Statistical Models and Computing

Statistical Methods for Solving Scientific Problems in Disease Modeling and Environmental Science

Murali Haran

Associate Professor
Department of Statistics
Penn State University

Penn State Statistics Club, February 2011

What do I work on?

1. Monte Carlo methods.
2. Models for spatial data.
3. Complex computer models.
4. Lots of scientific problems ("interdisciplinary research") that use all of the above. Working on interesting scientific problems invariably involves solving challenging statistical problems, new statistical research.

[This is what I do when I am not working on teaching-related or administrative duties.]

Interdisciplinary research problems

"The best thing about being a statistician is that you get to play in everyone's backyard." — John Tukey.

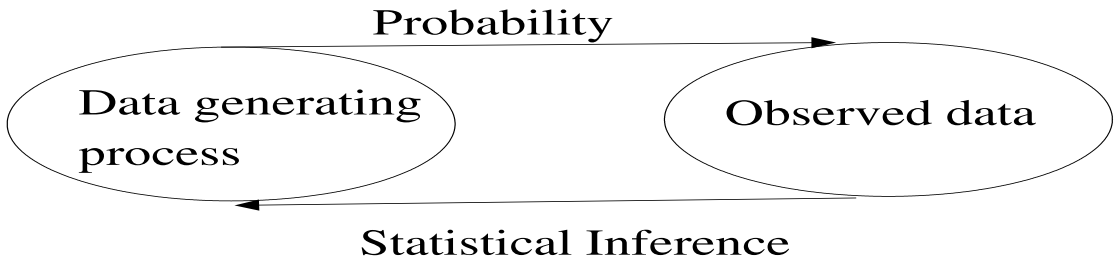
I work with many smart Penn State scientists:

1. Infectious disease models: studying the dynamics of disease transmission. Collaborations with biologists, disease modelers.
2. Climate science: using climate (computer) models and data to learn about climate change and predicting future climate. Collaborations with geoscientists.
3. Environmental statistics: crop epidemics, studying how ecological factors can impact coastal aquatic ecosystems. Collaborations with ecologists, geographers, plant pathologists.

Reminder: probability and statistical inference

Scientific research based on data, whether from experimental studies or observational studies, can be summarized as follows:

Given the data (what we have observed), what can we infer about the process that generated the data?



This Talk

- I will now describe some problems that I work on and try to give you a rough idea of the role statistics and computing have to play in solving them.
- This work involves a collaboration between me, other scientists at Penn State, and Ph.D. students in the statistics department.

Models for Infectious Disease Dynamics

- Studying disease dynamics is of fundamental scientific importance, e.g. in studying population and evolutionary biology.
- Infectious diseases have an immense impact on human health. e.g. measles, dengue, AIDS, malaria.
- Diseases that affect plants and animals can seriously impact agriculture and conservation.
- Understanding disease dynamics helps in management, vaccination, epidemic control strategies.

Disease dynamics models can be very useful for both basic science as well as for managing/controlling diseases.

What is a mathematical model?

- A mathematical construct used to help us better understand real world systems.
- Simple example: physics models describing how changing the temperature but keeping the pressure constant will change the volume of a gas over time.
- Obviously not possible to fully describe the real world *perfectly* via mathematical equations, so mathematical models are an approximation to reality.

What is a statistical model?

- A mathematical model which now also allows for *stochasticity*, a fancy word for randomness.
- By allowing for stochasticity, statistical models can often characterize the real world in a more realistic fashion.

SIR models

Basic SIR models classify individuals as one of susceptible (S), infected (I) or recovered (R).

- Individuals are born into the susceptible class.
- Susceptible individuals have never come into contact with the disease and are able to catch the disease, after which they move into the infected class.
- Infected individuals spread the disease to susceptibles, and remain in the infected class (the infected period) before moving into the recovered class.
- Individuals in the recovered class are assumed to be immune for life.

Gravity T-SIR model

- Model extends the discrete time-series SIR (T-SIR) model (Bjornstad et al.2002; Grenfell et al. 2002) with explicit formulation of the spatial transmission between different host communities.
- Notation:
  - $I_{k,t}$ - number of infected individuals in city k at time t.
  - $S_{k,t}$ - number of susceptible individuals in city k at time t.
  - $d_{k,j}$ - distance between cities k and j.
  - $N_{k,t}$ - population of city k at time t.
  - $B_{k,t}$ - local number of new hosts (births) in city k at time t.
  - $L_{k,t}$ - number of infected people moved to city k at time t.
  - T cities, K time points.

Modeling incidences

Following Xia, Bjornstad and Grenfell (2004):

- Number of incidences of a disease at time $t+1$ for city k,

  $I_{k,t+1} = \text{Poisson}(\lambda_{k,t+1})$, where $\lambda_{k,t+1} = \beta_t S_{k,t}(I_{k,t} + L_{k,t})^{\alpha}$.

- $\{\beta_t\}$ specified only via 26 parameters (26 = number of biweeks in a year), to allow for differences in seasonal transmission. Assumed to be same every year.
- $\alpha, \{\beta_t\}$ are local transition parameters.

Modeling susceptibles

- Number of susceptible individuals at time $t+1$ for city k is then modeled via balance equation (Bartlett, 1957):

$$S_{k,t+1} = S_{k,t} + B_{k,t} - I_{k,t+1}$$

- Finally, unobserved number of infected immigrants moved to city k at time t is modeled as:

$$L_{k,t} = \text{Gamma}(m_{k,t}, 1),$$

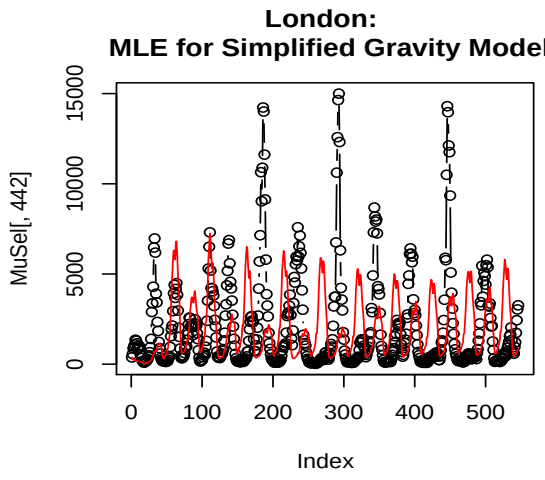
where

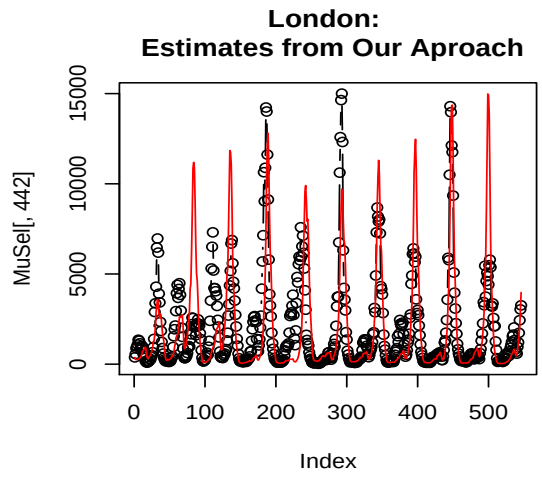
$$m_{k,t} = \theta N_{k,t}^{\tau_1} \sum_{j=1, j\neq k}^{K} \frac{(I_{jt})^{\tau_2}}{d_{k,j}^{\rho}}, \quad \theta, \tau_1, \tau_2, \rho > 0.$$

