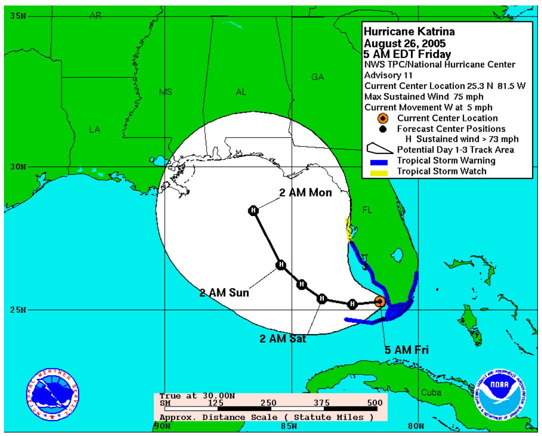
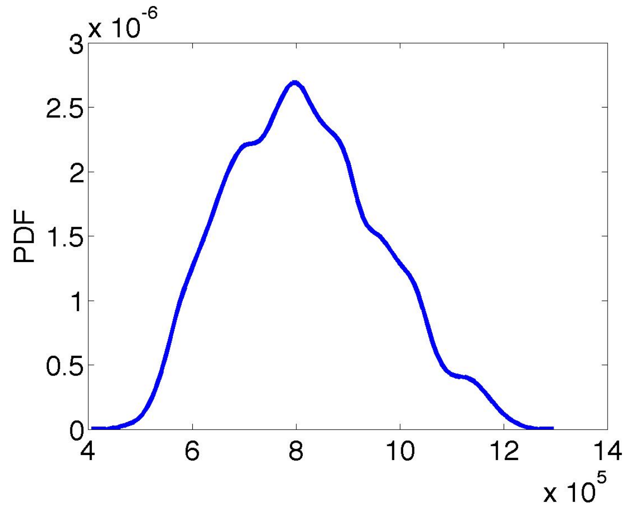
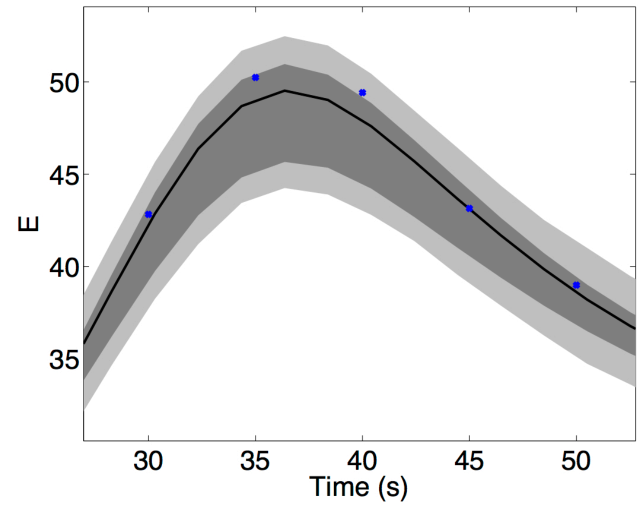
Uncertainty Quantification for Biological Models

Ralph C. Smith
Department of Mathematics
North Carolina State University







Essentially, all models are wrong, but some are useful, George E.P. Box, Industrial Statistician.

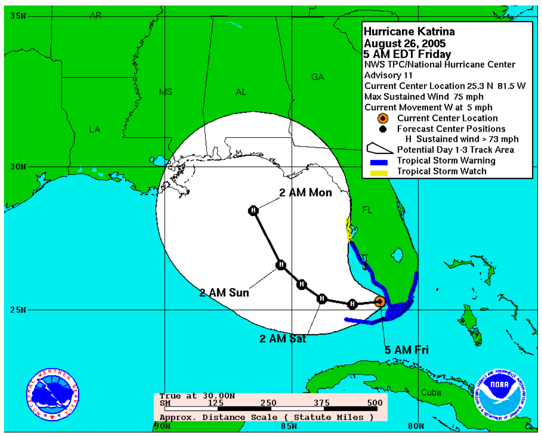
Support: DOE Consortium for Advanced Simulation of LWR (CASL)
NNSA Consortium for Nonproliferation Enabling Capabilities (CNEC)
National Science Foundation (NSF)
Air Force Office of Scientific Research (AFOSR)

