An Introduction to Computer Model Emulation and Calibration

Murali Haran

Department of Statistics, Pennsylvania State University

Spatially-varying Stochastic Differential Equations
Math Biosciences Institute (MBI) Workshop
Ohio State University, Columbus, Ohio. July 2015.

Talk Summary

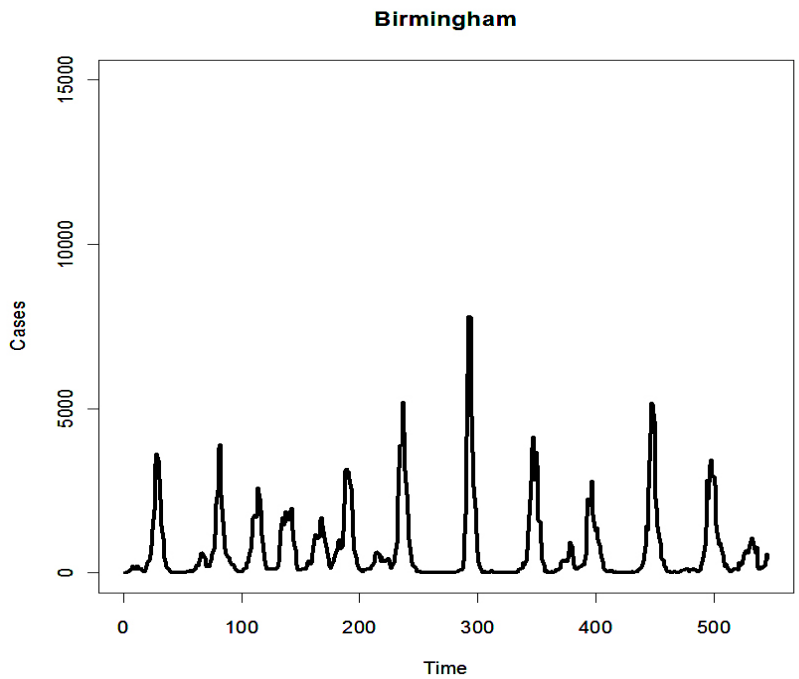
- Statistical inference for deterministic or stochastic models is often intractable
  - One way to learn about properties of these models may be to simulate them on a computer
  - Statistical inference for model parameters therefore involves observations of the process and output from the computer model at various parameter settings
  - Model simulations may also be computationally expensive
- Emulation: statistical approximation to model
- Calibration: inference for model parameters
- I will describe Gaussian process-based methods for emulation and calibration

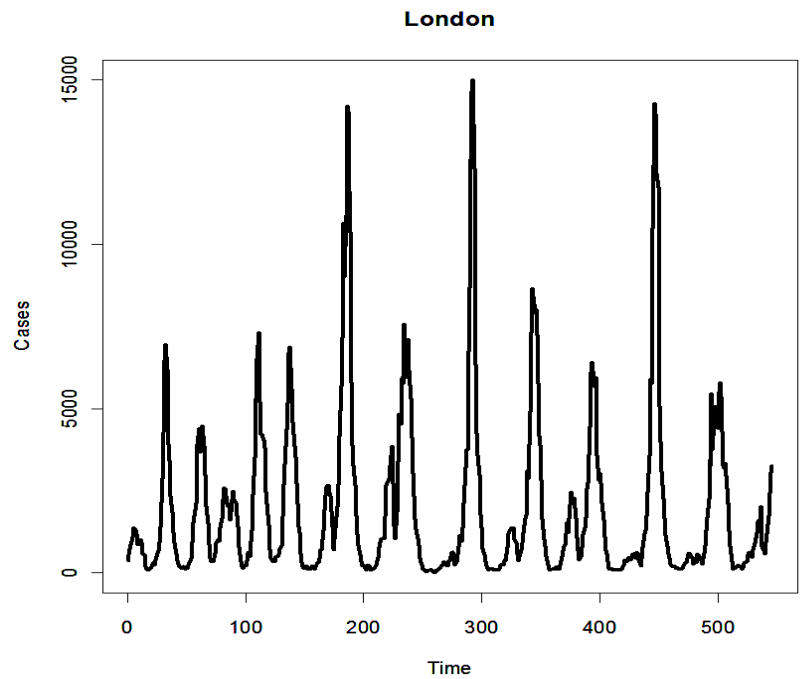
Motivating Examples

1. Gravity TSIR (time series susceptible-infected-recovered) model for measles infections
   - Stochastic model
   - Relatively simple: can write it down in closed form
   - Data/model output: time series of # infected at 546 time points $\times$ 952 cities
   - Can do likelihood-based inference in principle but likelihood evaluations are computationally expensive
2. Model for an ice sheet (e.g. West Antarctic).
   - Deterministic (system of partial differential equations)
   - No analytical form available
   - Data/model output: spatial binary data of ice presence-absence at various locations

Common to both: treat the model as a black box – only use model simulations, not its analytical description.

