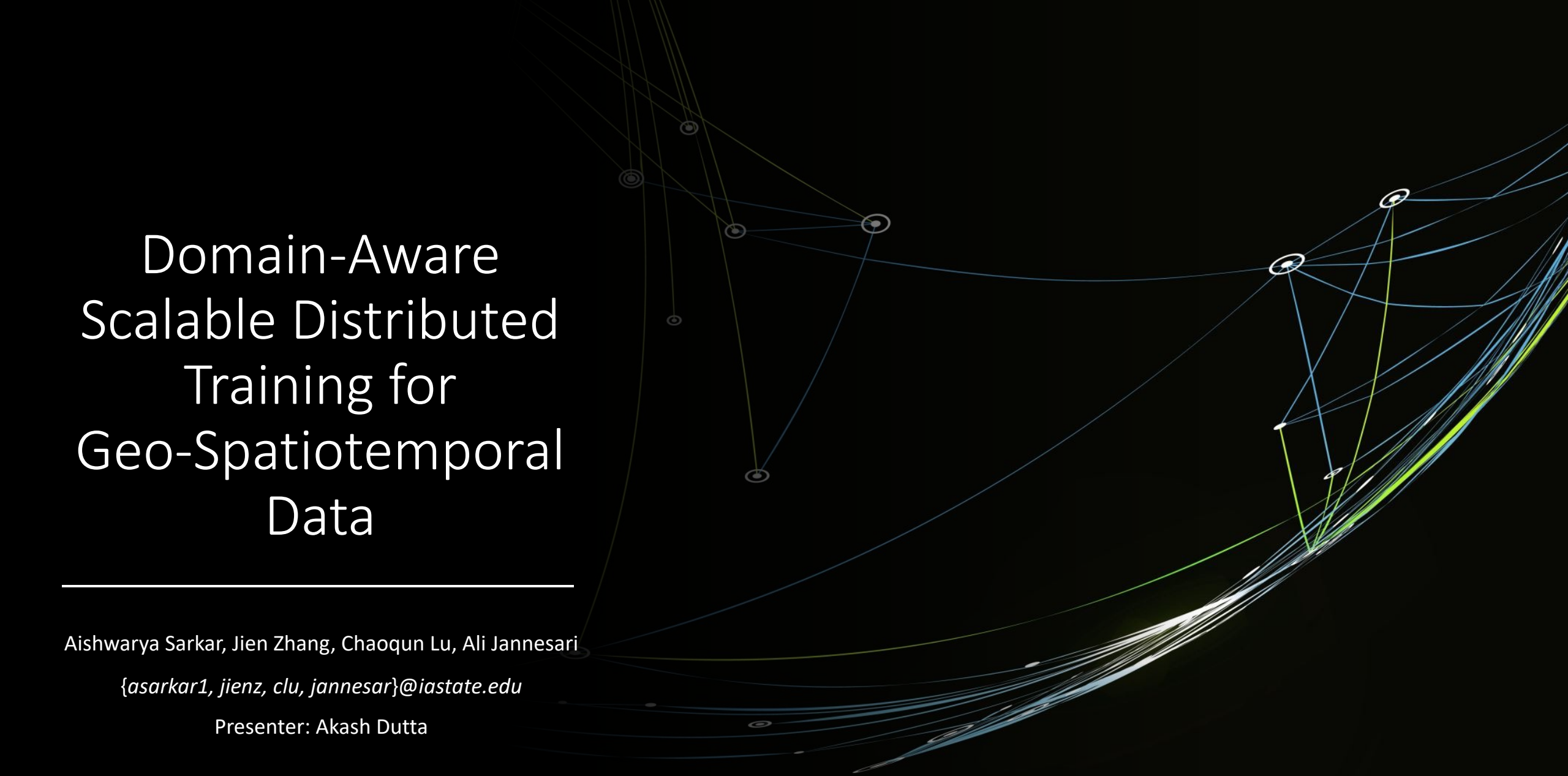
Domain-Aware Scalable Distributed Training for Geo-Spatiotemporal Data

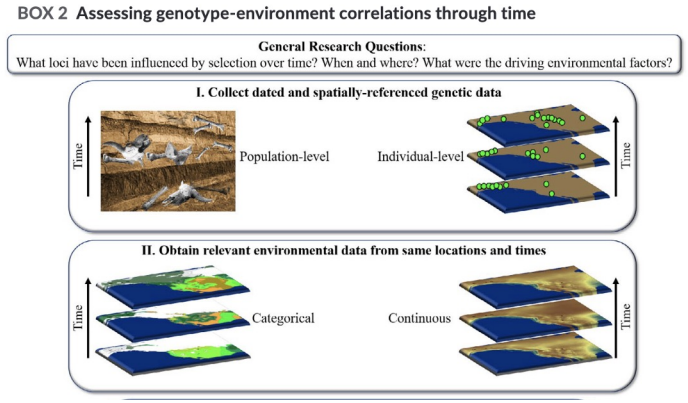
Aishwarya Sarkar, Jien Zhang, Chaoqun Lu, Ali Jannesari