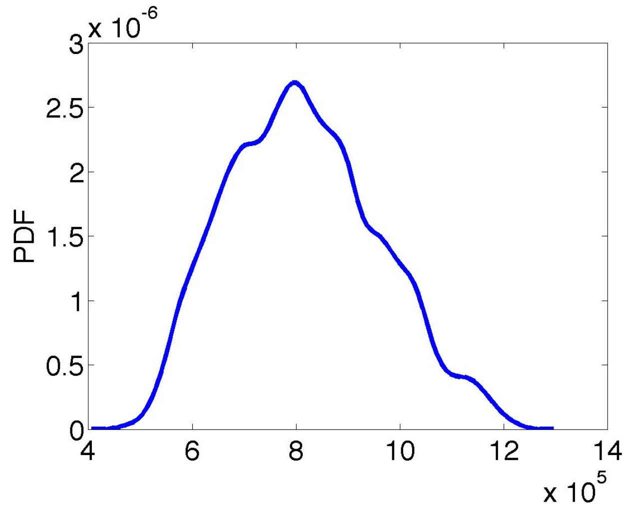
Uncertainty Quantification for Biological Models

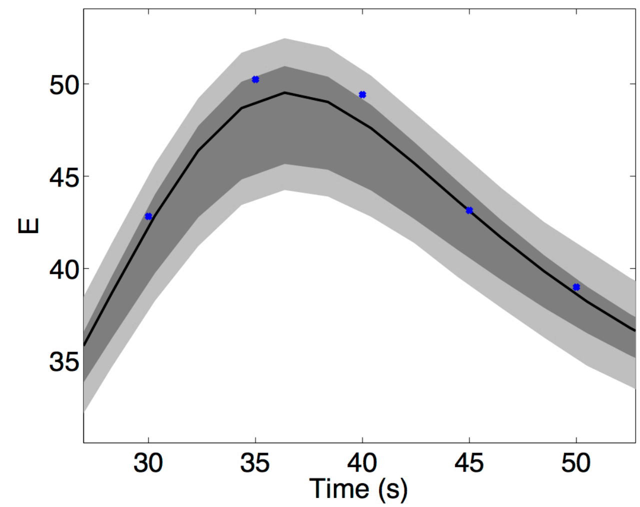
Ralph C. Smith

Department of Mathematics

North Carolina State University







”We”:

Zhengzheng Hu, Nate Burch, Allison Lewis, Kathleen Schmidt, Nikolas Bravo, Mami Wentworth (NCSU)

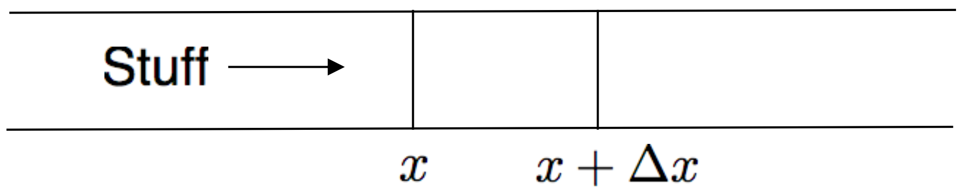
Michael Hays, Billy Oates (Florida State University)

Brian Williams (LANL), Russell Hooper, Brian Adams, Vince Mousseau (Sandia)

Emre Tatli and Yixing Sung (Westinghouse)

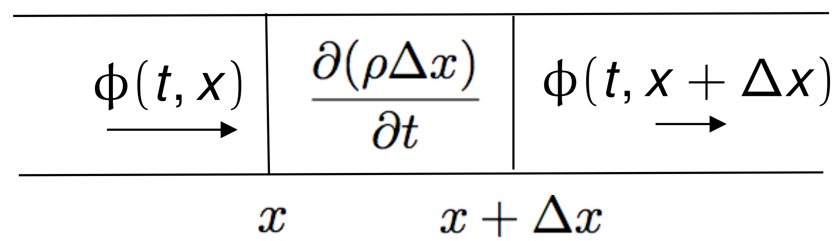
Modeling Strategy

General Strategy: Conservation of stuff



$$\frac{dStuff}{dt} = \text{Stuff in - Stuff out + Stuff created - Stuff destroyed}$$

Continuity Equation:



$$\frac{\partial(\rho \Delta x)}{\partial t} = \phi(t, x) - \phi(t, x + \Delta x)$$

$$\Rightarrow \lim_{\Delta x \to 0} \frac{\partial \rho}{\partial t} = \lim_{\Delta x \to 0} \frac{\phi(t, x) - \phi(t, x + \Delta x)}{\Delta x}$$

$$\boxed{\Rightarrow \frac{\partial \rho}{\partial t} + \frac{\partial \phi}{\partial x} = 0}$$

Density: $\rho(t, x)$ - Stuff per unit length or volume

Rate of Flow: $\phi(t, x)$ - Stuff per second

More Generally:

$$\boxed{\Rightarrow \frac{\partial \rho}{\partial t} + \frac{\partial \phi}{\partial x} = \text{Sources - Sinks}}$$