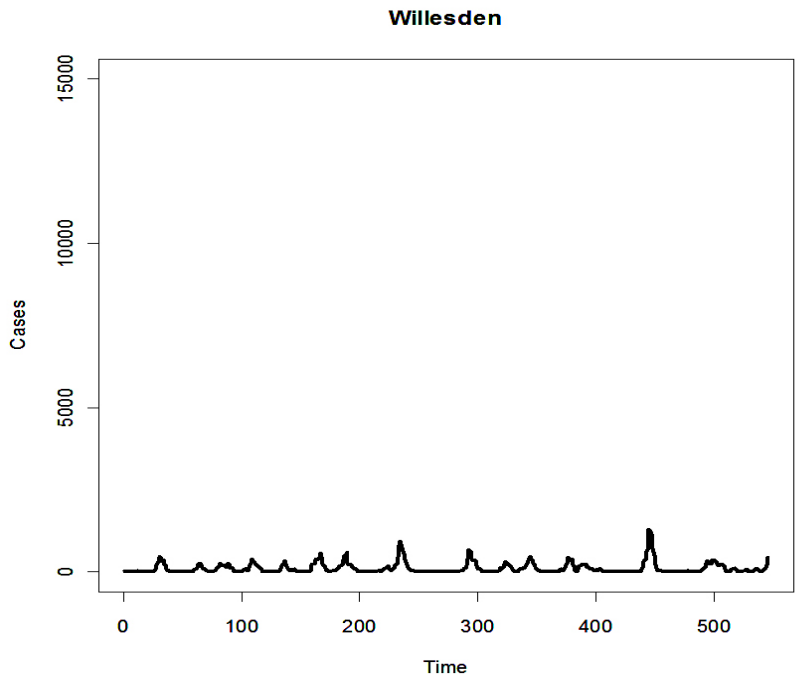
Measles Data: London and Birmingham

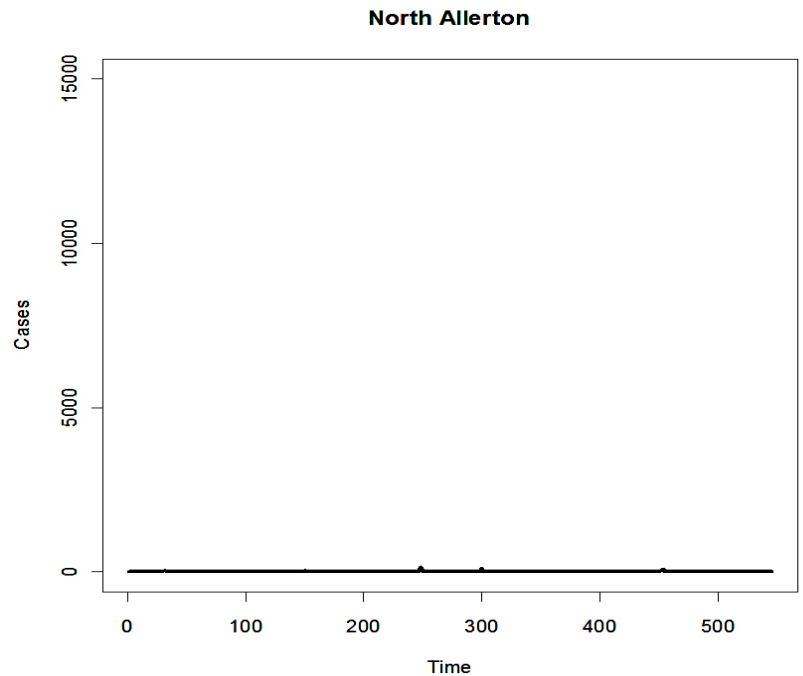




Time series of number of people infected by measles
Source: UK Registrar General's biweekly data for 952 cities in England and Wales for years 1944-1966

Measles Data: Willesden and North Allerton





Total of 952 time series (one for each city) of varying sizes and levels of number of infected.

Basic SIR Model

Compartmental model for dynamics:

- Individuals move from one class to another at various transition rates
- Transitions may be determined by size and birth rate in a city, distances among individuals
- SIR models: ODEs, PDEs, stochastic versions, with lots of sophisticated variants (cf. Keeling and Rohani, 2007)

Gravity TSIR Model

Information/data:

- I_{kt} : number of infected individuals in city k at time t
- S_{kt} : number of susceptible individuals in city k at time t
- L_{kt} : number of infected people moved to city k at time t
- d_{kj} : distance between cities k and j
- N_{kt}, B_{kt} : size and birth rate of city k at time t

Parameters to infer:

- Local transition parameters α and β. Use established ways to infer these non-spatial parameters (Bjørnstad et al. 2001).
- **Spatial dynamics "gravity" parameters** θ, τ_1, τ_2 and ρ. Estimation problem of interest.

Details: Gravity TSIR Model

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

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

- $I_{k(t+1)}$ increases with I_{kt}, S_{kt}, and number of infected immigrants coming to city k at time t (L_{kt}).
- $\{\beta_t\}$ are 26 different parameters that are repeated every year to allow differences in seasonal transmission (26 = number of biweeks in a year).