{*asarkar1, jienz, clu, jannesar*}*@iastate.edu*

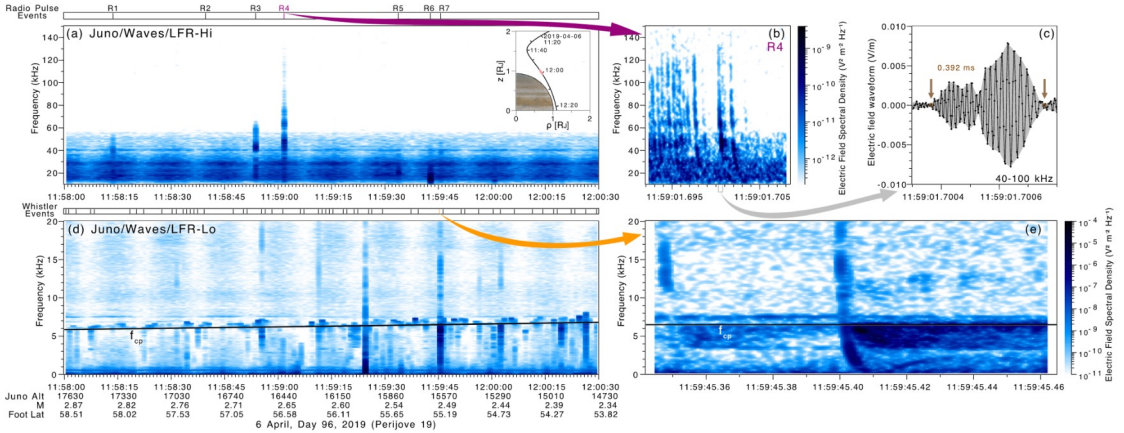
Presenter: Akash Dutta



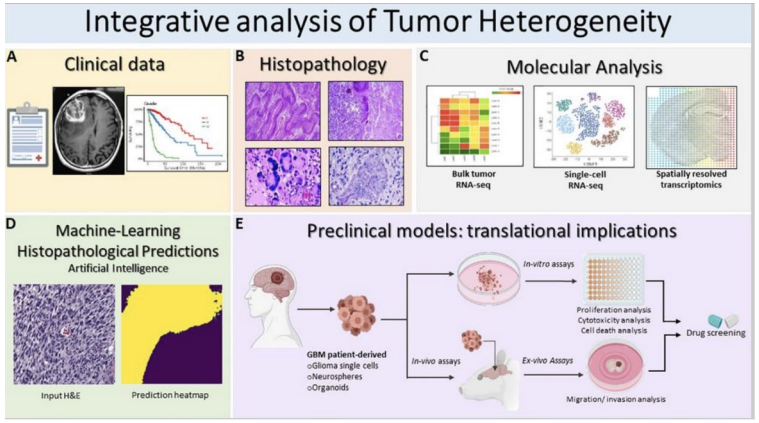
Growth in Spatiotemporal Data



Fenderson et al., . "Spatiotemporal landscape genetics: Investigating ecology and evolution through space and time." Molecular Ecology 29.2 (2020): 218-246.



Imai, Masafumi, et al. "High-spatiotemporal resolution observations of Jupiter lightning-induced radio pulses associated with sferics and thunderstorms." Geophysical Research Letters 47.15 (2020): e2020GL088397.

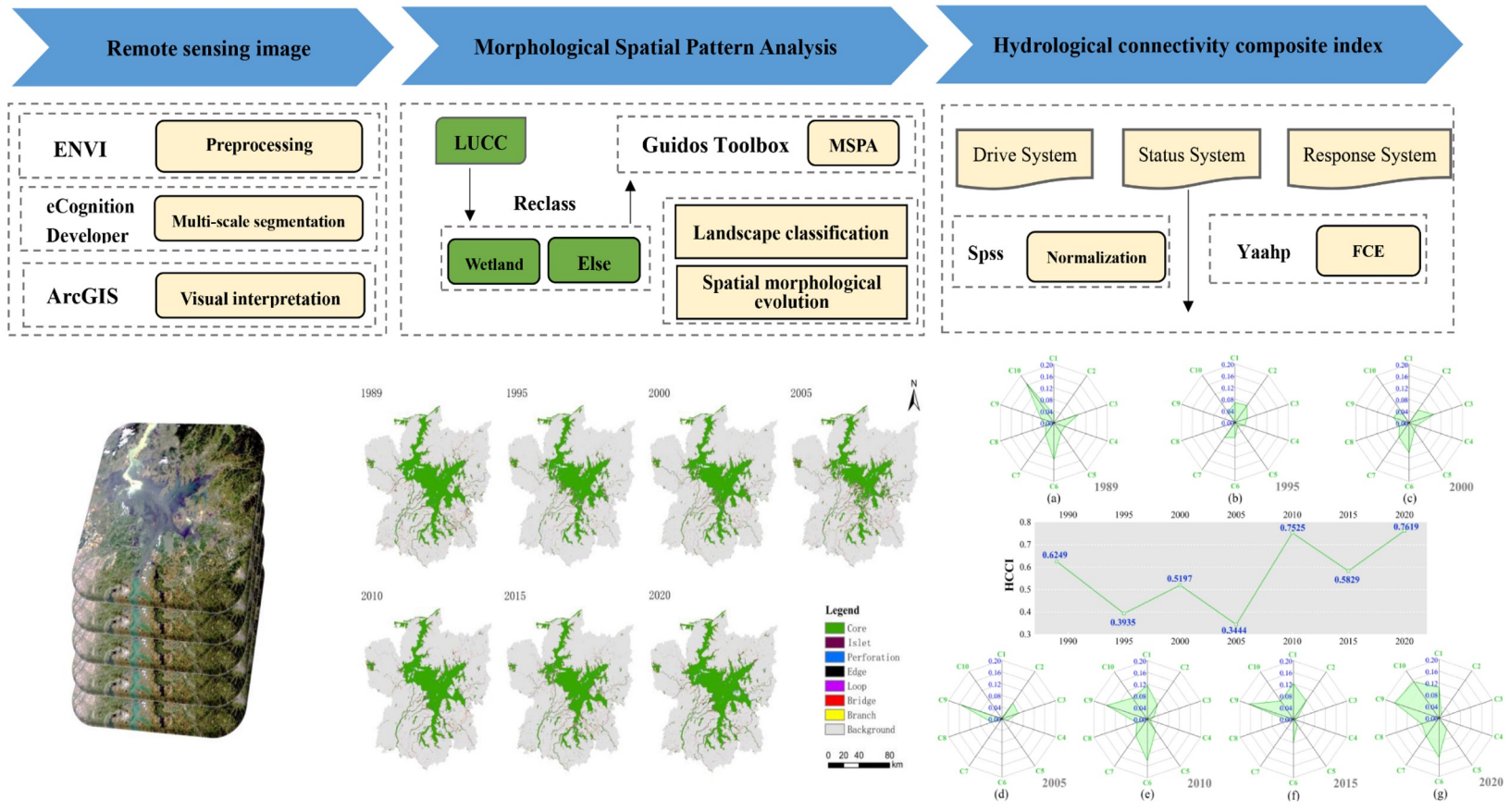


Comba, Andrea, et al. "Uncovering Spatiotemporal Heterogeneity of High-Grade Gliomas: From Disease Biology to Therapeutic Implications." Frontiers in Oncology 11 (2021).

FIGURE 3: The ID and locations of stations in our experiment.

Luo, Xianglong, et al. "Spatiotemporal traffic flow prediction with KNN and LSTM." Journal of Advanced Transportation 2019 (2019).

Spatiotemporal Data in Hydrology



Bhatti, Ahmad Zeeshan, et al. "Spatiotemporal hydrological analysis of streamflows and groundwater recharge for sustainable water management in Prince Edward Island, Canada." World Water Policy 7.2 (2021): 253-282