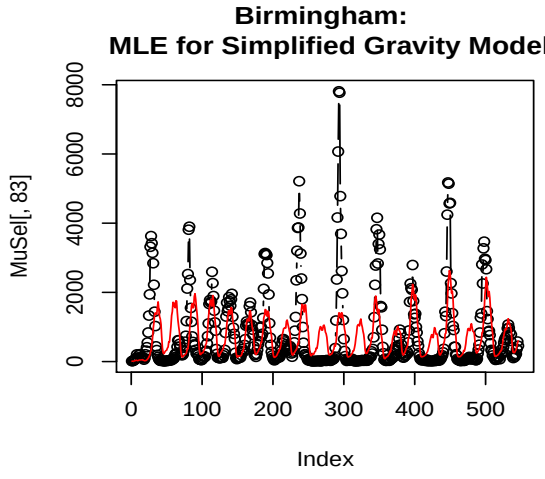
Statistical inference

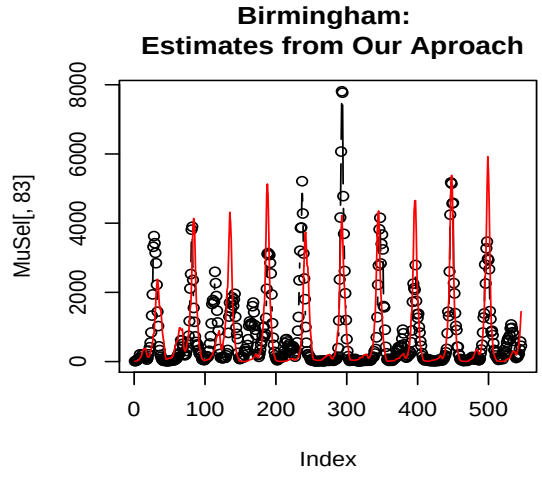
- **Measles data**
  - The UK Registrar General's data for 952 cities in England and Wales for years 1944-1966 of biweekly incidences of measles. Very rich spatio-temporal data.
  - Data for number of susceptibles from standard susceptible reconstruction algorithms (Fine and Clarkson 1982a, Schenzle 1984, Ellner et al. 1998, Bobashev et al. 2000, Finkenstadt and Grenfell 2000).
- **Parameters of the model**:
  - Reliable estimates of local transition parameters α and $\{\beta_t\}$ are assumed known from previous work (Bjornstad et al. 2001).
  - Gravity parameters $\theta, \tau_1, \tau_2, \rho$ are unknown.
- **Goal**: Infer unknown gravity parameters.

Data Example: London, Birmingham

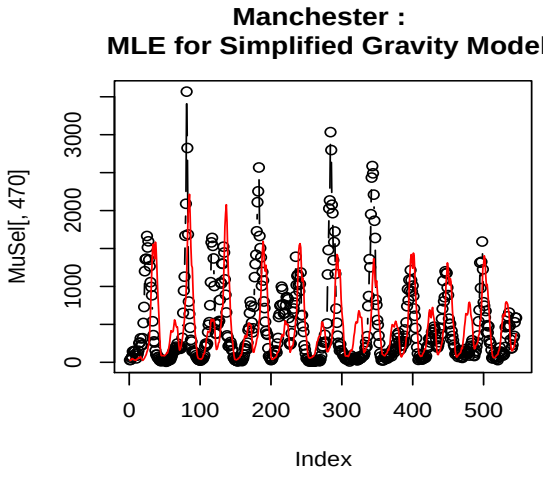


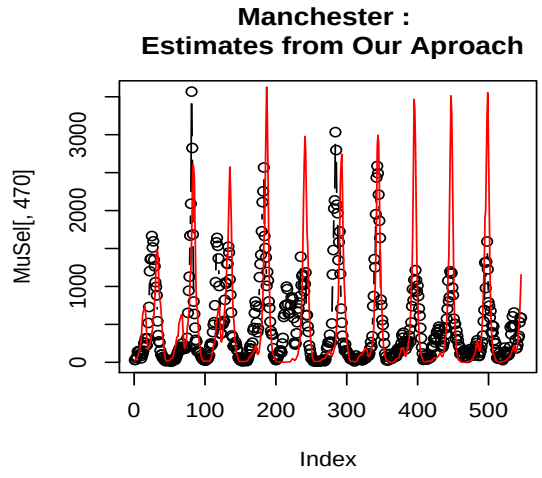


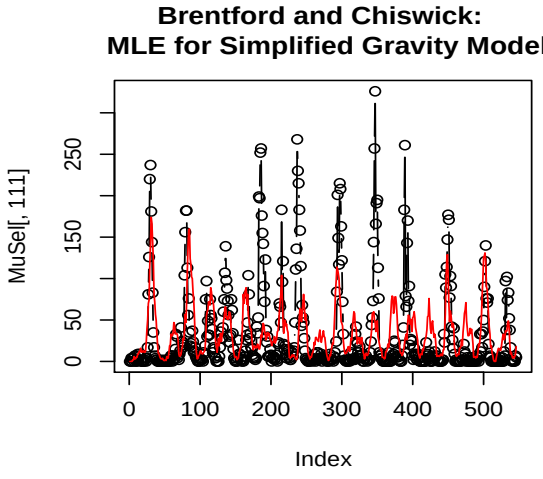


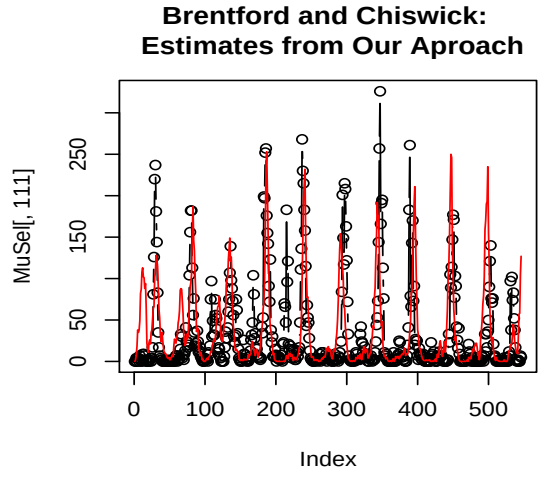


Data Example: Manchester, Brentford/Chiswick

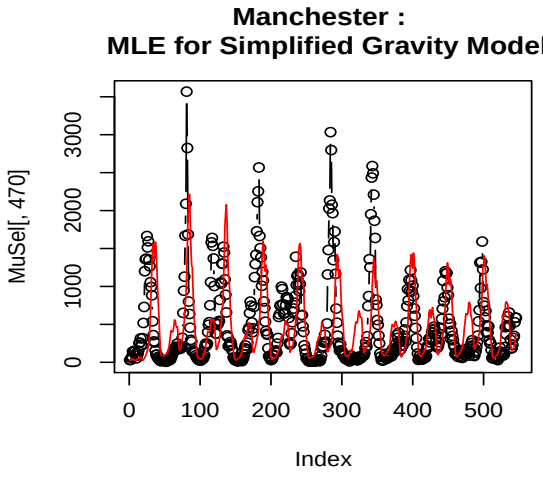


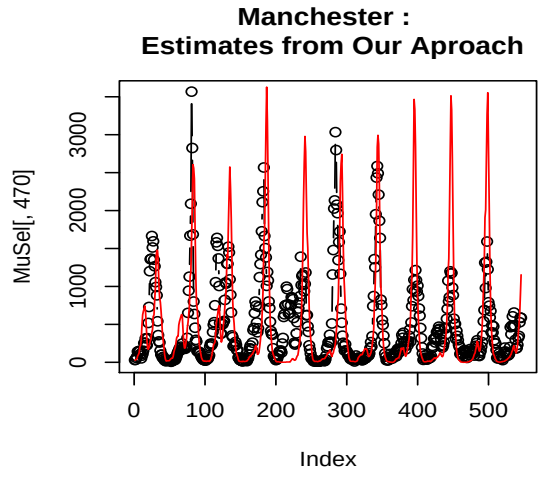


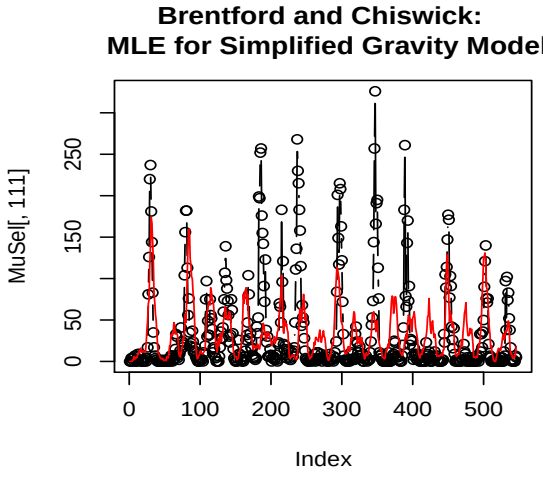


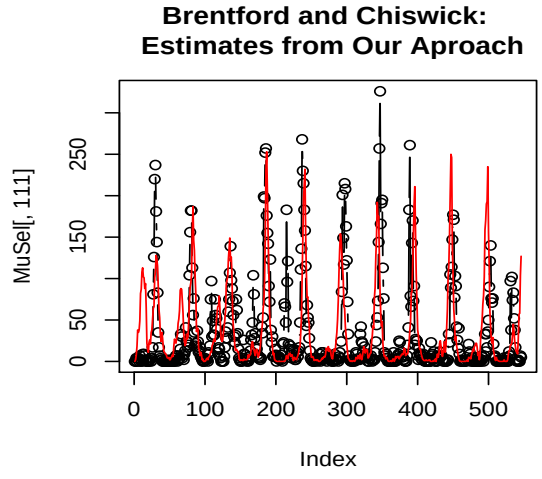


Data Example: Manchester, Brentford/Chiswick









What Now?

