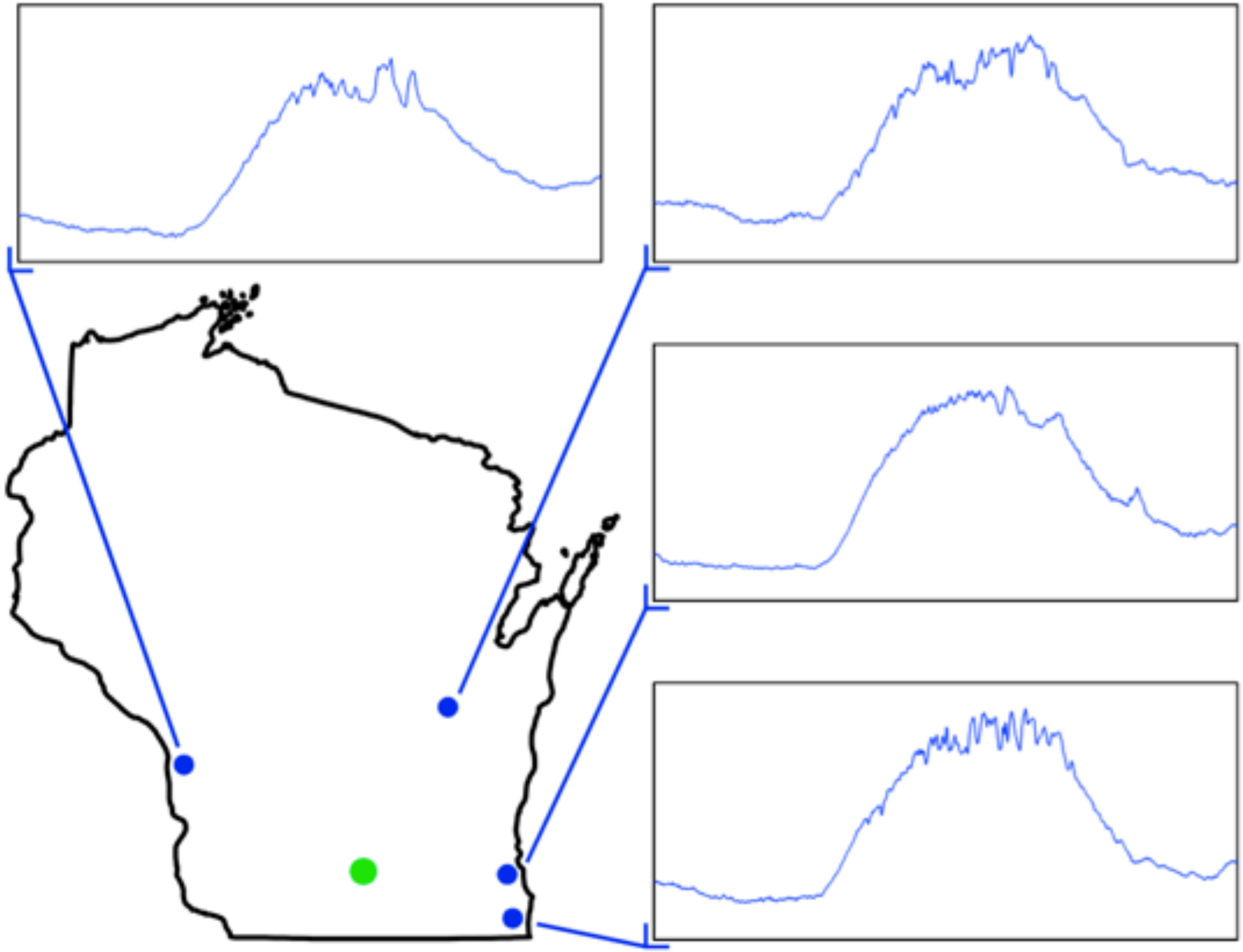
Nonstationary modeling of space-time processes

