Multi-Scale Spatiotemporal Mining of Atmospheric Data

Amy McGovern
Schools of Computer Science and Meteorology
amcgovern@ou.edu

Why?

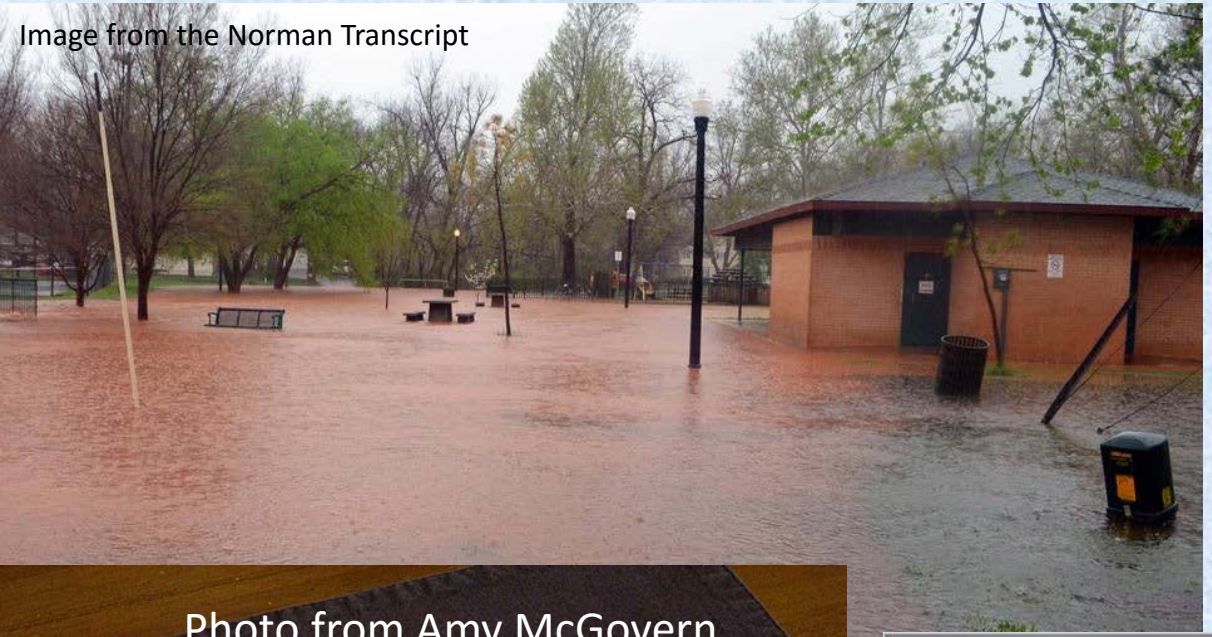


Image from the Norman Transcript


©2010 John R. Leeman

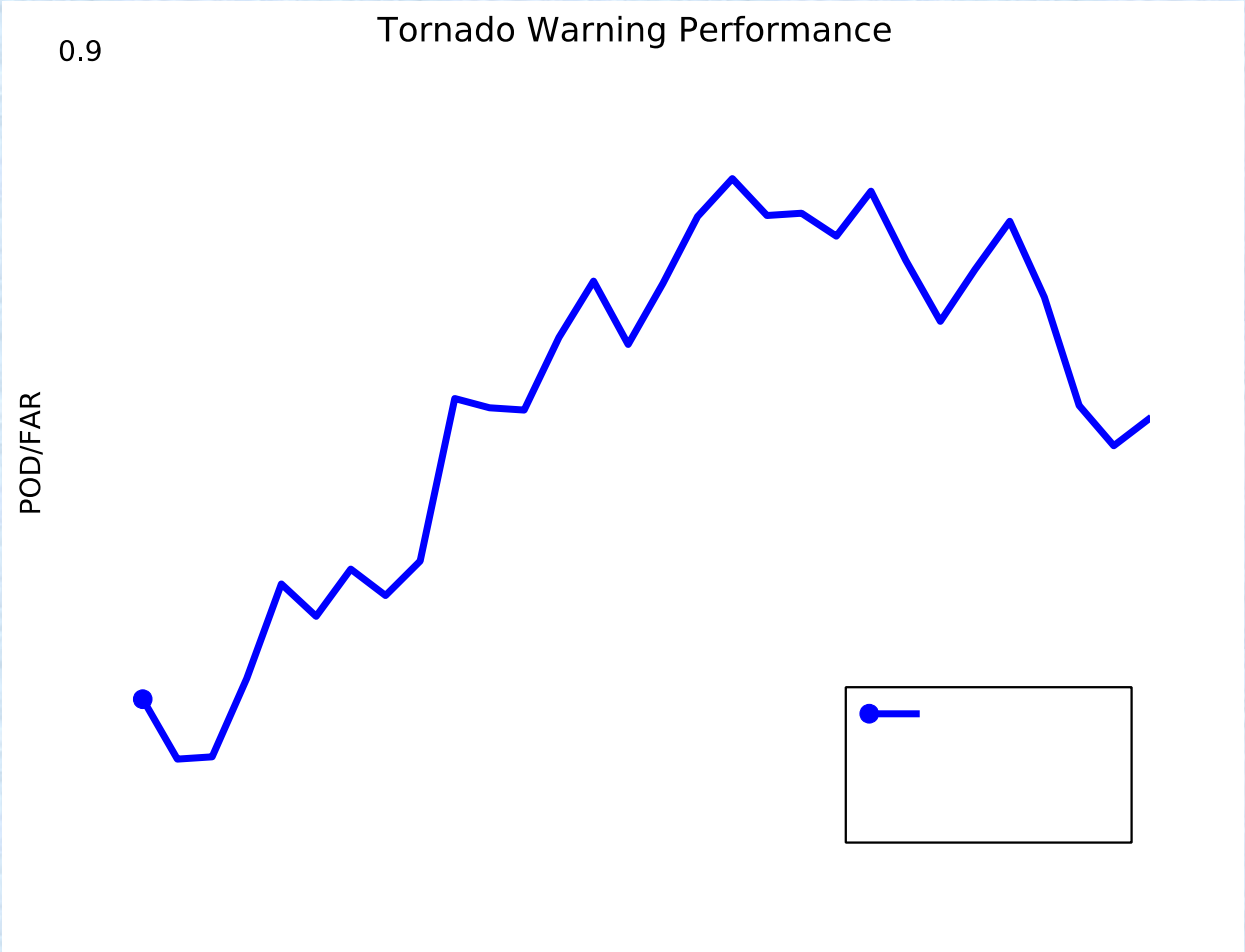

Photo from Amy McGovern


Image Courtesy: New York Times

Goals

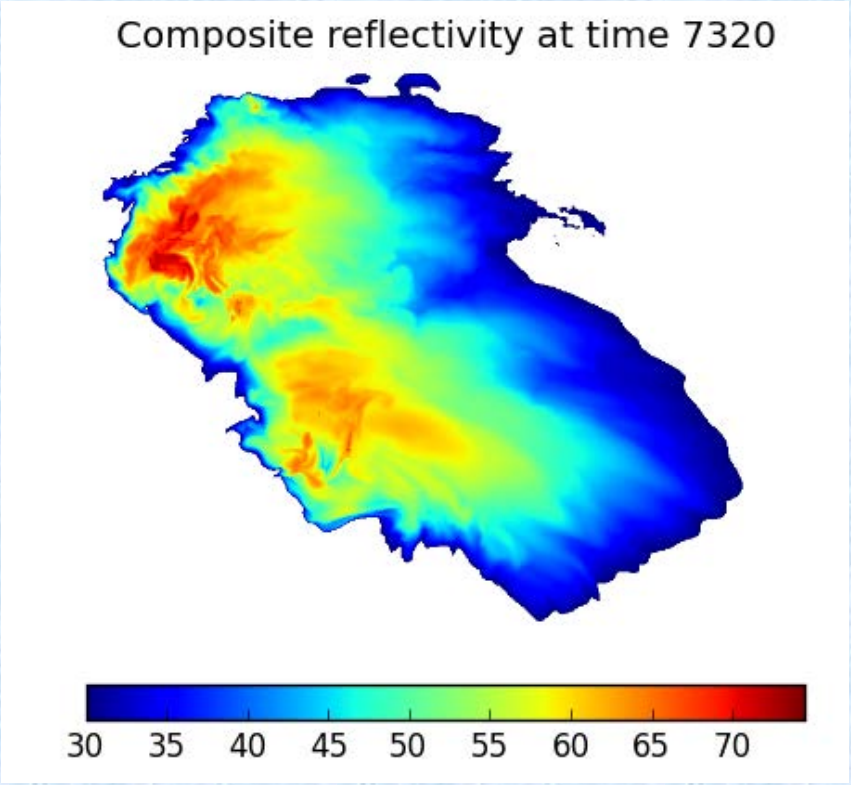
1. Fundamentally transform our understanding of the causes of high-impact weather
   - **Tornadoes**
   - Turbulence
   - Polar vortices
   - Hail
   - Severe wind
   - Thunderstorm initiation
   - Rainfall (floods, drought)
2. Improve the prediction and modeling of these events

Tornado prediction performance



Big weather data: Tornadoes

- 100 m horizontal resolution in center of nested grid
- Stretched vertical grid
- 1536 × 1536 × 99 ≈ 234 million model grid points
- Data saved at every grid point every 30 seconds of simulated time



Data Intensive Computing: Tornadoes

- Simulations
  - Run the simulation
  - Post-process the simulation
- Data mining and visualization
  - Visualizations/Data verification
  - Vortex identification and extraction
  - Object identification and extraction
  - Generating metadata for data mining
  - Data mining