- We now have a statistical model describing the spread of infectious disease + data on the actual spread of infectious diseases.
- We can use statistical inference techniques:
  - Maximum likelihood
  - Bayes

  This allows us to estimate the parameters of the model and to see whether the model "fits" the data well. If it does, the model can be used for science and disease management.

Statistics and Computing

- Fitting the models is very challenging.
- Studying how the model behaves is very challenging.
- We need computers as well as sophisticated new statistical approaches to solve these problems.
- Sketch follows...

Interesting Problems: Models for Infectious Disease

Models for computer experiments:

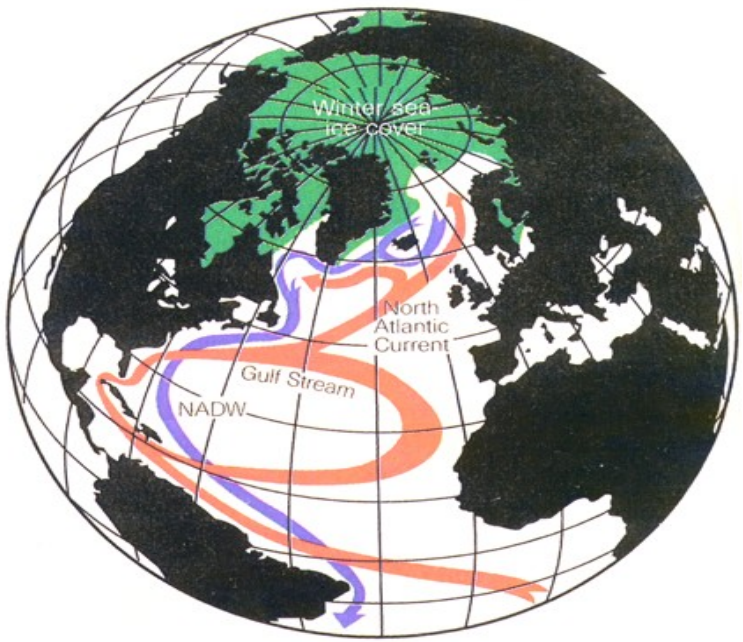
- Complex computer models are often used to mimic reality.
- Examples:
  - General Circulation Models (GCMs) are used to model climate systems.
  - Complicated disease dynamics models are used to model the spread of infectious diseases.
- Given inputs, the model will run and produce output. For e.g. at certain inputs of climate system characteristics, the GCMs will provide temperature fields all over the world over hundreds of years.

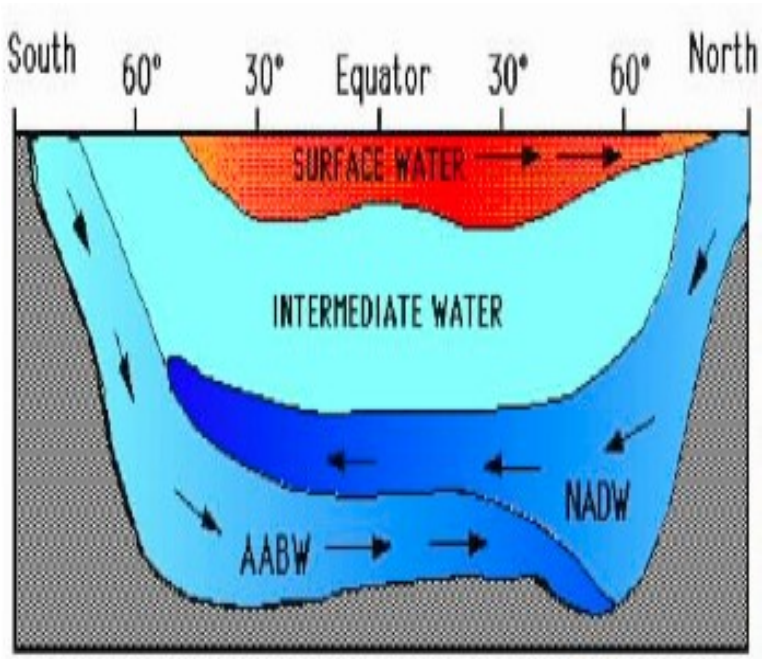
Models for computer experiments

- Complex computer models can be extremely computationally expensive. e.g. GCMs can take months to run. EMICs (Earth system models of intermediate complexity) can take days or weeks to run.
- Cannot see what happens at all interesting input settings.
- If we have observations (of reality), we may want to learn about the computer model inputs most 'compatible' with reality. e.g. we can compare measurements of temperature values across the world to the climate model output to figure out which sets of inputs (climate characteristics) produce output most like reality.
- Parameters may be of interest, and the model + parameter estimates can be used for doing predictions.

Example: climate change research

- What is the risk of human induced climate change?
- Example of climate change: potential collapse of meridional overturning circulation (MOC).
- An MOC collapse may result in drastic changes in temperatures and precipitation patterns.



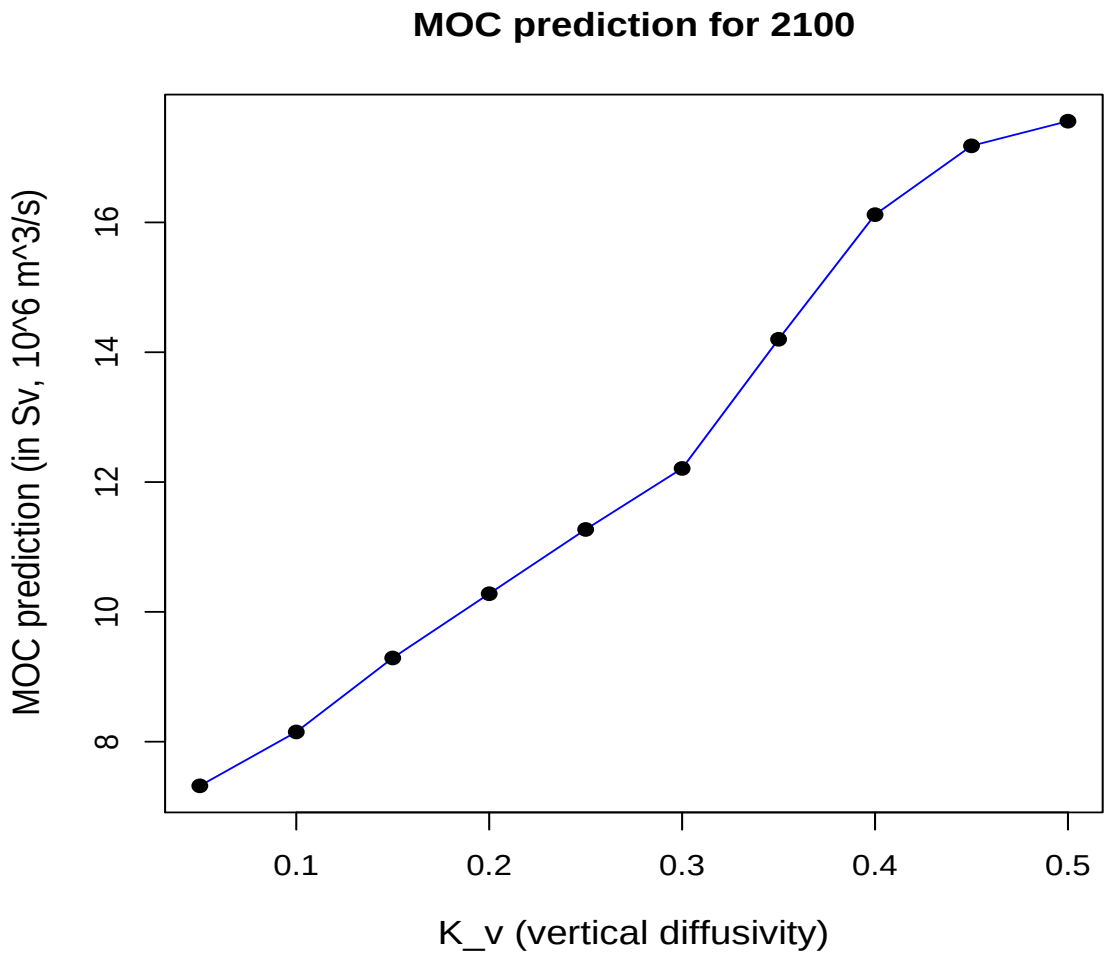


(plots: Rahmstorf (Nature, 1997) and Behl and Hovan)