Joe Guinness

University of Chicago

Department of Statistics



A) Project Overview

- Modeling and interpolation of environmental data (temperature)
- Finding reasonable models for high frequency (1 minute increment) data
- Participants: Myself and my advisor, Michael Stein
- Finished a project with univariate data
- Sponsor: Rao Kotamarthi
- Goal: Develop computationally feasible methods for fitting space-time models to environmental data

B) Science Lesson

- We have a probabilistic model for the temperature process in time and space that depends on a number of parameters.
- Using the temperature records, we can estimate the parameters of the model via maximum likelihood
- Then we can provide predictions of what the temperature might have been at other points in space, as well as uncertainties for those predictions

C) Parallel Programming Model

- Don’t really have one yet!
- However, I think there are some obvious things that can be done
- Currently, all of our code for our likelihood calculations is in matlab

D) Computational Methods

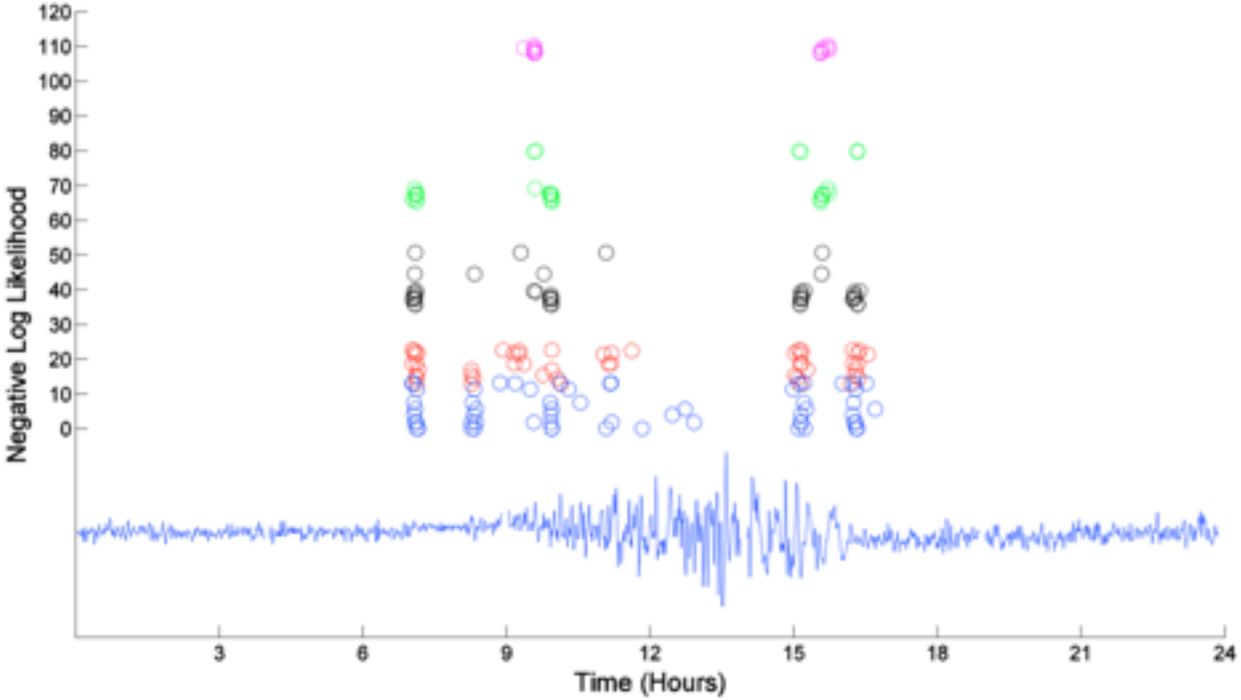
- Our likelihood approximation makes use of the FFT whenever possible
- The optimization procedure consists of a continuous optimization nested inside of a discrete optimization
- The discrete optimization is much harder and much slower