Approaches to Extract Pattern

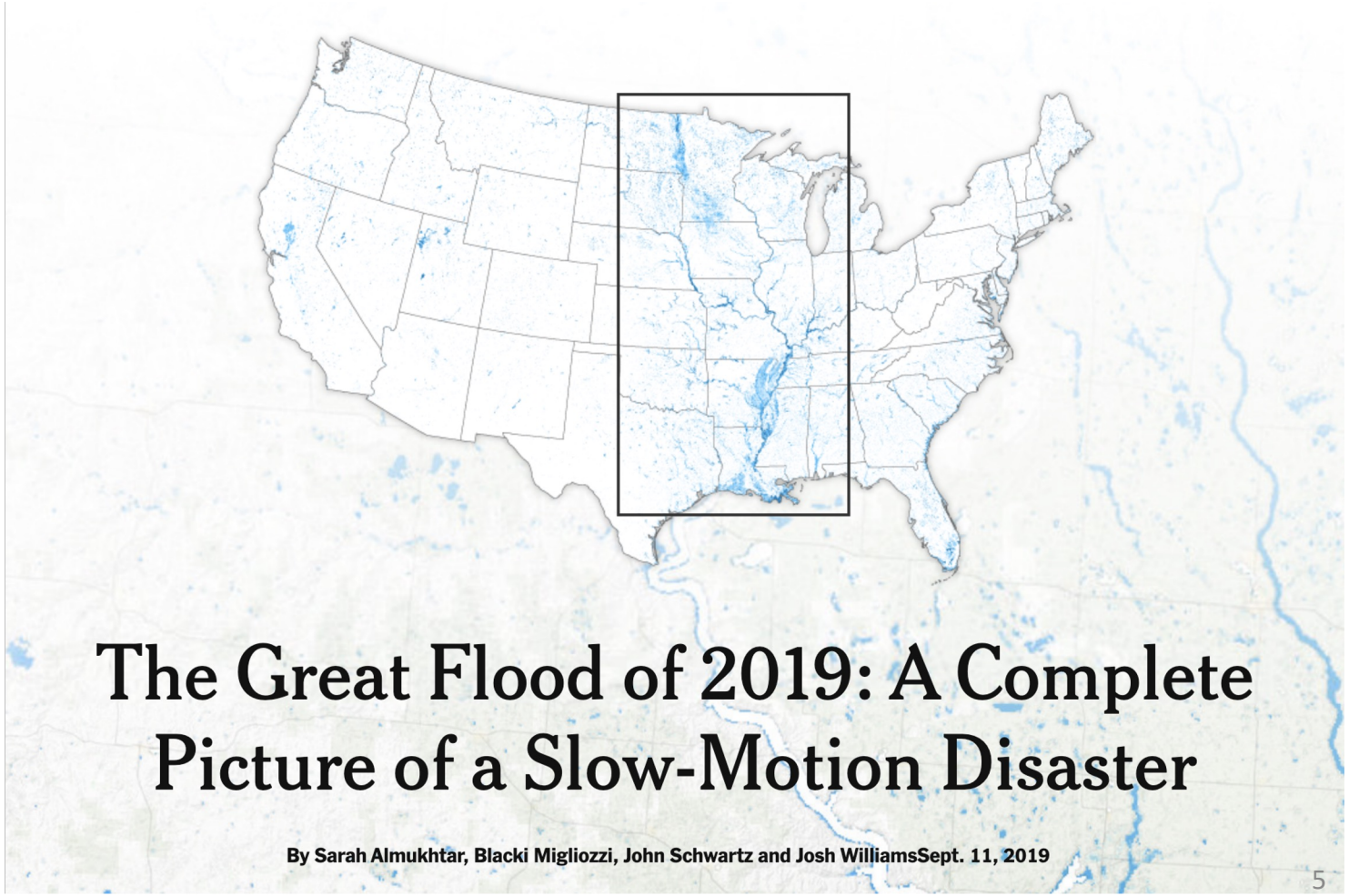
1. Domain:
   - Use domain-specific theories and equations
   - Too complicated
   - Difficult to solve deterministically
   - Requires massive computational resources
   - Not scalable

2. Data-driven:
   - Purely data driven
   - Lack of domain knowledge tends to sub-optimal performance
   - Difficult for policy makers to interpret – “black box”
   - Can be expensive for large dataset

Use Case: Flood Prediction

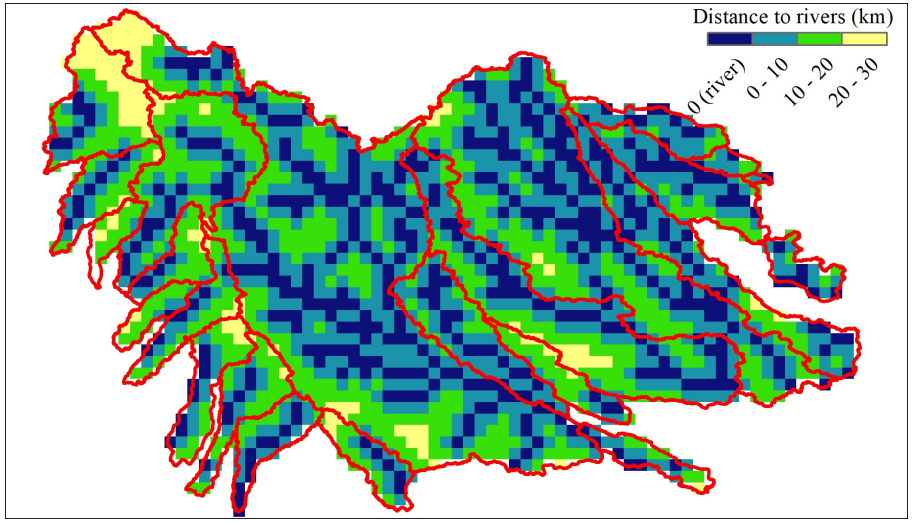
- The most frequent natural disaster
- Irreparable damages to farmlands and infrastructure
- In 2019, ¼ million acres of farmland was underwater for 4 months
- Performance still low

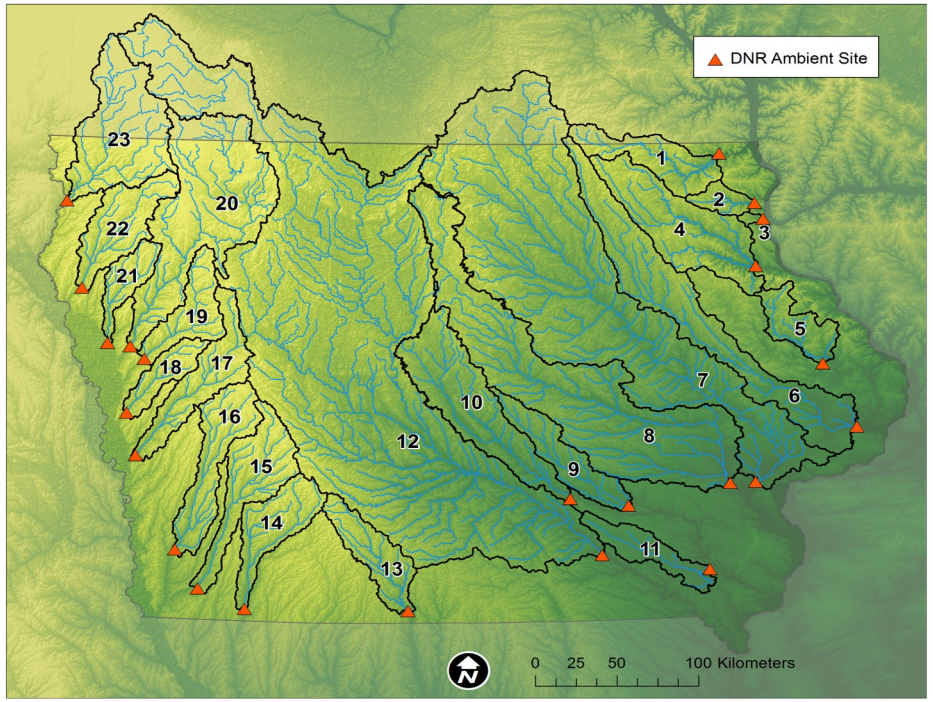
(notoriously tricky)