Motivation-MOC

- MOC phenomenon: Movement of water from equator to higher latitudes, deep water masses created by cooling of water in Atlantic, resulting in sea ice formation. Result is denser salt water, which sinks, causing ocean circulation.
- MOC weakening results in disruptions in the equilibrium state in the climate, may lead to major temperature and precipitation changes and shifts in terrestrial ecosystems.
- Predictions of MOC strength can be made for particular climate parameter settings, e.g. vertical diffusivity, $\mathbf{K_v}$.
- $\mathbf{K_v}$ cannot be measured directly.

MOC predictions versus K_v



- MOC predictions are clearly much lower as K_v values get small. Small $K_v \Rightarrow$ weak MOC.

Learning about the MOC via tracers

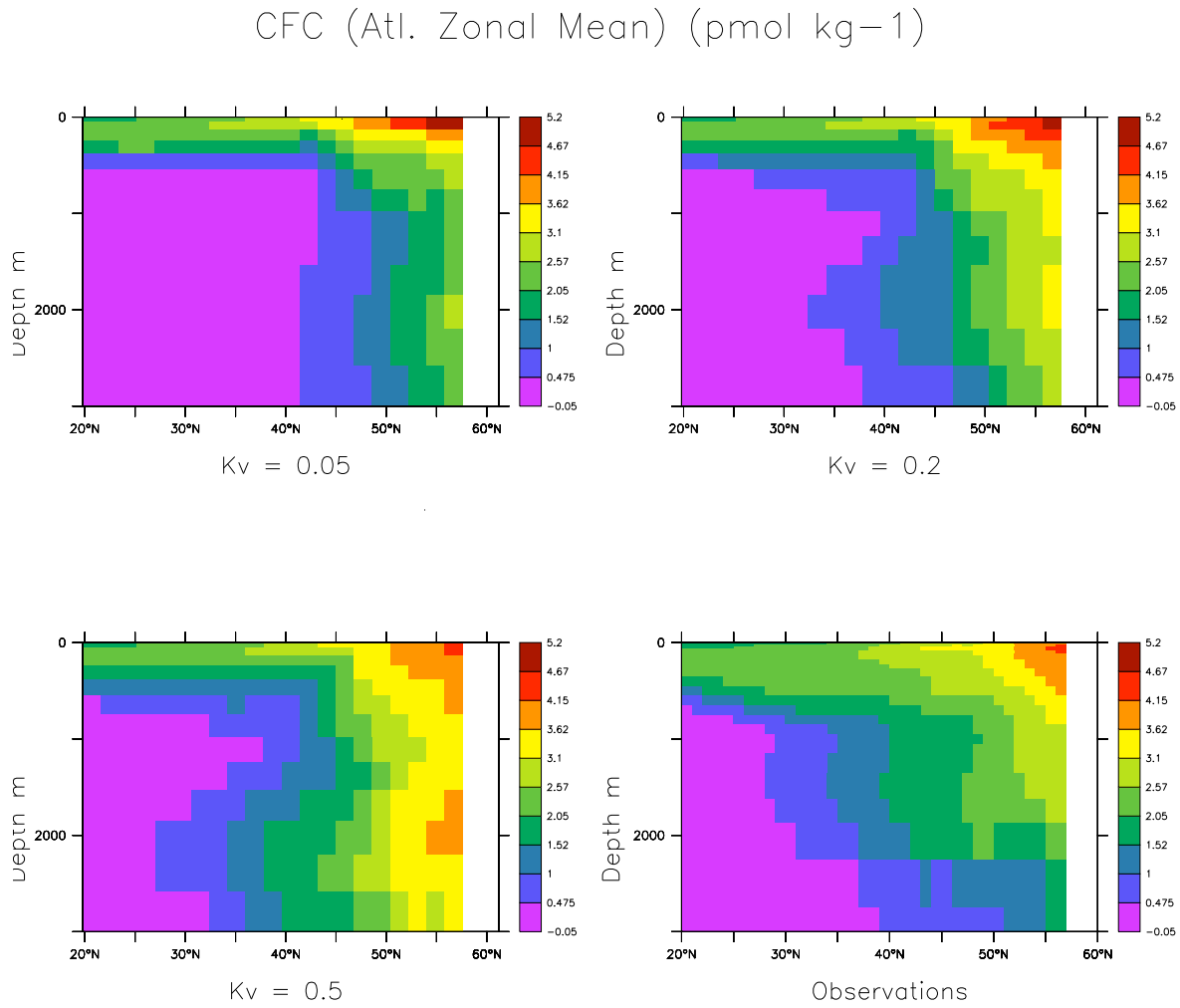
Two sources of indirect information:

- Climate models: output at different parameter settings.
- 'Tracers' of climate parameters: spatio-temporal data.
- Trichlorofluoromethane (CFC11) and Carbon-14 (C-14) are 'stable tracers' for $\mathbf{K_v}$ so we can use data and model runs for CFC11 and C-14 to learn about $\mathbf{K_v}$.
- Observations of both tracers collected across globe in the 1990s, locations consist of latitude, longitude, and depth values, aggregated over longitudes: spatial data.
- Second source of information: climate model output at different values of $\mathbf{K_v}$.
- 3706 observations and 5926$\times$ 6 data points from model.
- Latitude between -80 S and 60 N, depths from 0 to 3000m.

Statistical inference

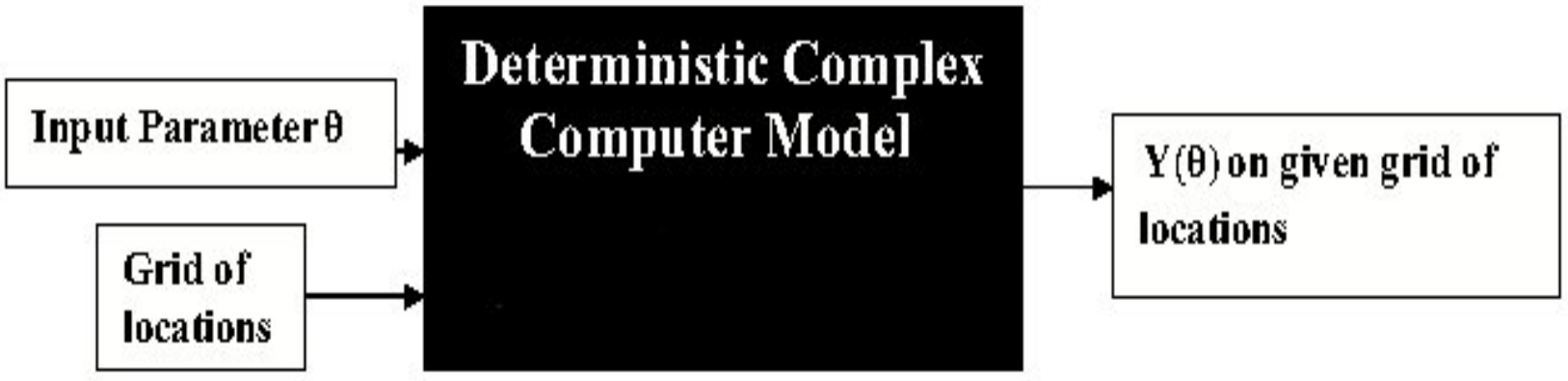
- **Goal**: Infer important climate characteristics (parameters) that drive major climate systems. Focus on $\mathbf{K_v}$.
- Sources of information
  - Physical observations of climate system: spatial data on CFC-11 and C-14. Notation: $Z(\mathbf{s})$, $\mathbf{s}$ =location.
  - Output from complex climate models at several different climate parameters from University of Victoria(UVic) Earth System Climate Model (Weaver et. al. 2001). Notation: $Y(\mathbf{s}, \theta)$, θ= climate parameter.
- Want to learn about $\mathbf{K_v}$ based on Zs and Ys for both CFC-11 and C14.