Simulation details

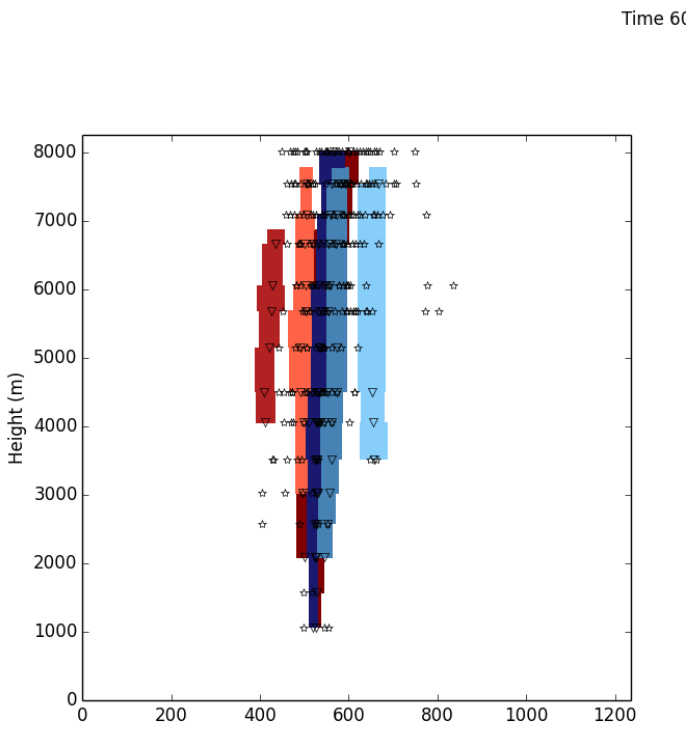
- Run the simulations:
  - Goal of 100 simulations
  - Requires a supercomputer!
    - 4.5 hours on 3600 cores (darter)
    - 6-8 hours on 1800 cores (schooner)
- Post-process the simulations
  - 1 hour, 128 cores (darter)
  - 1-2 hours, 4 cores (schooner)

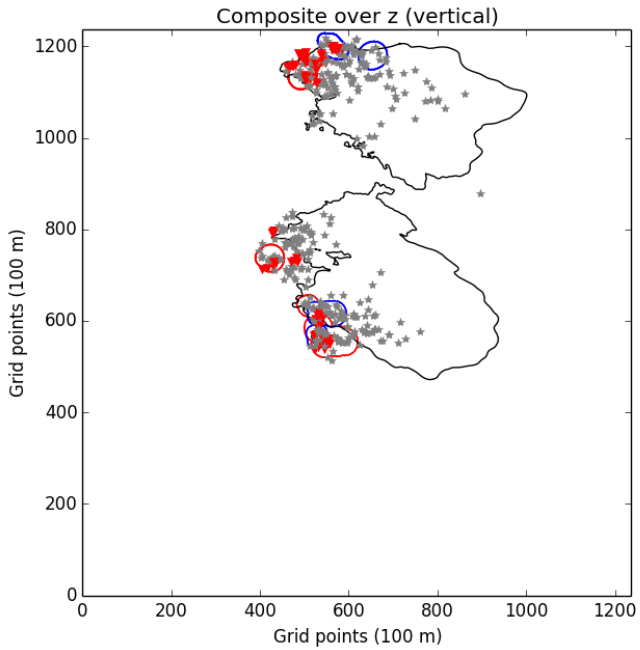
Data Intensive Computing: Tornadoes

- Simulations
  - Run the simulation
  - Post-process the simulation
- Data mining and visualization
  - Visualizations/Data verification
  - **Vortex identification and extraction**
  - Object identification and extraction
  - Generating metadata for data mining
  - Data mining

Vortex identification and extraction

- Purpose: Identify the vortices to find the tornadoes and mesocylones
- Computational challenges:
  - Each process takes ~15 minutes
    - Primarily mathematical (gradient descent)
  - 17 vertical levels x 300 time steps x 3 parameter settings = 15300 independent jobs
  - IO bound:
    - Memory footprint fairly minimal
    - Reads in a LOT of netcdf files and outputs a LOT of small files





OSCER's solutions

- Vortex identification and extraction went from ~1 week of running time to a few hours
  - Consultation with Henry identified key inefficiencies
- Henry has identified key issues to fix in other steps as well
  - In process of being implemented as we transition to schooner
- OSCER helped my students and I learn about HPC
  - Our simulations outgrew boomer so we looked to XSEDE
  - We are moving back to schooner to better share data within OU
  - OSCER continued to help us even when our work was on XSEDE machines

Research Outcomes

- Created unique set of supercell simulations
  - 252 storms at a 500 m horizontal resolution
  - ~60 storms (in progress) at 100 m horizontal resolution
  - Shared with researchers at OU and outside of OU
- Developed and applied unique spatiotemporal data mining methods to a wide variety of high-impact weather phenomena
  - NSF funded projects: CAREER, REU supplements
  - NSF pending projects: BIGDATA
  - Publications: 17 publications facilitated by HPC knowledge that wouldn't be possible without OSCER's help

Research Outcomes: Students

- OSCER has helped to train many of my MS and PhD students!
  - Now located at:
    - Amazon
    - National Center for Atmospheric Research (NCAR)
    - National Hurricane Center
    - Seagate
    - And more