Problem Statement and Dataset

- Pixelated map of a mid-west U.S. state at 5 arc min ~ 2000 pixels
- 23 Watersheds with USGS[1] observation sites (▲)
- They vary characteristically



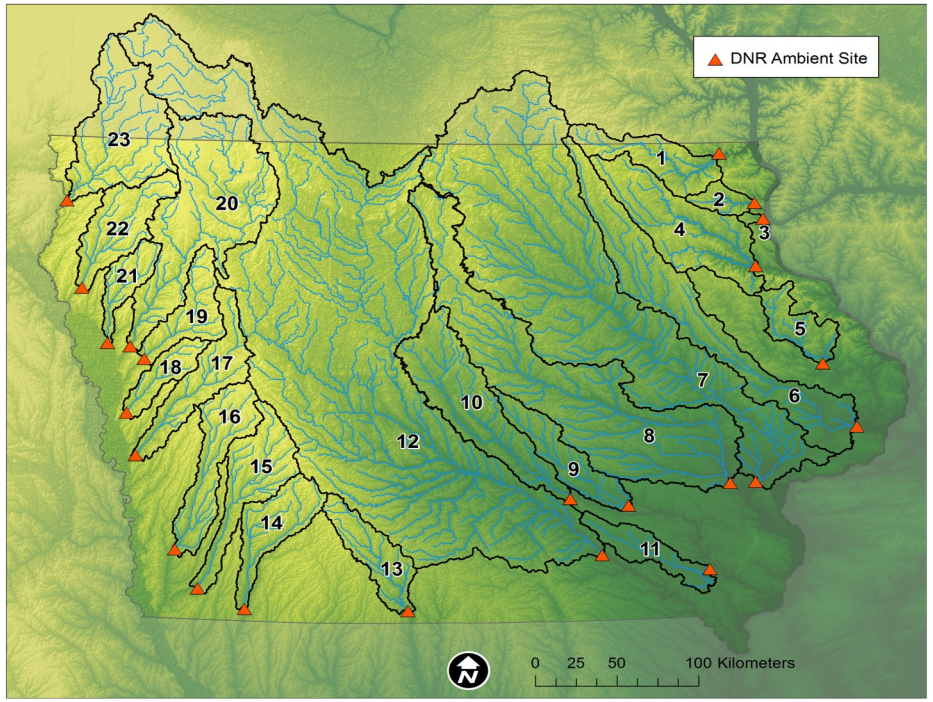


Jones, Christopher S., et al. "Iowa stream nitrate and the Gulf of Mexico." PloS one 13.4 (2018): e0195930.

[1] United States Geological Survey

Challenges

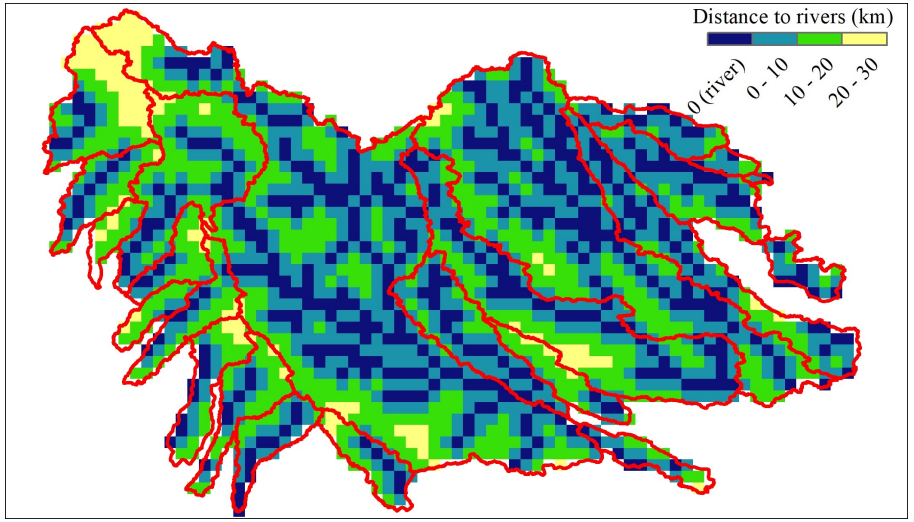
- Low prediction performance
- Each region needs separately trained models for more accurate prediction -- **expensive**
- Both deep learning and domain models take days to train on the whole data

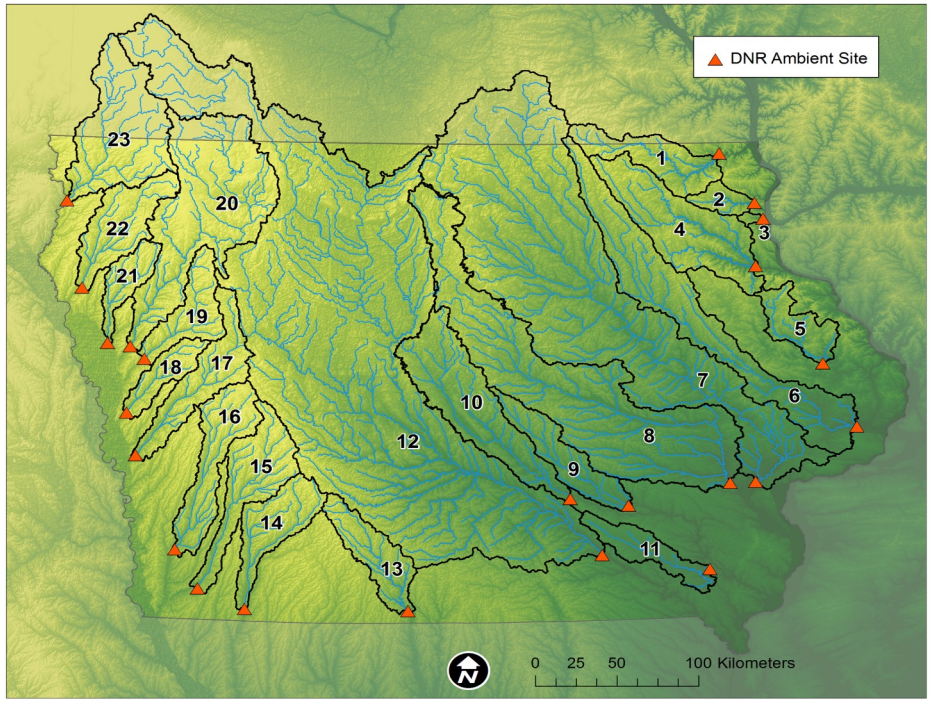


Jones, Christopher S., et al. "Iowa stream nitrate and the Gulf of Mexico." PloS one 13.4 (2018): e0195930.