CFC example



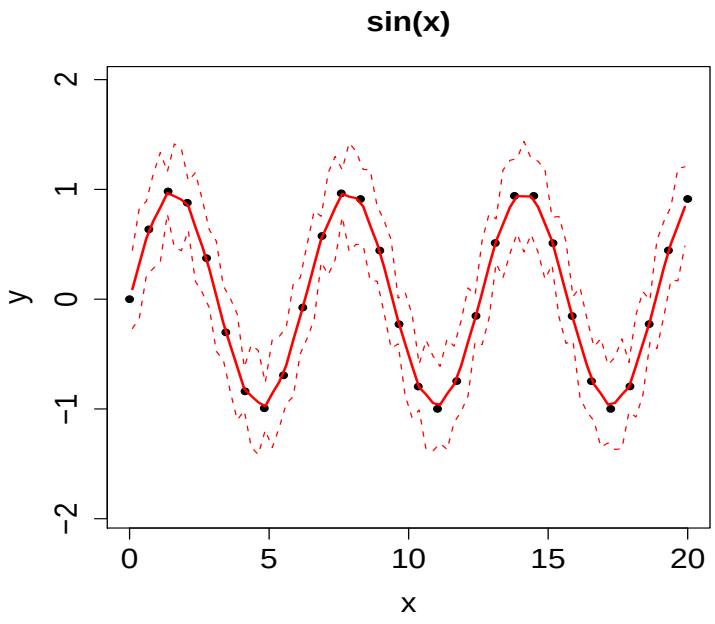
- Bottom right: observations.
- Remaining plots: climate model output at 3 settings of K_v.

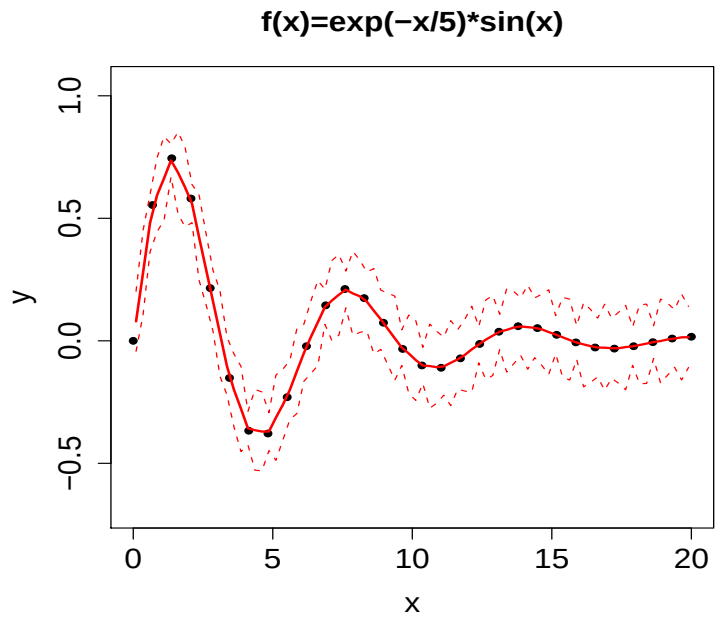
Computer model emulation



- **Emulation** involves replacing a complicated computer model with a simpler approximation.
- Statistical solution: Use Gaussian processes (a statistical model) to 'emulate' the computer model and obtain estimates of the output at input settings we did not try: run the model at several input settings, collect the output, and fit a Gaussian process model to the output.

GP model for emulation: toy examples





Suppose we ran the two toy computer models at 'input' values *x* equally spaced between 0 and 20 to get model output (black dots). Can we predict between black dots?

Pretend we don't know the model (functions). The red curves are interpolations using a simple GP model, along with some idea of uncertainty. Nice fit for both functions!

GP model for emulation: climate model

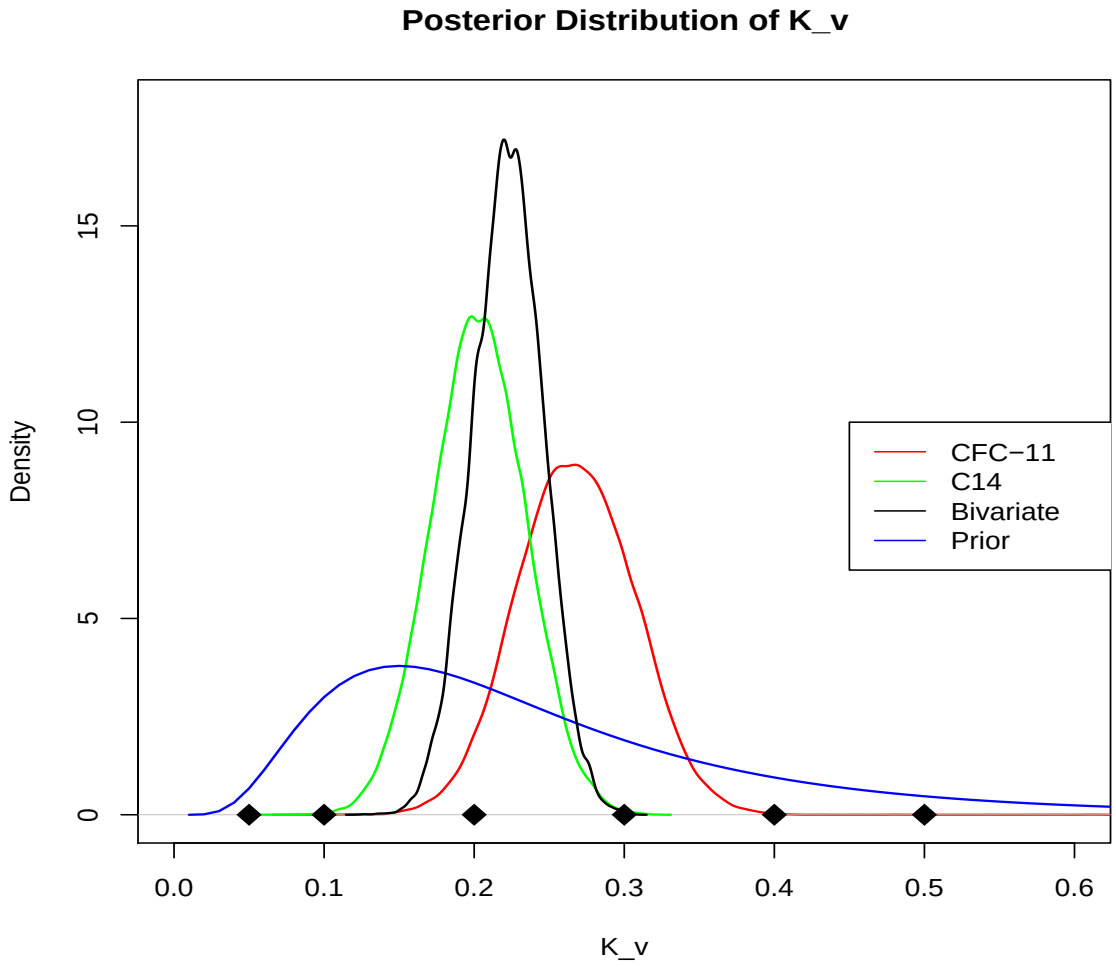
- Unlike the toy example, the output from the climate model is much more complicated — for each $\mathbf{K_v}$ we have a spatial field (not a single point). We fit a (more sophisticated) Gaussian process model to the climate model output. This serves as the emulator for the climate model.
- We can now use this GP model instead of the very complicated climate model — this provides a connection between $\mathbf{K_v}$ and the climate model output, in this case the tracers CFC-11 and C-14.
- Model for the observed CFC-11 and C-14: can use the GP model + allow for additional sources of uncertainty and bias.