(Xia, Bjørnstad and Grenfell, 2004; Jandarov, Haran, Bjørnstad and Grenfell, 2014)

Details: Gravity TSIR Model (cont'd)

- Number of susceptible individuals at time $t+1$ for city k, $S_{k(t+1)} = S_{kt} + B_{kt} - I_{k(t+1)}$.
- Number of infected immigrants (latent) at time t for city k

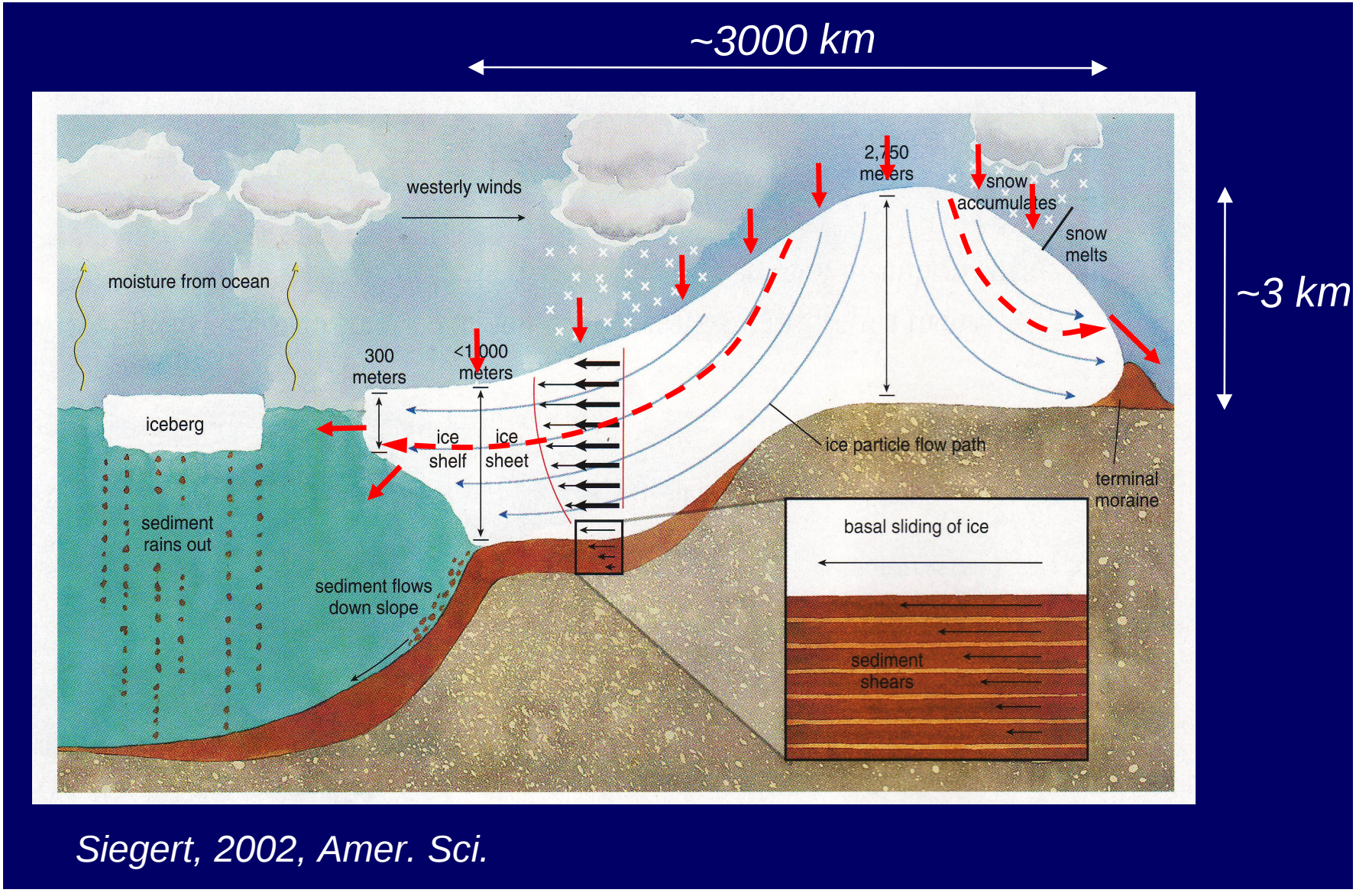
$$L_{kt} \sim \text{Gamma}(m_{kt}, 1), \text{ where } m_{kt} = \theta N_{kt}^{\tau_1} \sum_{j=1, j\neq k}^{K} \frac{(Ijt)^{\tau_2}}{d_{kj}^{\rho}}.$$

- L_{kt} increases with size of city k, number of infected people in all other cities, taking into account distances.

Challenges

- Dimensions of the data (TK): 546*952 = 519,792.
- Number of infected immigrants $\{L_{k,t}\}$ are unobserved.
- The likelihood function
  $\mathcal{L}(\theta, \tau_1, \tau_2, \rho; \{I_{kt}\}) = \int \mathcal{L}(\theta, \tau_1, \tau_2, \rho, \{L_{k,t}\}; \{I_{kt}\}))d\{L_{k,t}\}$.
  - Involves integrating over 519,792 latent variables.
  - Expensive calculations per iteration.
- Possible solution: simulate from this model for a variety of different parameter values. Treat this as a computer model emulation-calibration problem.