Problem Statement and Dataset

- Input – Pixelated map, Distance, Precipitation
- Goal – To predict water discharge





Jones, Christopher S., et al. "Iowa stream nitrate and the Gulf of Mexico." PloS one 13.4 (2018): e0195930.

[1] United States Geological Survey

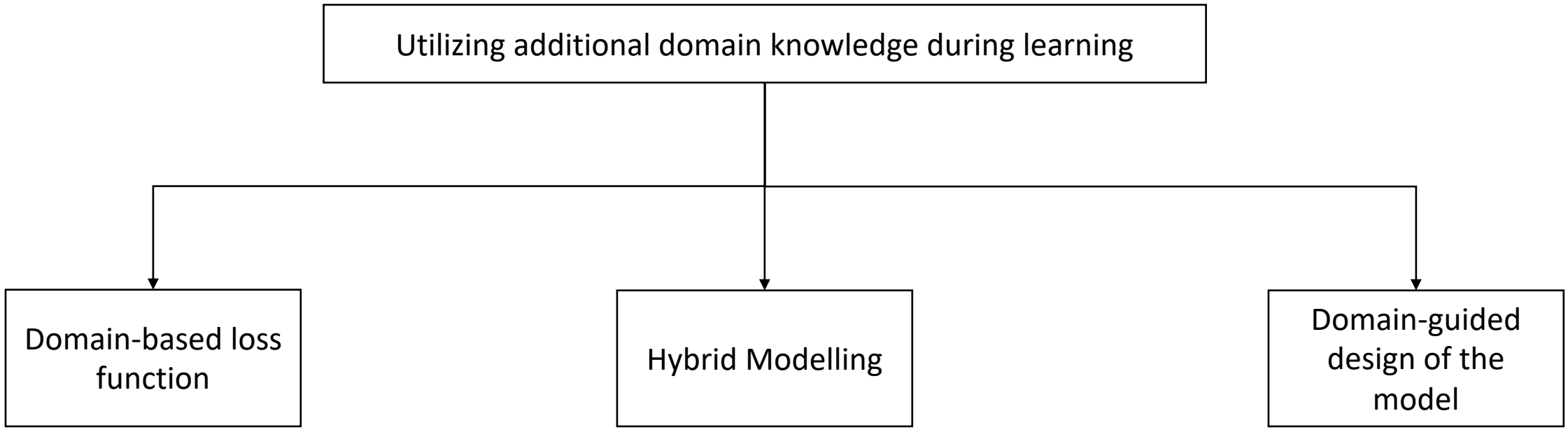
Introducing Dom-ST

Domain-Aware Spatiotemporal Network

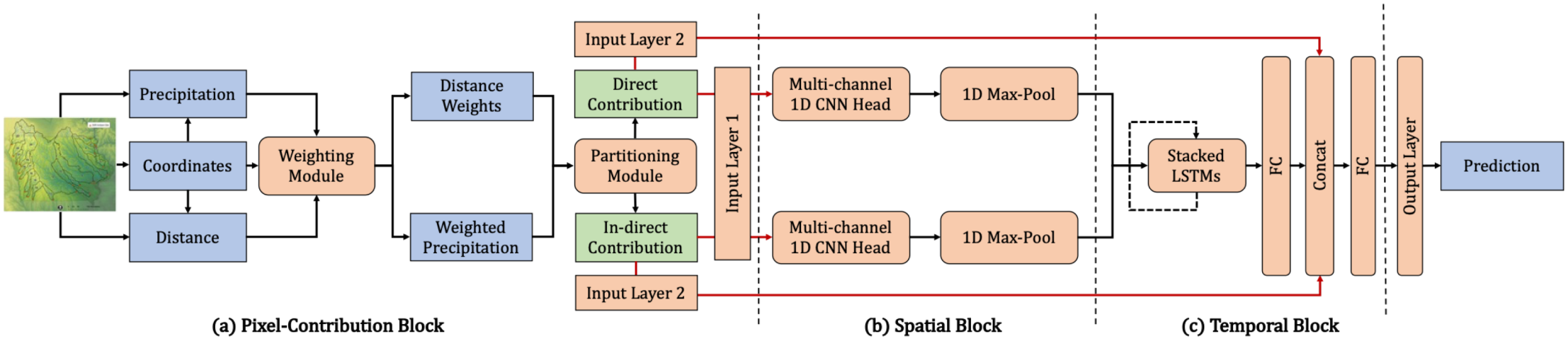
Introducing Dom-ST

Domain-Aware? Spatiotemporal Network

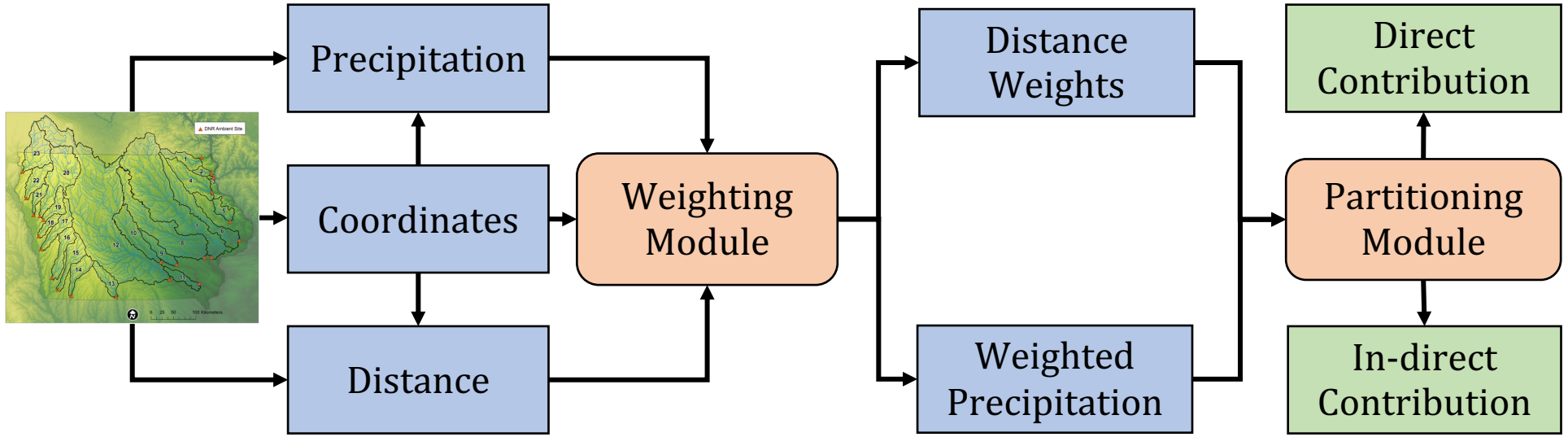