K_v inference

Here's what we have done:

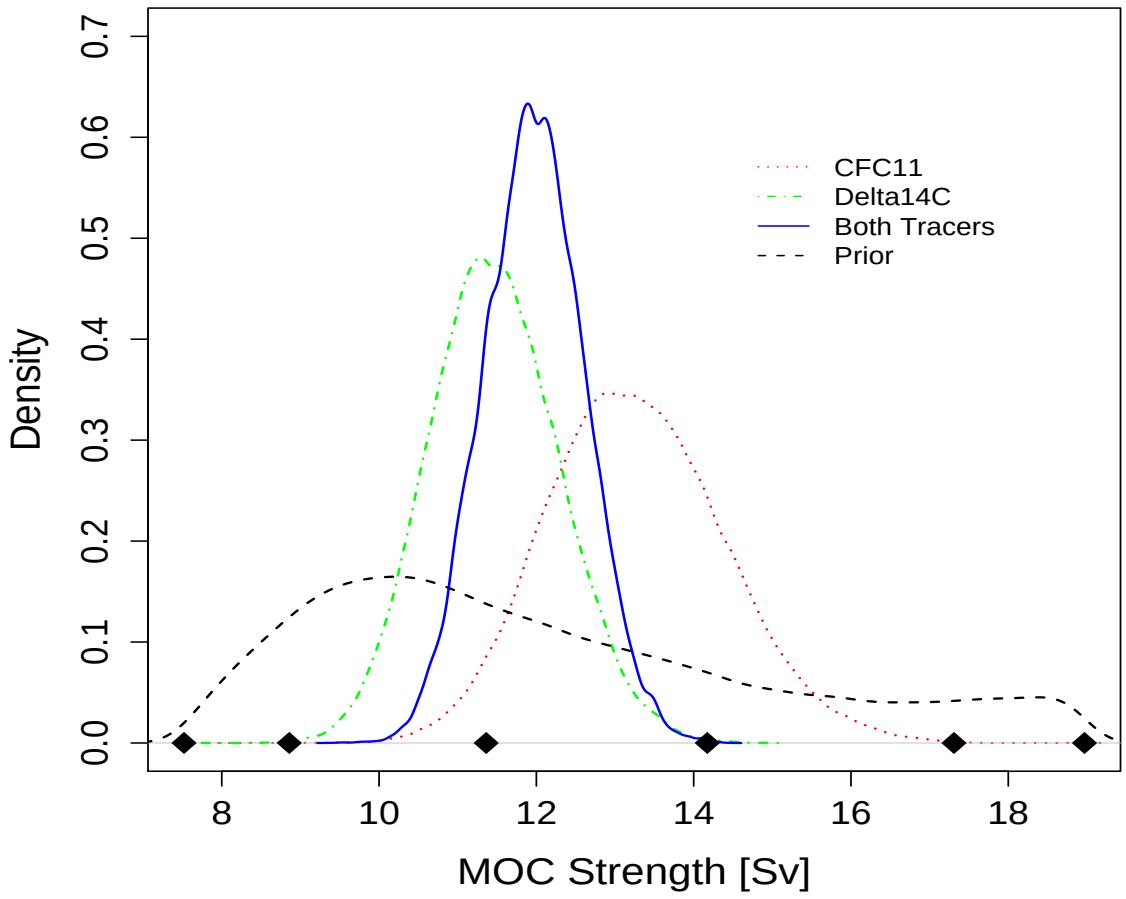
- Taken the climate model output, CFC-11 and C14 spatial fields at several values for $\mathbf{K_v}$, and fit a GP model.
- Now, assume this GP model + model for error, biases is the model for the observations of CFC-11 and C14.
- Since we have a probability model, we can perform inference for $\mathbf{K_v}$ based on the data. That is, we can use statistical techniques to learn about the values of $\mathbf{K_v}$ most compatible with all the information we have. This information will be in the form of a (posterior) probability distribution for $\mathbf{K_v}$.
- *Lots* of complicated computing — matrix operations, optimization, Markov chain Monte Carlo (MCMC).

Results for K_v inference



Probability density functions (pdfs): the prior pdf (assumption *before* using data), and posterior pdfs (*after* using the tracers.)

Results for MOC predictions in 2100



Predictions of MOC in 2100: weaker than current MOC.

Some references

- Kennedy, M.C. and O'Hagan, A.(2001), Bayesian calibration of computer models, *JRSS(B)*.
- Sanso, B. and Forest, C.E. and Zantedeschi, D (2008) , Inferring Climate System Properties Using a Computer Model, *Bayesian Analysis (with discussion)*.
- Higdon (1998) A process-convolution approach to modelling temperatures in the North Atlantic Ocean, *Environmental and Ecological Statistics*.
- Royle, J.A. and Berliner, L.M. (1999) A hierarchical approach to multivariate spatial modeling and prediction, *Journal of Agricultural, Biological, and Environmental Statistics*.
- Bhat, K.S., Haran, M., Tonkonojenkov, R., Keller, K. (2009) "Inferring likelihoods and climate system characteristics using climate models and multiple tracers."

Interesting problems (Part II)

Models for spatial data

- Spatial data: where they are located is critical to the scientific study involving the data.
- A model for spatial data utilizes the location of the data points, for e.g. by assuming that data points closer together are more closely related (dependent) than data further away.

Examples:

- Concentrations of PM2.5 (air pollutants) across the U.S.
- Disease rates by county.
- Abundance of plant/animal species across Pennsylvania.

A model for crop epidemics

Wheat spikes with symptoms of Fusarium head blight

- Fusarium Head Blight (FHB or “Scab”) is a very destructive disease that affects wheat crops.

Forecasting FHB

High levels of mycotoxins ('vomitoxin') accumulate in the kernels, can cause major health problems in humans, animals. Losses of hundreds of millions of dollars per year. Hence, important to forecast such epidemics well.

- Focus on forecasting for the same year. Farmers can look up website to see if there is high risk of FHB.
- Based entirely on weather conditions. Predict FHB three weeks later based on weather today.
- Example: What is the probability of an FHB epidemic on July 10th at latitude 69.79 and longitude 46.00?
  Answer: 0.43.

Forecasting FHB: old approach

- How is this probability obtained ?
  - Obtain weather conditions at latitude 69.79 and longitude 46.00 via radar (Rapid Update Cycle).
  - Use a model to obtain probability of an epidemic.
- Model is obtained from data collected at experimental plots set up by plant pathologists across 12 states.
- Weather information and crop disease (FHB) rates are observed at each plot.
- Using knowledge of disease biology and data from experimental plots, can obtain a "logistic regression" model that predicts an FHB epidemic based on weather.
  $p(s,t) = P$(epidemic in site s at time t) is a function of weather (temperature, humidity, precipitation etc.) at site s at time t.

FHB risk and flowering

- Two processes are involved in determining risk of FHB:
  - **Weather factors**: High humidity, rainfall, optimal temperatures create ideal conditions for an FHB epidemic.
  - **Flowering dates**: An outbreak can only occur within a small window of time around the flowering date for the crop.
- An FHB epidemic is most likely when optimal weather conditions occur prior to or during flowering.
- If flowering and weather conditions for FHB do not coincide, no epidemic. Rule of thumb: these weather conditions must occur within 2 days (plus or minus) of flowering date for an outbreak. (A “perfect storm”.)

Forecasting model for estimating risk of FHB

- Using model and weather data at any location, get $p(s, t)$.
- These risks only describe the predisposition or *susceptibility* of a site to an FHB epidemic.
- FHB is primarily spread to adjacent regions when spores of the fungi are windblown or splashed onto the heads of the crops, need to model **space-time dependence.**

Sources of information

Case study: North Dakota, 2005.

I. Experimental Plots:

Severity	Temperature	Humidity	Precipitation

- Set up by plant pathologists to determine relationship between relative humidity, temperature, precipitation and severity of FHB.
- Data available from 1990-2005: 12 plots total in 7 states.