Ice Sheet Physics



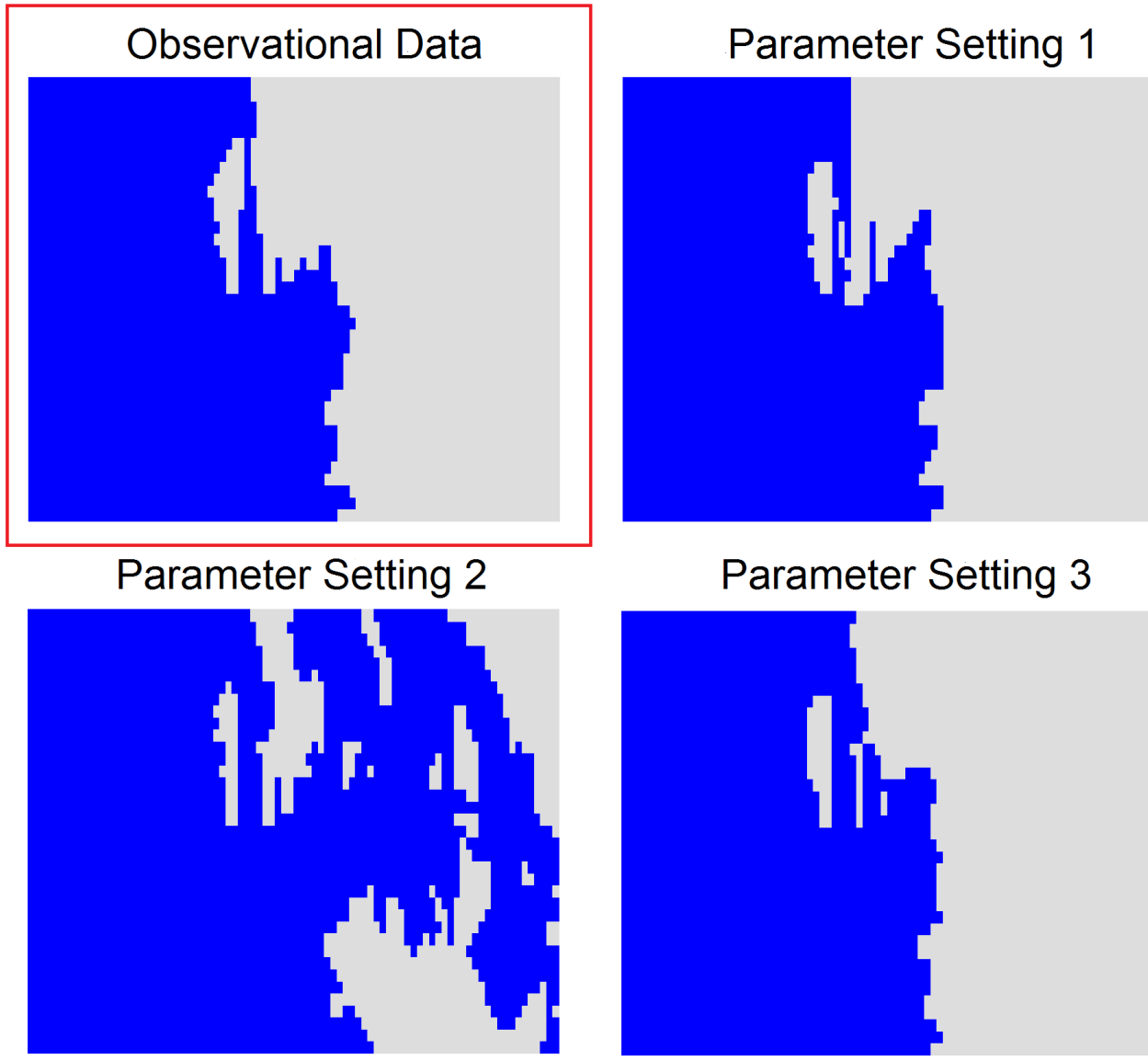
Siegert, 2002, Amer. Sci.

Ice Sheet Model Parameters

The ice sheet model is complex

- Model equations predict ice flow, thickness, temperatures, and bedrock elevation, through thousands to millions of years.
- Examples of key model parameters:
  - Ocean melt coefficient: sensitivity of ice sheet to temperature change in the surrounding ocean
  - Strength of the “calving” process. Calving = where ice breaks off and transitions from attached to floating
  - “Slipperiness” of the ocean floor