E) I/O Patterns and Strategy

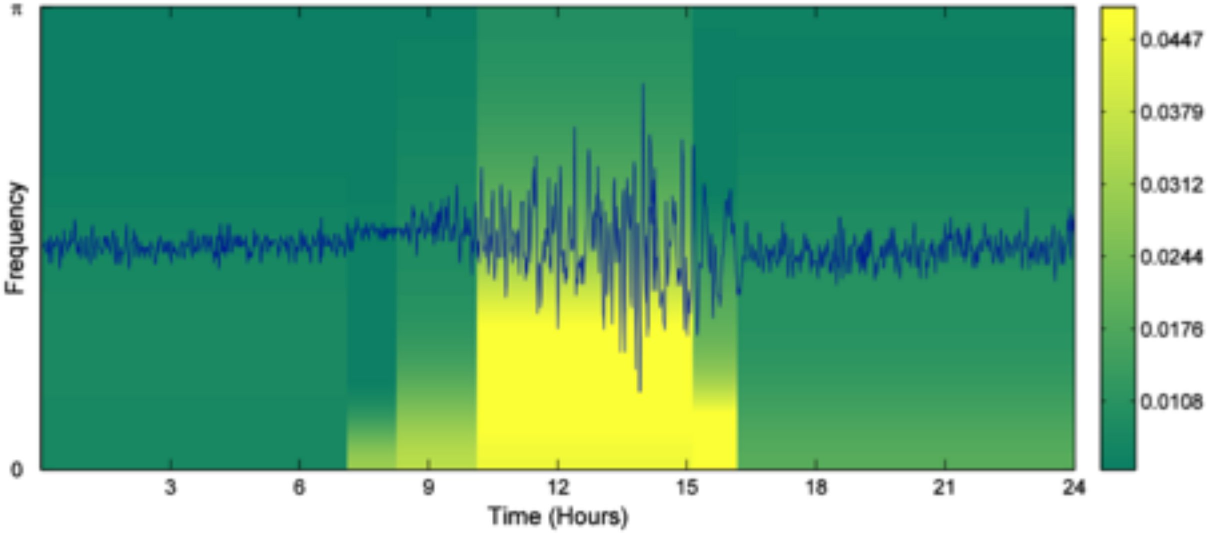
- Input I/O and output I/O patterns (one file per MPI process?, pNetCDF? HDF5?, etc) - ???
- Size of data: Roughly 20 sites at which data are collected, 1440 observations per day, per site, and we’re looking at several days now, but would like to do more
- Outputs could be much larger, depending on resolution of interpolation in space
- Checkpoint / Restart capabilities: what does it look like? ???
- Current status and future plans for I/O ???

F) Visualization and Analysis

- Two plots on the next page as examples
- The visualization problems are really interesting to us, so we’re hoping to learn more



Negative Log Likelihood
120
110
100
90
80
70
60
50
40
30
20
10
0
3
6
9
12
15
18
21
24
Time (Hours)



π
Frequency
0
3
6
9
12
15
18
21
24
Time (Hours)
0.0447
0.0379
0.0312
0.0244
0.0176
0.0108

G) Performance

- What tools do you use now to explore performance (Tau, DynInst, PAPI, etc) ???
- Slowest part of the optimization is the discrete optimization (finding the best partition)
- Current bottleneck to better scaling is that I need to learn more about parallel processing!
- What features would you like to see in perf tools (ease of instrumentation, different measurements, embedded vis, etc) ???
- Future plans are to parallelize as much as is possible and useful

H) Tools

- Debugging is done in the matlab user interface

I) Status and Scalability

- In the univariate case (one site), I think the calculations would scale well to longer time intervals, but adding more sites is more difficult.
- Would like to finish the multivariate case in a year, complete with parallelization
- What are your top 5 pains? (be specific)
  - Not knowing how to implement parallelization
  - We have some knowledge on where it can be useful but would like to improve understanding
- Current scalability achieved by tailoring our model so that the likelihood calculation makes use of the FFT
- Future plans: parallelize!

J) Roadmap

- For statisticians, serious space time modeling is pretty new, so we're hoping to show that that it can be done and that the computation problem is feasible for large datasets
- We are not sure what scientific questions or answers will come out of this.
  - We have a "hunch" that environmental processes are not well understood on short time scales – minute-by-minute or shorter variations.