Domain-aware Deep Learning

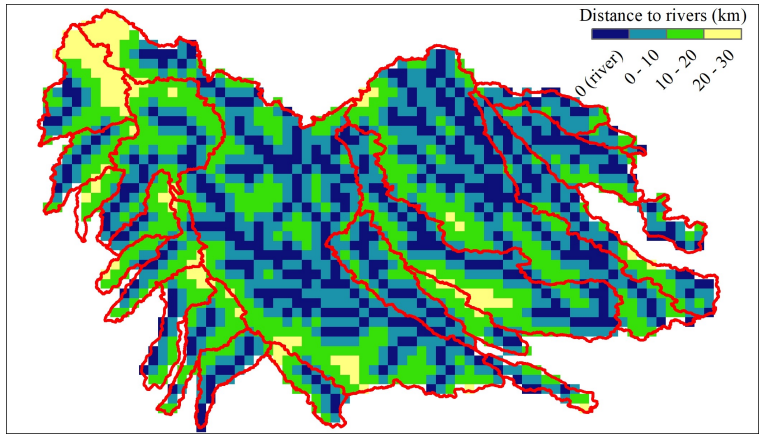


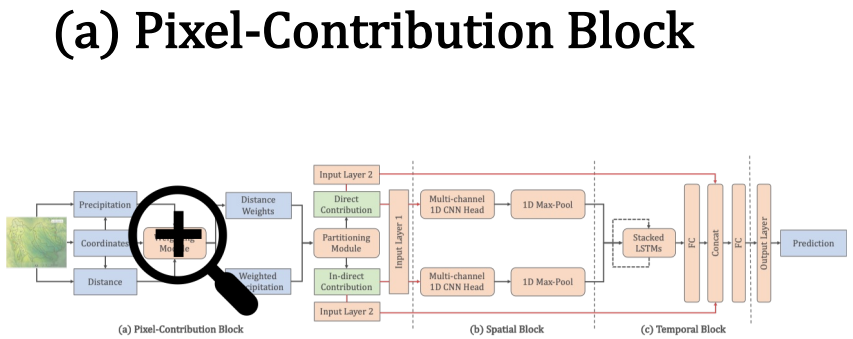
Network Architecture of Dom-ST

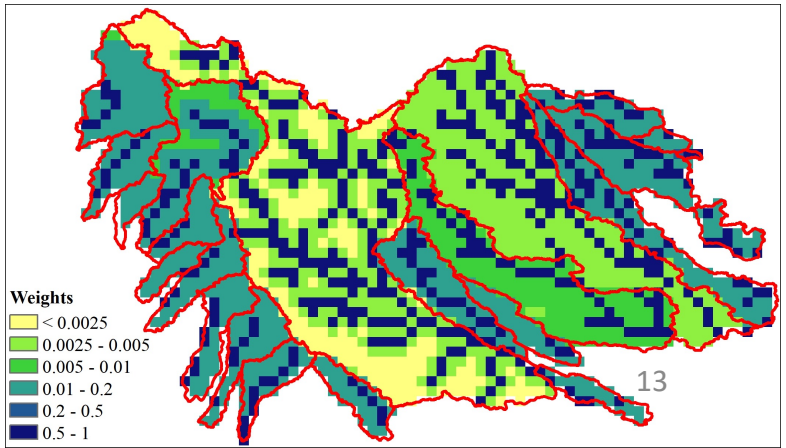


Pixel Contribution Block

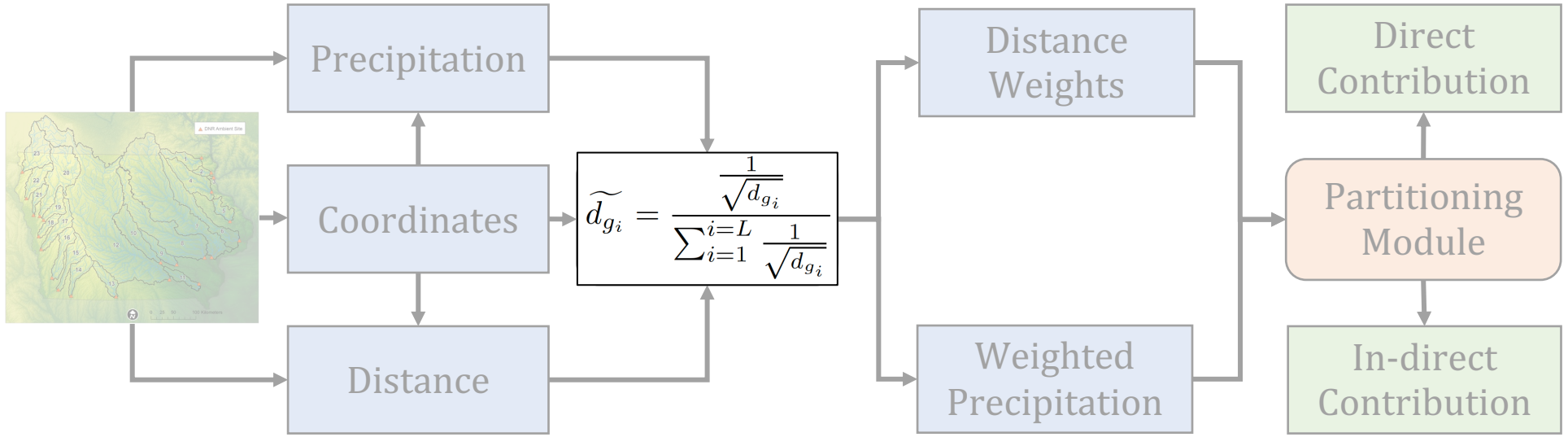


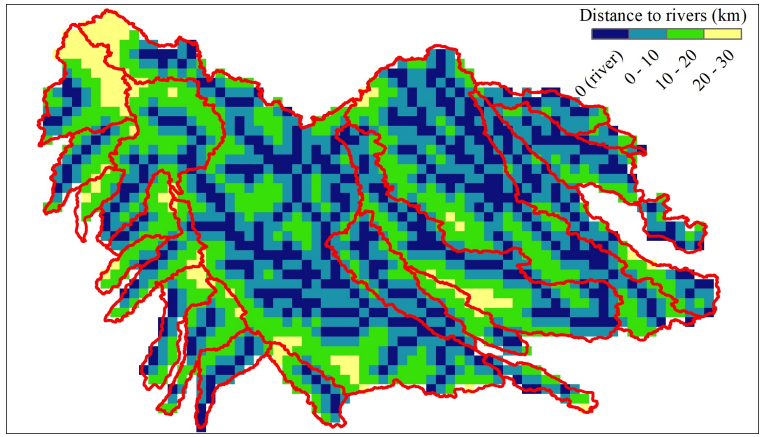


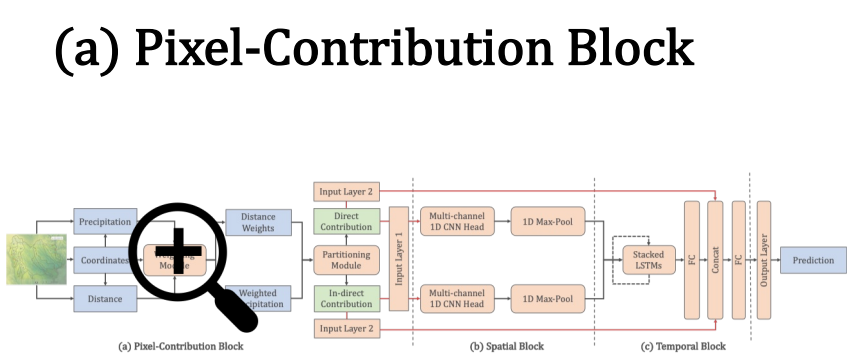


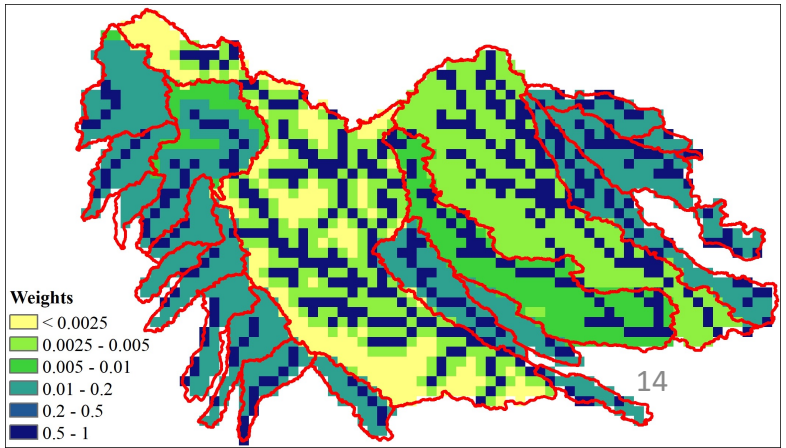


Pixel Contribution Block

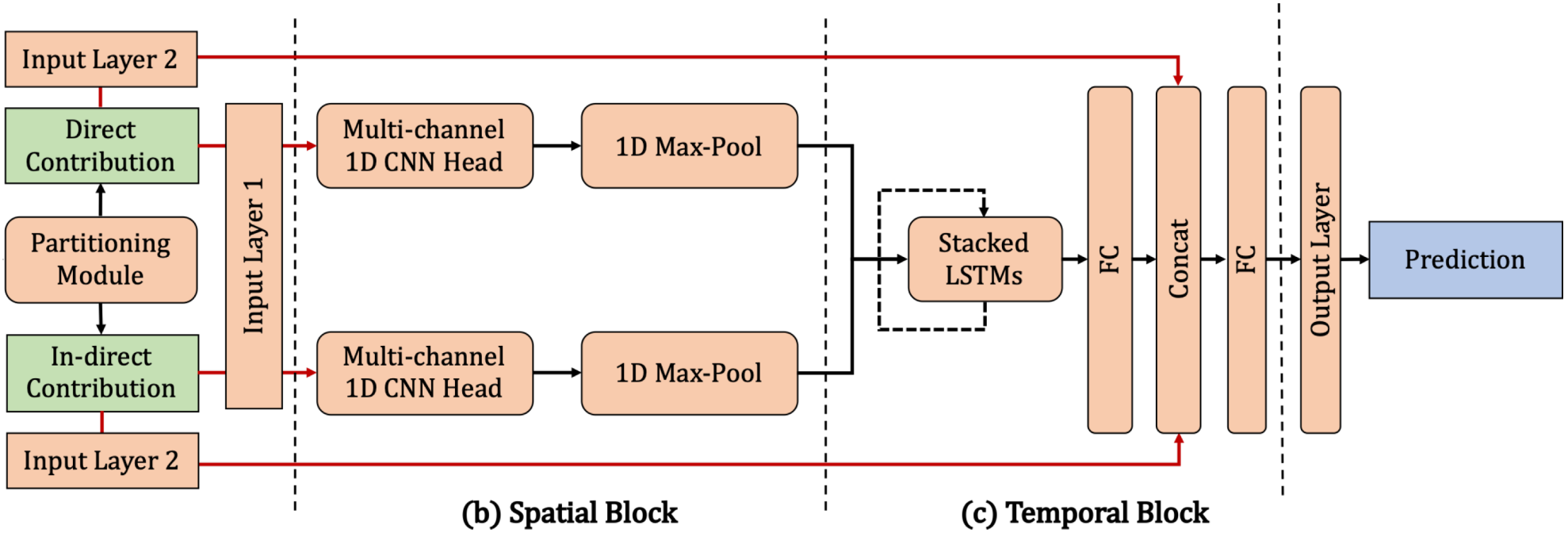


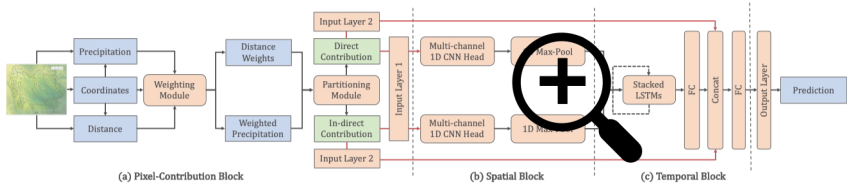