West Antarctic Ice Sheet Example

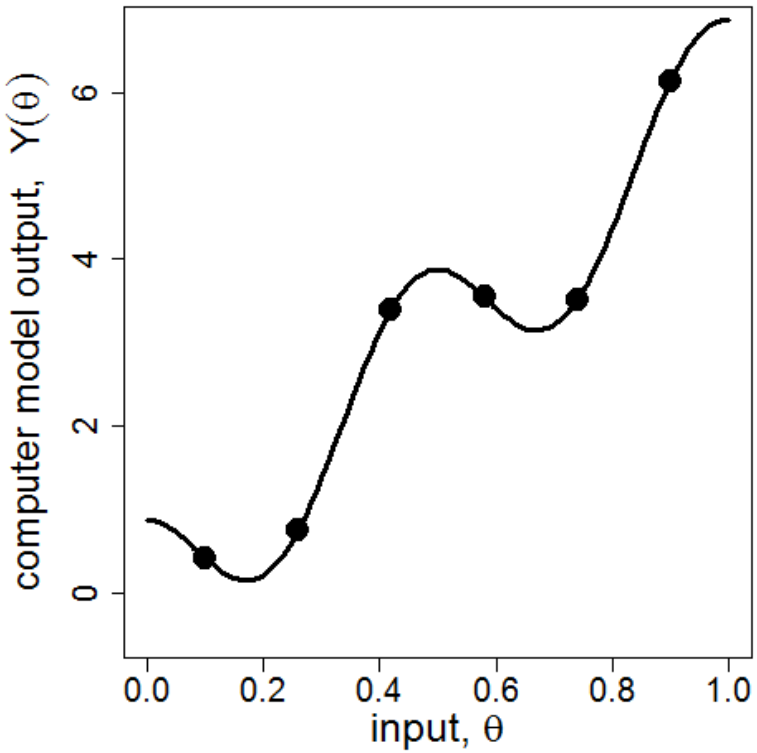


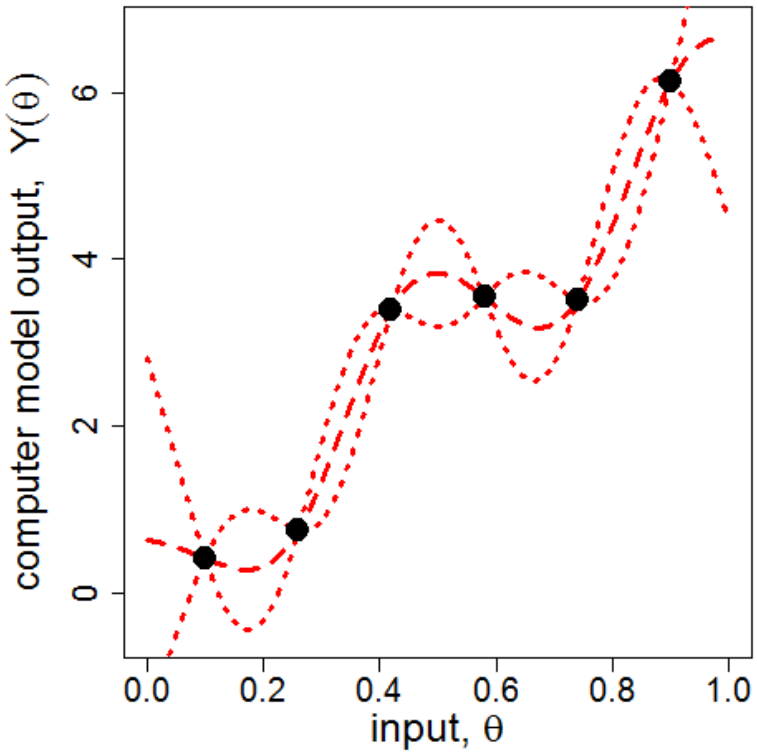
Two-stage Approach to Emulation-Calibration

1. Emulation: Run model at various parameter settings. Fit a Gaussian process to these model runs to obtain a model approximation: for a new parameter value, can approximate model output.
   Original reference: Sacks et al. (1989) though interpolation idea dates back to original kriging methods from the 1950s.
2. Calibration step: Infer parameters using emulator and observations, while accounting for data-model discrepancy.
   Original reference: Kennedy and O'Hagan (2001)

(Liu, Bayarri and Berger, 2009; Bhat, Haran, Olson, Keller, 2012)

