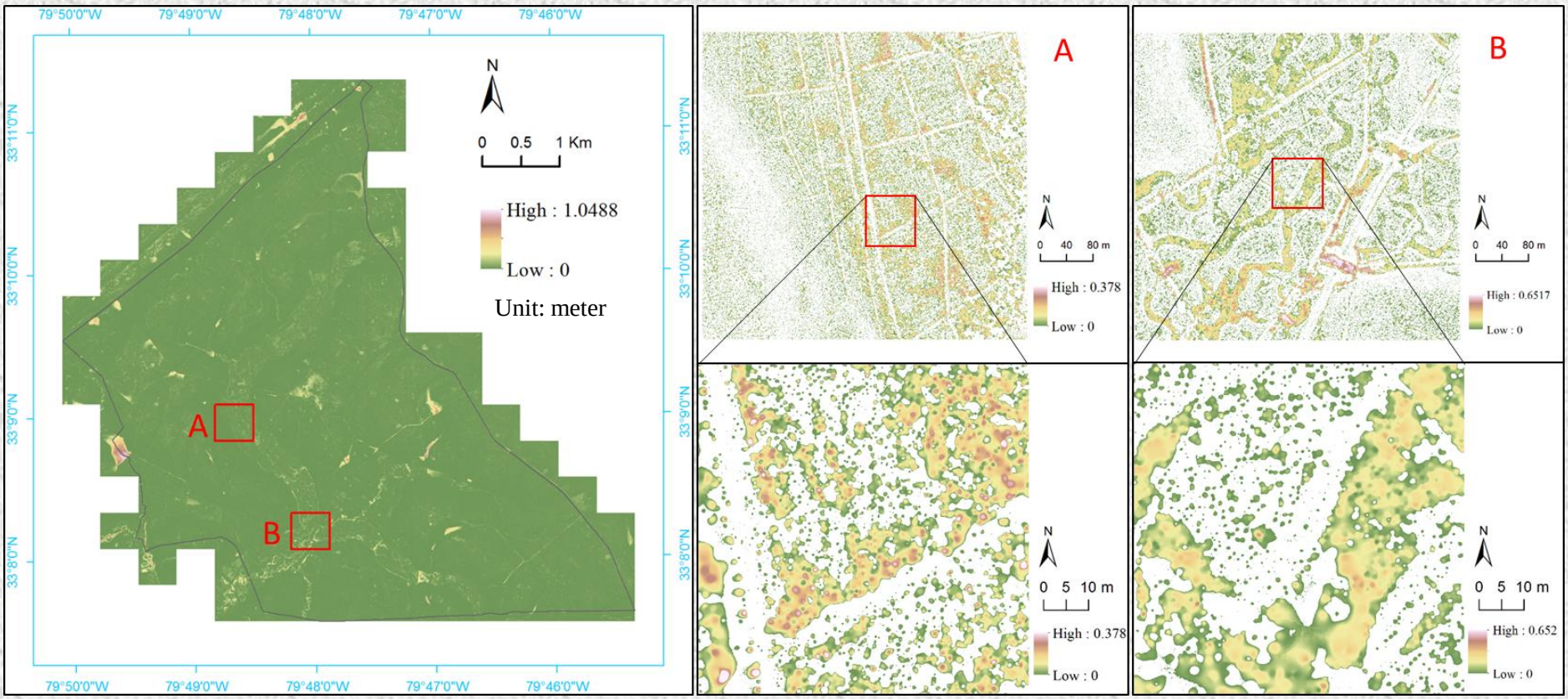


Figure adapted from Zheng et al. 2018

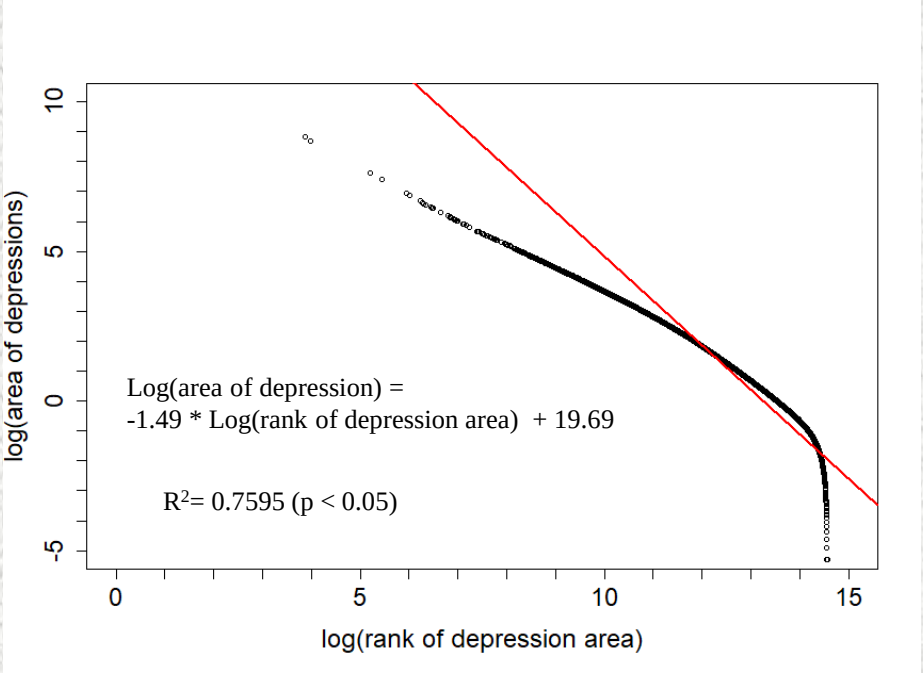
Results

- Depressions (or hollows)



Results

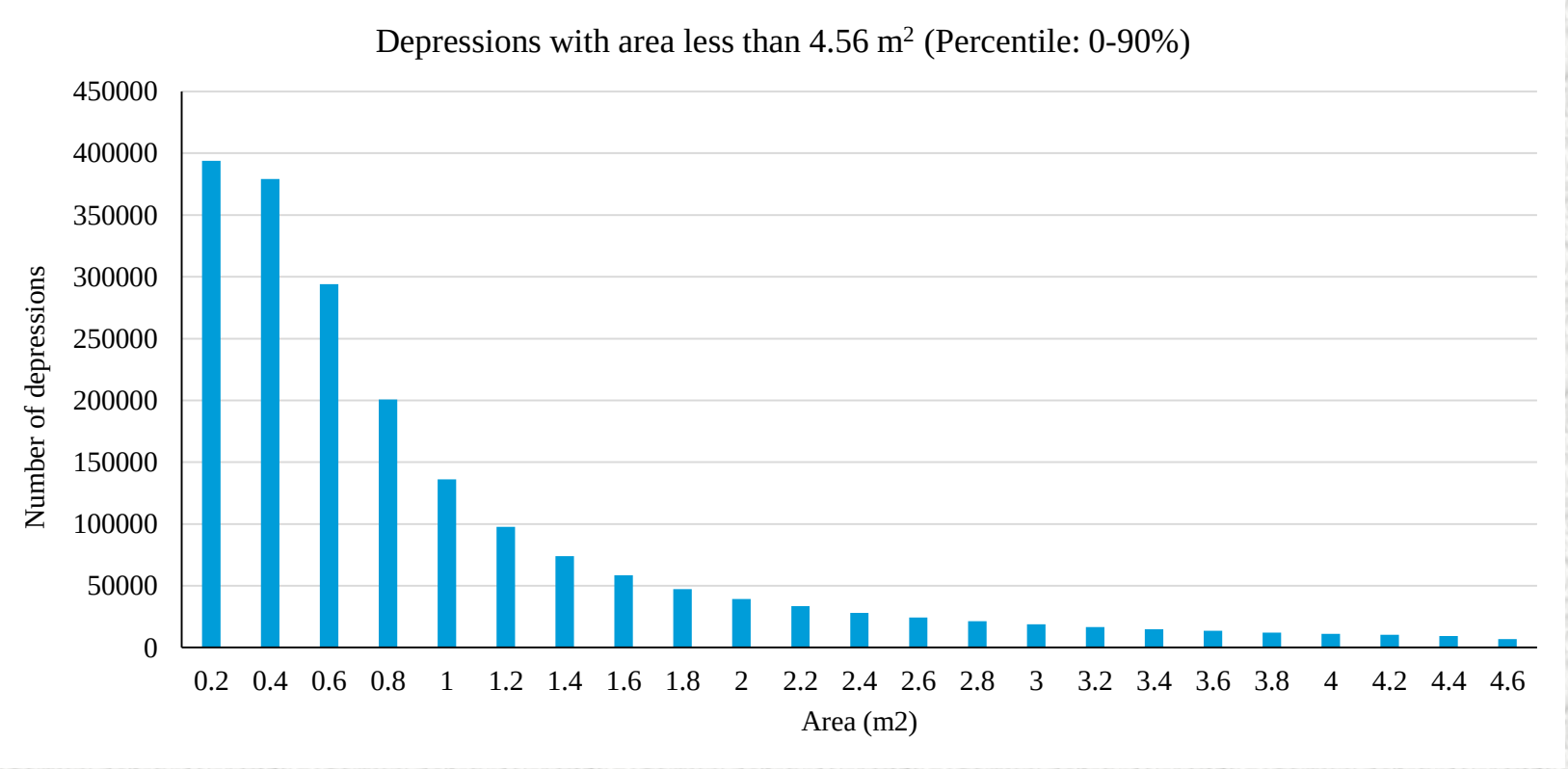
- Depressions (hollows)
  - Total number of depressions: 2,156,432
  - Total area of depressions: 7.63 km²
  - Mean area: 3.54 m²
  - Median area: 0.61 m²
  - Maximum: 63,064.81 m²
  - **Long-tail issue**
    - Rank-size distribution