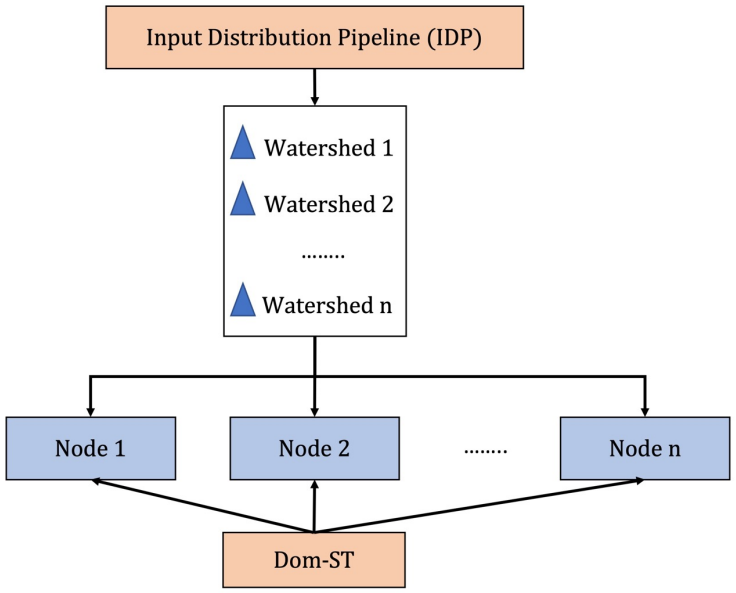


Multi-head Multi-channel CNN-LSTM

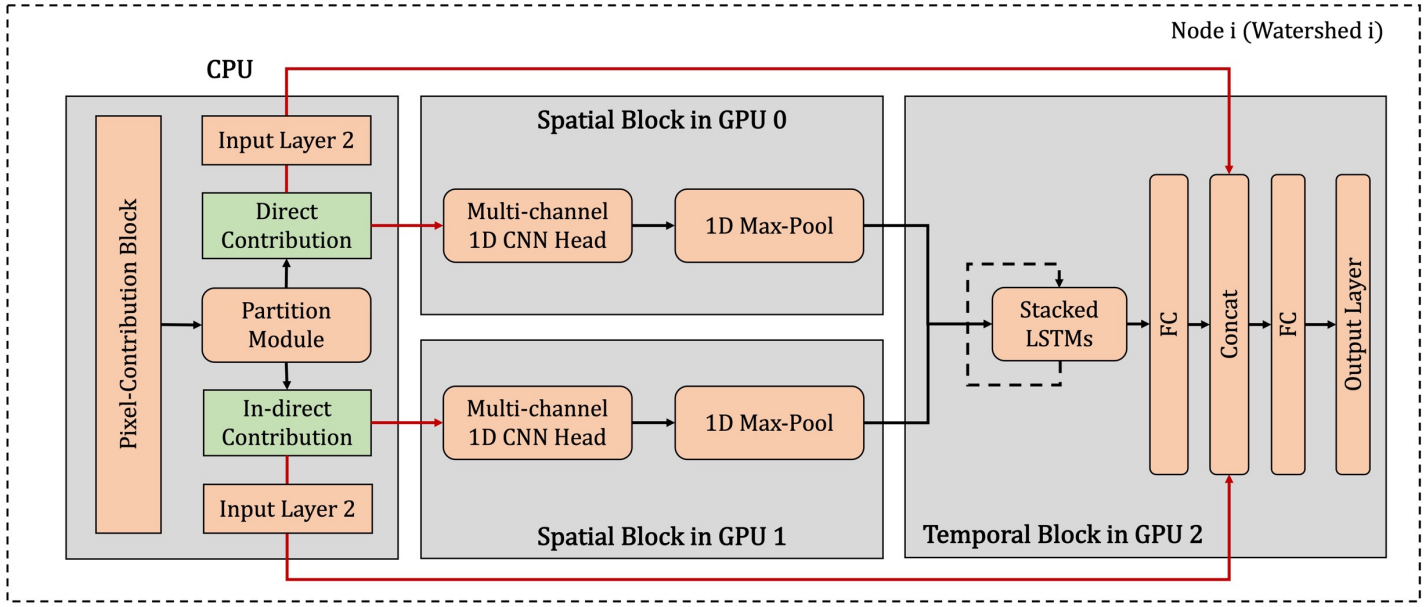




Domain-aware Distribution Strategy

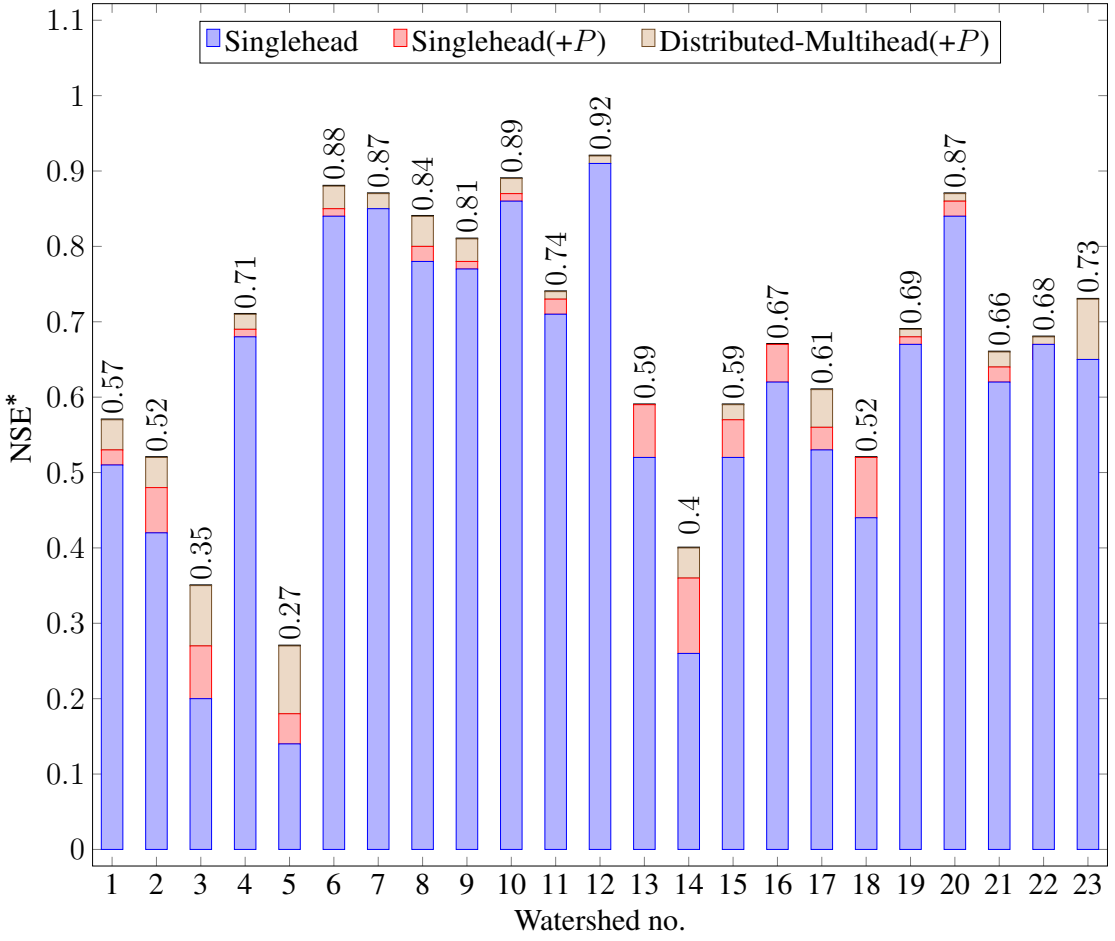


(a) Distributing Data



(b) Distributing Dom-ST

Evaluation



Approach	Time (T_{seq})	Time (T_{IDP})	Speedup
Singlehead(+P)	9.96 h	1.18 h	8.5x
Distributed-Multihead	5.49 h	0.44 h	12.6x

$$Overall\ Speedup = \frac{T_{seq}(Singlehead(+P))}{T_{IDP}(Distributed - Multihead)}$$

= 22.7x

*Nash-Sutcliffe Efficiency

Limitations and Future Work

- Need more climate data at high-frequency
- More domain-awareness
- More advanced distribution strategies
  - Introduce Mixed Precisions
  - Introduce Pipelining during training
  - Mitigate load balancing issues in IDP

Conclusion

- A novel distributed training approach to accelerate a domain-aware spatiotemporal network
- Achieves an overall speedup of 22.7x in our study region
- The highest increase in individual NSE has been 93%
- Highest individual watershed speedup of 4.11x

Thank You!