Emulation: Toy e.g. with scalar, continuous output

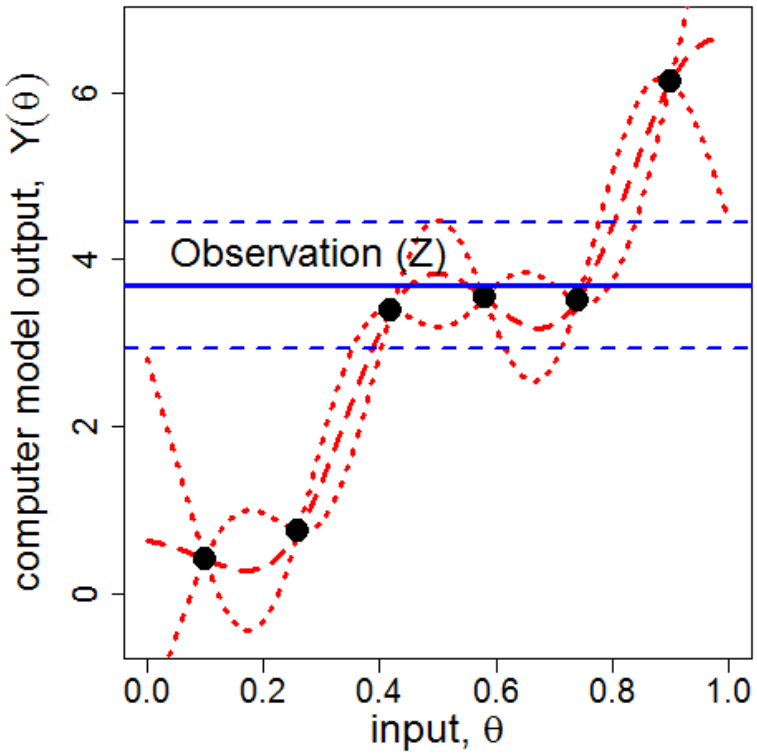




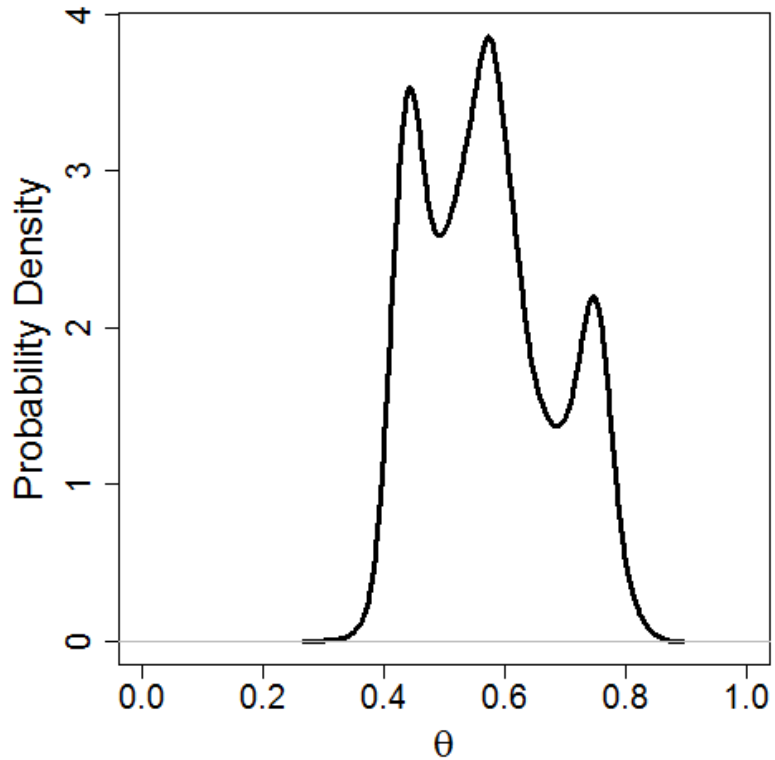
- Fitting a stochastic process to a realization of a curve with only discrete observations. This is an interpolation across *parameter space* (instead of time or space)
- Gaussian process dependence picks up nonlinearities, allows for “nonparametric” curve fitting
- To prevent perfect interpolation can add a nugget

Calibration

Toy example: model output, observations are scalars



Combining observation and emulator



Posterior PDF of θ given model output and observation

Note: multimodal posterior pdf, very common in practice. Much more useful than simply providing a "best parameter" estimate.

Summary of Statistical Problem

- Toy examples before had scalar computer model output
- Usually the computer model output is multivariate. Often time series, spatial, spatio-temporal, multiple curves/surfaces etc.

- **Goal**: Learn about model parameter(s) θ. We have two sources of information:
  - **Observations**: $\mathbf{Z} = (Z(\mathbf{s}_1), \ldots, Z(\mathbf{s}_n))^T$, where $\mathbf{s}_1, \ldots, \mathbf{s}_n$ locations (1D, 2D or 3D)
  - **Model output** $\mathbf{Y}(\theta_1), \ldots, \mathbf{Y}(\theta_p)$, where each $\mathbf{Y}(\theta_i) = (Y(\theta_i, \mathbf{s}_1), \ldots, Y(\theta_i, \mathbf{s}_n))^T$ is a vector of spatial data
- Often important to determine whether and how to transform the multivariate output $\mathbf{Y}$ into a summary.
  - Measles dynamics: biologists provide important "signatures" of disease dynamics, e.g. peak incidences, proportion of time without incidences.

Step 1: Computer Model Emulation

- Fit Gaussian process model for computer model output $\mathbf{Y}$ to interpolate the values at the parameter settings $\theta_1, \ldots, \theta_p$ *and* the spatial locations $\mathbf{s}_1, \ldots, \mathbf{s}_n$

$$\text{vec}\,(\mathbf{Y}) \sim N(\mathbf{X}\beta, \mathbf{\Sigma}(\xi_y)),$$

  vec$(\cdot)$ concatenates columns into one vector
- β and ξ_y estimated by maximum likelihood, $\hat{\beta}, \hat{\xi}_y$.
- Covariance interpolates across spatial surface *and* input space.

Result: Obtain a probability model (= predictive distribution) for model output at any input parameter θ, $\eta(\theta, \mathbf{Y})$.

Step 2: Calibration

- Discrepancy $\approx$ mismatch between computer model output and data when parameters are perfectly calibrated and there is no observational error.
- Represents discretizations, "parameterizations" of complex processes, other simplifications + acknowledging that no model $=$ reality
- Discrepancy is enormously important. Without discrepancy term, θ inference will be wrong (Bayarri et al., 2009a,b). Hence needs to be modeled carefully.
- If discrepancy is too flexible, will be confounded with the model. Will necessarily require expert judgment to help "unconfound" this term from the rest of the model. Convenient to do this in a Bayes framework.
- Probability model for observations $\mathbf{Z}$ is then

$$\mathbf{Z} = \eta(\boldsymbol{\theta}, \mathbf{Y}) + \boldsymbol{\delta},$$

where n-dimensional spatial field $\boldsymbol{\delta}$ is model-observation discrepancy with covariance parameter ξ_δ.