(Figure plotted through sampling 50000 depressions)

Results

- 90% of the depressions are with area less than 4.56 m²
  - Number: 1,940,788
  - Total area: 1.65 km² (21.68%)



Concluding discussion

- Very high resolution DEMs provide substantial support for **delineating micro-level detail of topographic surfaces** such as hummocks or hollows in a forested wetland environment.
- **Spatial domain decomposition** strategies are pivotal in reaping the high-performance computing power for big spatial data analytics.
- The size of depressions presents a **long tail distribution** which highlights future discrimination of arbitrary depressions from actual ones.
- The high-performance and parallel computing solution proposed in this study demonstrated its ability to accelerate the **big data-driven GIS-based geospatial analytics**

References

- Bryan Farley, 2017, *Investigation of CO_2 and CH_4 Emission from Tital Freshwater and Non-Tidal Bottomland Forests*, Master Thesis, Department of Geography and Earth Sciences, the University of North Carolina at Charlotte, NC, USA (advised by **Craig J. Allan**)
- Amoah, Joseph KO, Devendra M Amatya, and Soronnadi Nnaji. 2013. "Quantifying watershed surface depression storage: determination and application in a hydrologic model." Hydrological Processes 27 (17):2401-2413.
- Barnes, Richard. "Parallel priority-flood depression filling for trillion cell digital elevation models on desktops or clusters." Computers & Geosciences 96 (2016): 56-68.
- Li, Zhilin, Qing Zhu, and Chris Gold. *Digital terrain modeling: principles and methodology*. CRC press, 2004.
- **Trettin, Carl C**, Brooke J Czwartacki, **Craig J Allan**, and Devendra M Amatya. 2016. "Linking freshwater tidal hydrology to carbon cycling in bottomland hardwood wetlands." In: Stringer, Christina E.; Krauss, Ken W.; Latimer, James S., eds. 2016. Headwaters to estuaries: advances in watershed science and management-Proceedings of the Fifth Interagency Conference on Research in the Watersheds. March 2-5, 2015, North Charleston, South Carolina. e-General Technical Report SRS-211. Asheville, NC: US Department of Agriculture Forest Service, Southern Research Station. 302 p.
- **Zheng, M., Tang, W.,** Lan, Y., Zhao, X., Jia, M., **Allan, C., Trettin, C.,** 2018. Parallel Generation of Very High Resolution Digital Elevation Models: High-performance computing for Big Spatial Data Analysis. Big Data in Engineering Applications. Springer.
- Kirwan, Matthew L., and J. Patrick Megonigal. "Tidal wetland stability in the face of human impacts and sea-level rise." Nature 504.7478 (2013): 53.

Acknowledgements

- NSF XSEDE supercomputing resource award (SES170007) "Accelerating and enhancing multi-scale spatiotemporally explicit analysis and modeling of geospatial systems"
- USDA Forest Service
- Dr. Carl Trettin From USDA Forest Service
- **2018-2022**, USDA Forest Service (Santee Experiment Forest), Development and Operation of a Web GIS-enabled Data Management System for the Santee Experimental Forest. [PIs: Wenwu Tang].

Thank you!
Questions?

gis.uncc.edu