II. RUC (Rapid Update Cycle) metereological information:

Latitude	Longitude	Weather 1-5

- Centroids of 20km$\times$20km sites across the entire country.
- Obtained daily from national weather system (devices on commercial aircraft, surface reporting stations etc.)

Sources of information (contd.)

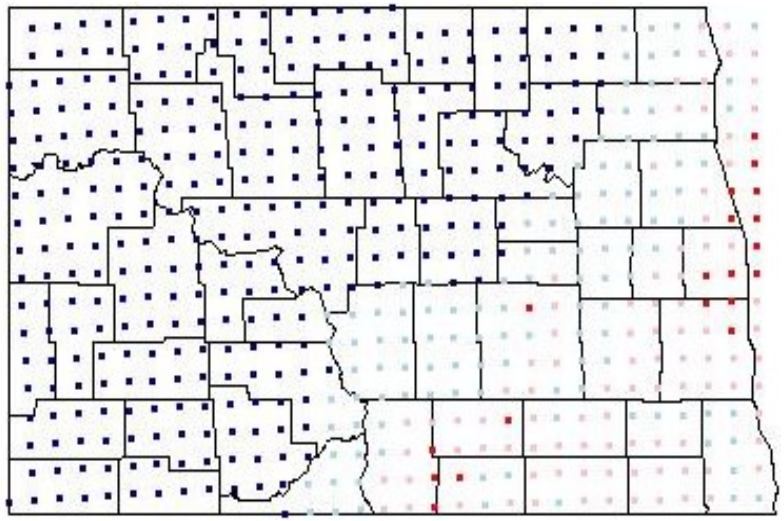
III. Flowering information: Data on flowering dates of wheat crops is available for several sites — surveyors either backtrack and estimate the flowering date after it has happened or predict when it is going to happen for a particular site. Note: this means flowering information is also only an estimate (based on observation+model).

Overview for North Dakota:

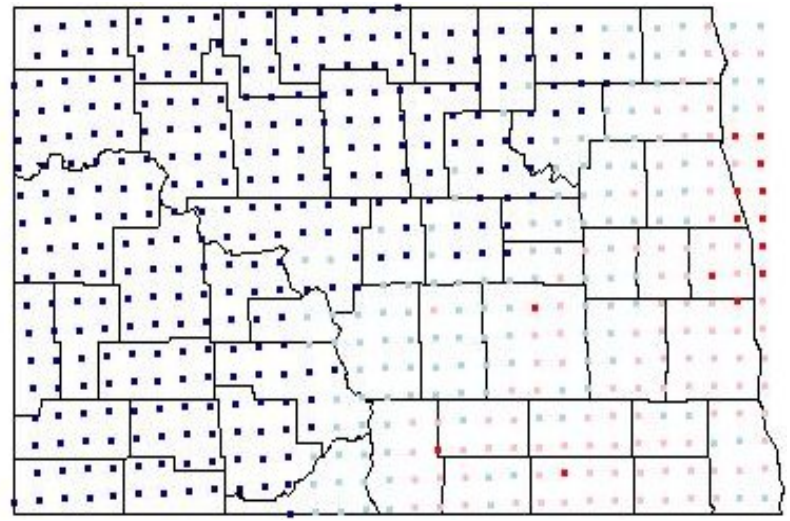
- 545 RUC sites where plant pathologists' risk predictions are available over a total of 44 days in the summer: June 8, 2006 through July 28, 2006 (though several days of data are missing during this period).
- Flowering information is available at 365 surveyed sites.

I. Susceptibilities: red=epidemic, blue=no epidemic

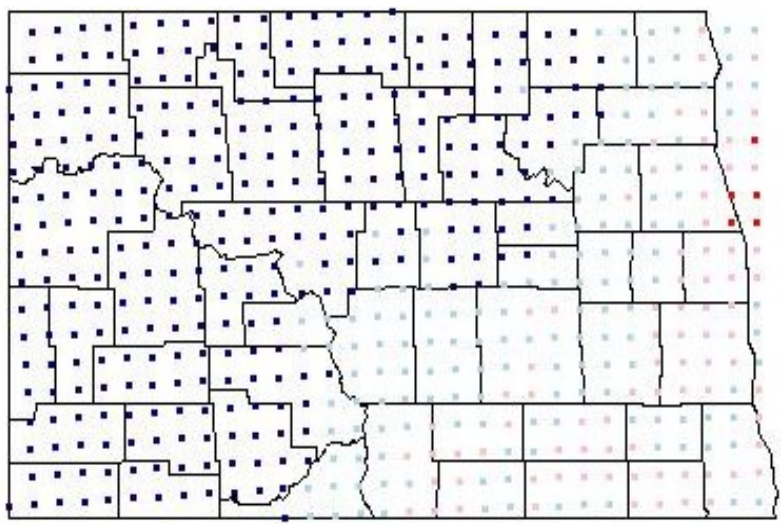
July 8



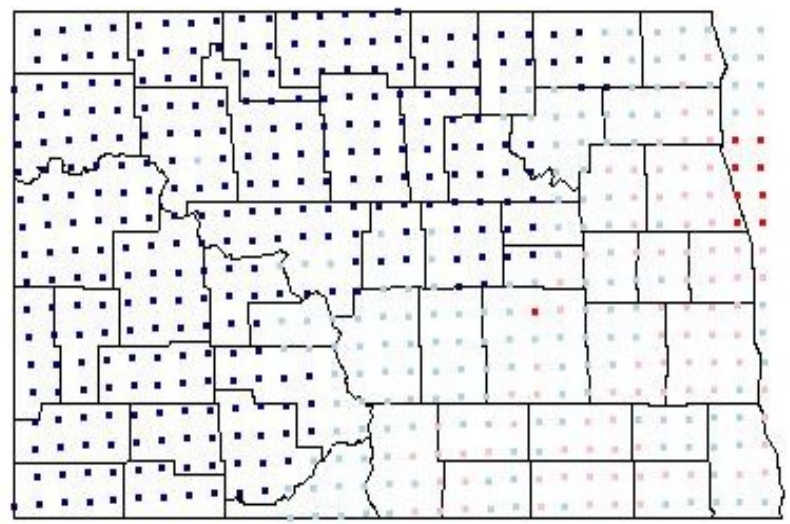
July 9



July 10



July 11



red⇒P(epidemic>0.5)

blue⇒P(epidemic<0.1).