Step 2: Calibration (Inference)

- Inference for θ based on posterior distribution

$$\pi(\theta, \xi_\delta | \mathbf{Z}, \mathbf{Y}, \hat{\xi}_y) \propto \underbrace{\mathcal{L}(\mathbf{Z}|\mathbf{Y}, \theta, \xi_\delta, \hat{\xi}_y)}_{\text{likelihood given by above}} \times \underbrace{p(\theta) \times p(\xi_\delta)}_{\text{priors for } \theta \text{ and } \xi_\delta}$$

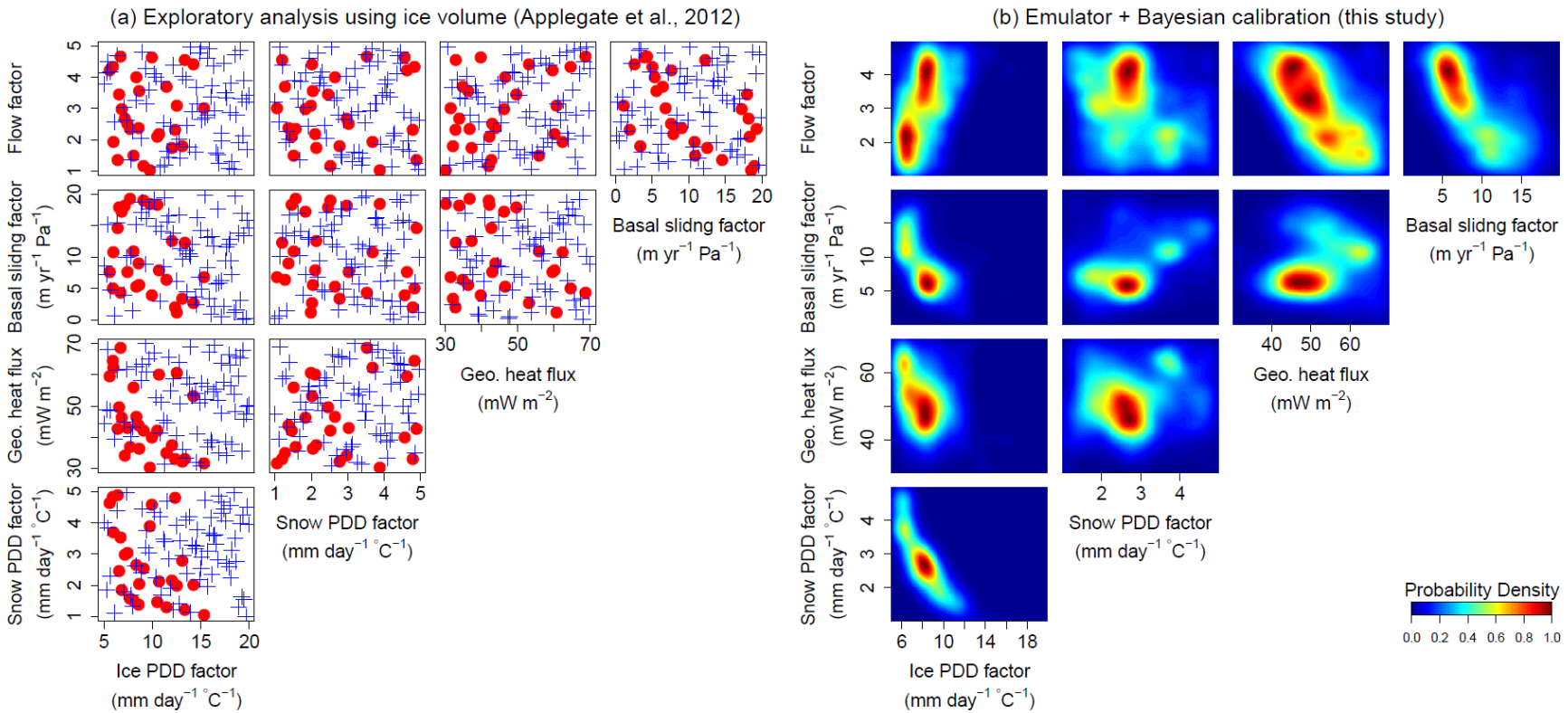
  with parameter $\hat{\xi}_y$ fixed from emulation step.
  MCMC or other methods to infer posterior.
- Note: may be useful to think of $\mathcal{L}(\mathbf{Z}|\mathbf{Y}, \theta, \xi_\delta, \hat{\xi}_y)$ as $\hat{\mathcal{L}}(\mathbf{Z}|\mathbf{Y}, \theta, \xi_\delta, \hat{\xi}_y)$, an "inferred" or estimated likelihood function. Inferred from computer model output across multiple parameter settings + discrepancy model.

How Does Statistical Rigor Help Scientists?

Ad-hoc non-statistical methods may seem reasonable, e.g. methods based on finding parameter settings that minimize a distance between the data and the model output. But being statistically rigorous has important benefits:

1. We account for (epistemic) uncertainties in emulation
2. We provide *real* probability distributions
   - Very important for projections based on the model: approximate projections averaging across the pdf of θ.
   - We provide more interpretable results.
3. Use all available information: often reduces uncertainties.

Example of Sharper Results

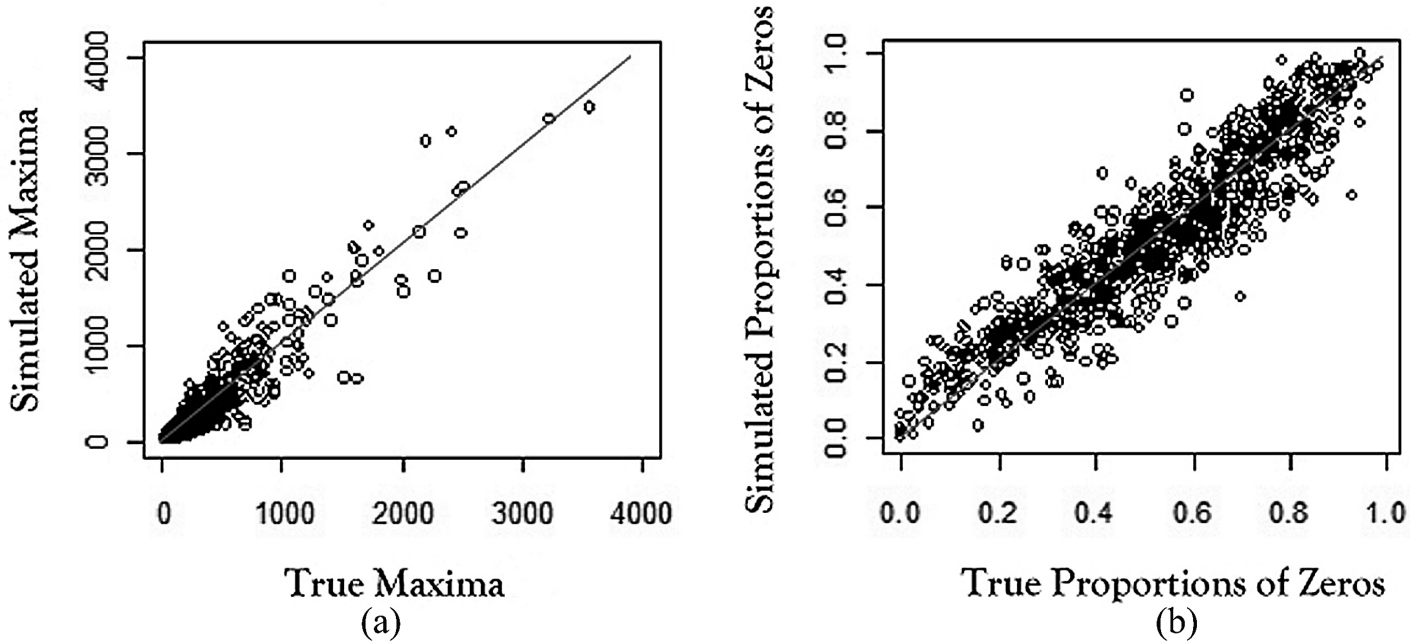


Left: previous ad-hoc methods. Right: statistical calibration

Ice sheet example (cf. Chang, Applegate, Haran, Keller, 2014; Chang, Haran, Pollard, Applegate, 2015)

Fitting Biological Characteristics using GP-approach

Measles dynamics example (Jandarov, Haran, Bjornstad, and Grenfell, 2014)



Fitted model better captures important characteristics of the data.

Concluding Remarks

- I have described methods for approximating and performing statistical inference for complex models.
- *In principle* can do this for a wide variety of models, stochastic or deterministic.
- Important to account for sources of uncertainty: e.g. interpolation uncertainty, model-data discrepancy, measurement error. Bayesian methods are helpful (central) for making this easy to specify.
- Can use inferred parameter distributions to study the process of interest, e.g. projections of future ice sheet, infectious diseases under different vaccination regimes
- When data are high-dimensional, non-Gaussian, need other methods. e.g. dimension reduction, GLMMs
- If simulations are fast, Approximate Bayesian Computing (ABC) (cf. Beaumont et al., 2002) may be worth exploring

Acknowledgments

Collaborators:

- Won Chang, University of Chicago Statistics
- Roman Jandarov, University of Cincinnati Biostatistics
- David Pollard, Earth and Environmental Systems Institute (EESI), Penn State U.
- Patrick Applegate, EESI, Penn State U.
- Klaus Keller, Geosciences, Penn State U.
- Roman Olson, The University of New South Wales

This work was partially supported by the following grants:

- Bill and Melinda Gates Foundation
- The Network for Sustainable Climate Risk Management (SCRiM), **NSF GEO-1240507**.
- **NSF CDSE/DMS-1418090** Statistical Methods for Ice Sheet Projections

References

- Sacks, Welch, Mitchell, Wynn (1989) “Design and analysis of computer experiments,” *Statistical Science*
- Kennedy and O’Hagan (2001) “Bayesian Calibration of Computer Models,” *J of Royal Stat Soc (B)*
- Calibration with high-dimensional, non-Gaussian spatial
  - Jandarov, R., Haran, M., Bjornstad, O.N. and Grenfell, B. (2014) “Emulating a gravity model to infer the spatiotemporal dynamics of an infectious disease.” *J of the Royal Stat. Society (C)*
  - Chang, W., Haran, M., Olson, R., and Keller, K. (2014) Fast dimension-reduced climate model calibration, *Annals of Applied Statistics, arXiv:1303.1382*.
  - Chang, W., Applegate, P., Haran, M. and Keller, K. (2014) Probabilistic calibration of a Greenland Ice Sheet model using spatially-resolved synthetic observations *Geoscientific Model Development*
  - Chang, W., Haran, M., Applegate, P., and Pollard, D. (2015) Calibrating an ice sheet model using high-dimensional non-Gaussian spatial data *on arxiv.org*