II. Flowering dates from survey

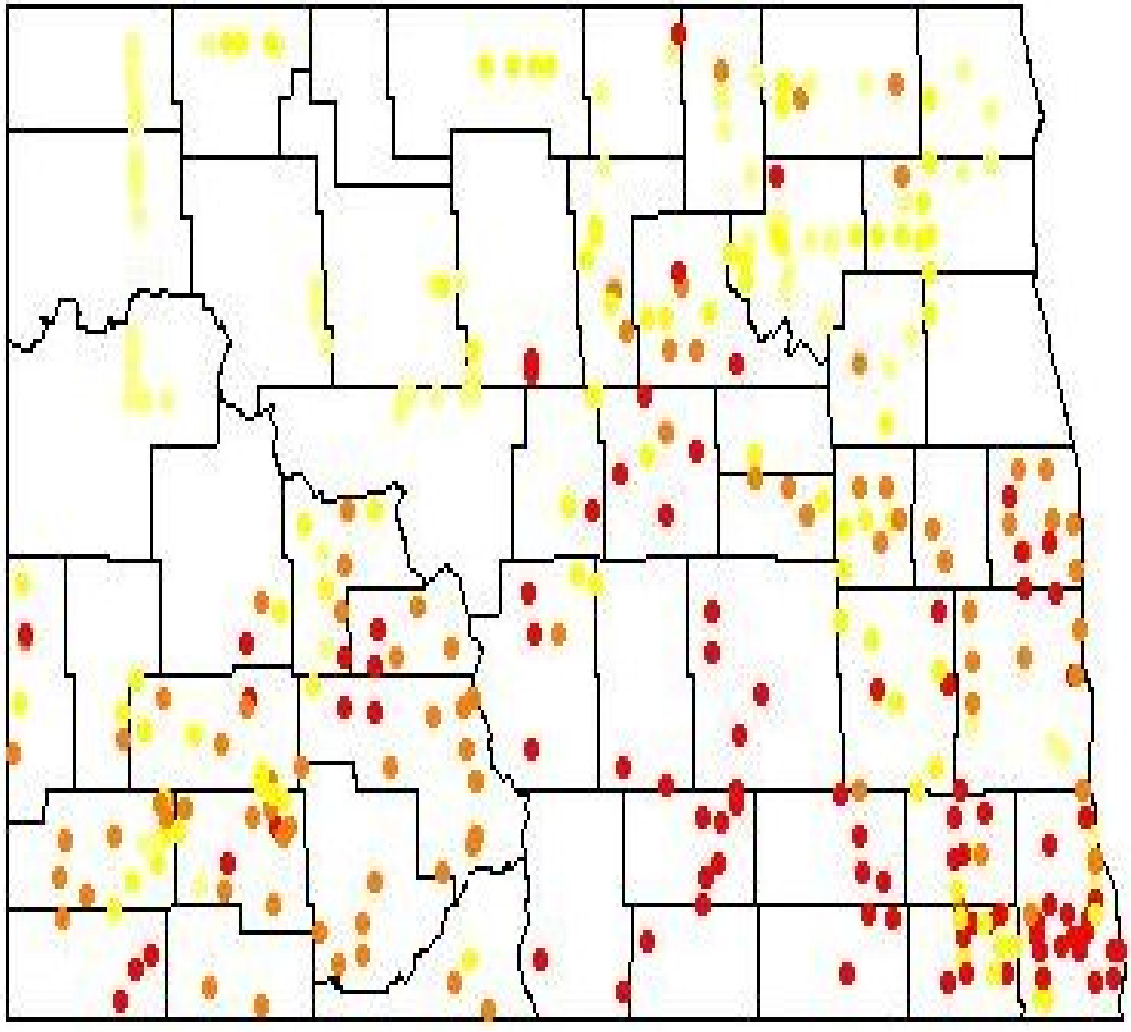


Figure: North Dakota flowering dates, 2005: red: <July 6, orange: July 6 to July 12, brown:July 13 to July 18, yellow: after July 18.

A model for FHB risk

- How do we connect these different sources of information and place them in one coherent framework? Note: Flowering dates information is available at *different locations* from the information where risks are available. Need to do statistical interpolation to co-locate flowering date information and risks.
- We also allow for the fact that several pieces of information going into the model have errors associated with them.
- We develop a **hierarchical Bayes model** that uses spatial models to integrate the multiple sources of information and account for variability in a systematic manner.

A model for FHB risk: outline

1. For RUC locations $\mathbf{s}$, FHB epidemic probabilities $p(\mathbf{s}, t)$.
2. Flowering dates at locations $\mathbf{u}_1$ and $\mathbf{u}_2$, $d(\mathbf{u}_1), d(\mathbf{u}_2)$ are spatially related. Using this, we can figure out (interpolate) the flowering date at $\mathbf{s}$ where $\mathbf{s}$ is an RUC location.
3. From (1) we have improved estimates $p(\mathbf{s}, t)$ and from (2) estimates of flowering dates $d(\mathbf{s})$ at the same locations. Using $d(\mathbf{s})$ we can find probability t is a flowering date.
4. Final estimates of risk: $p(\mathbf{s}, t) \times$ Prob(t is a flowering date). We can also provide errors for these estimates. Important!

Haran, M., Bhat, K.S., Molineros, J, and De Wolf, E. (2009) "Estimating the risk of a crop epidemic from coincident spatiotemporal processes," *Journal of Agricultural, Biological and Environmental Statistics*.

Interesting problems (Part III)

Statistical Computing (Markov chain Monte Carlo):

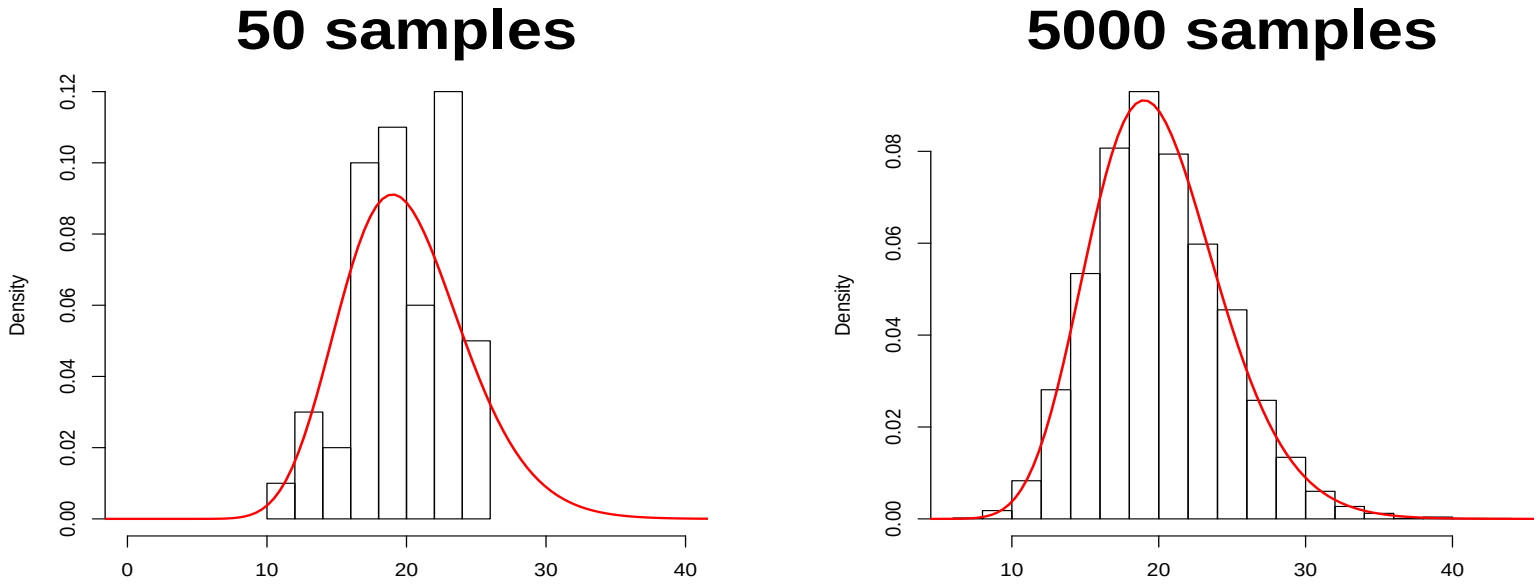
- When the statistical model used is complicated, it becomes challenging to fit the model to the data. That is, it becomes challenging to perform statistical inference since it may involve doing very high dimensional integrals. [You need to learn about *maximum likelihood inference* and *Bayesian inference* to see why.]
- Markov chain Monte Carlo (MCMC) is an approach that can help perform inference for lots of very hard problems by using computer simulations. An important research area: lots of theory, lots of computing.

Toy Monte Carlo example

X $\sim \pi$ where π is a Gamma density, Gamma($\alpha = 20, \beta = 1$).

$\mathbf{E}_\pi \mathbf{X} = \mathbf{20}$ and $\Pr(X > 22) = \mathbf{E}_\pi \mathbf{I}(\mathbf{X} > \mathbf{22}) = \mathbf{0.306027}$.

Simulate: $X_1, \ldots, X_N \overset{iid}{\sim}$Gamma(20,1). Red=true density.



N= 50: estimates $E_\pi X$=19.70698 and $\Pr(X > 22)$=0.34.

N=5000: estimates $E_\pi X$=20.05199 and $\Pr(X > 22)$=0.3102.

The last slide (I promise)

- My research on computer experiments, spatial models and Monte Carlo methods lets me work on lots of interesting research problems from different fields.
- Other examples I did not discuss: infectious disease dynamics, modeling invasive plant species.
- Statistics research offers opportunities to combine interests in science, statistical modeling, theory (opportunities to do hard mathematics), and computer science.
- Lots of other fascinating research problems being worked on by my colleagues, Ph.D. students and postdoctoral reseachers, for example at Penn State Statistics, many other departments, research labs and government agencies around the country.