Example 1: Weather Models

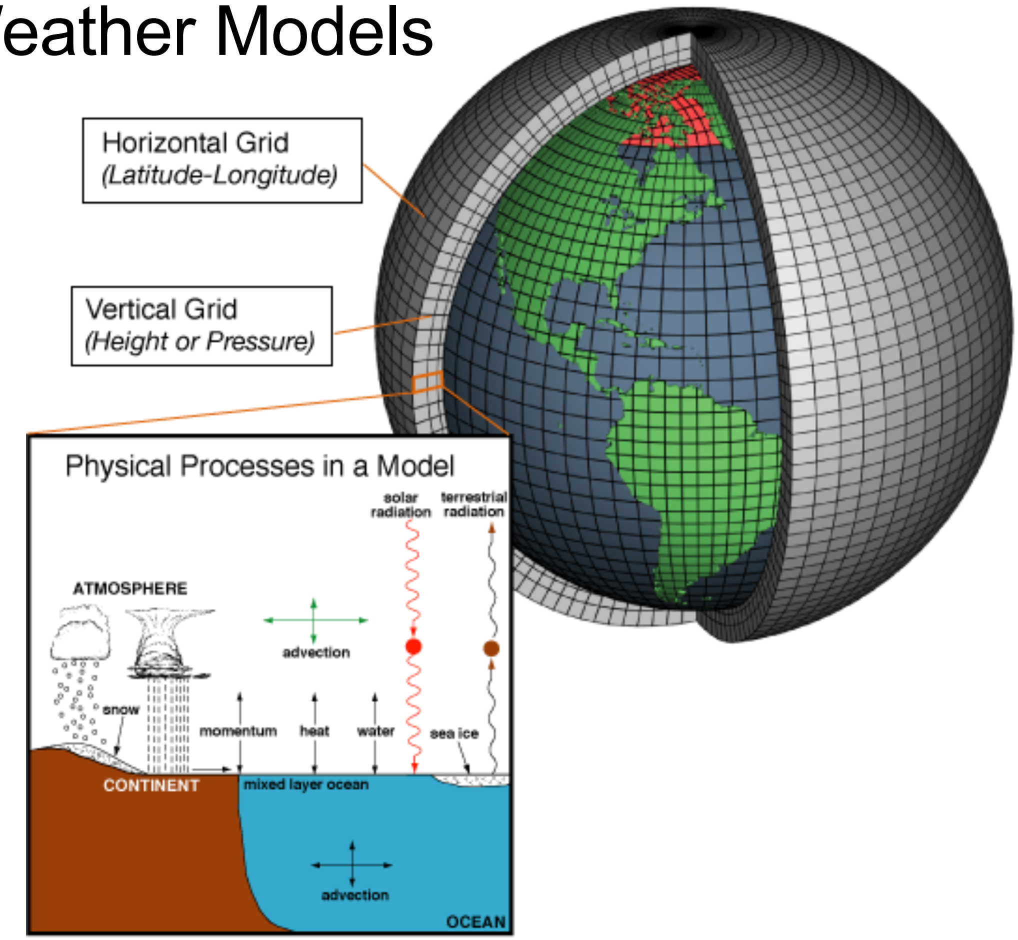
Challenges:

• Coupling between temperature, pressure gradients, precipitation, aerosol, etc.;

• Models and inputs contain uncertainties;

• Numerical grids necessarily larger than many phenomena; e.g., clouds

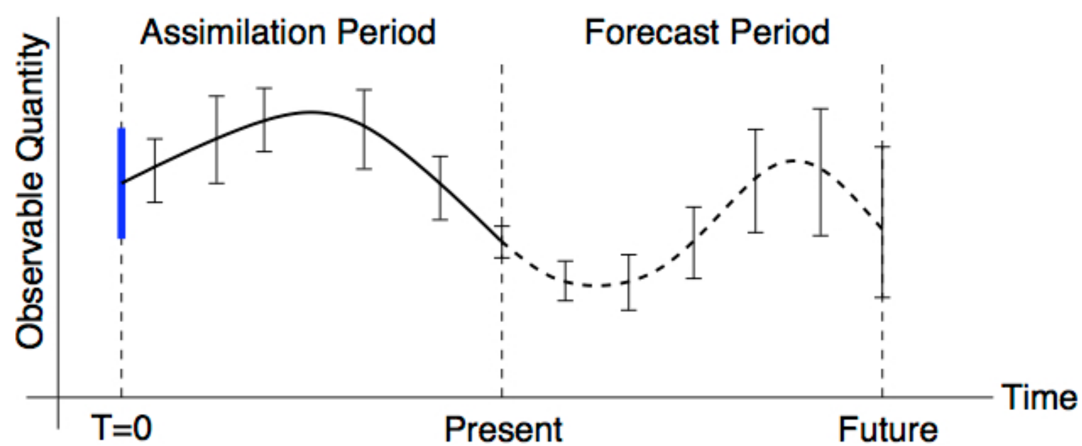
• Sensors positions may be uncertain; e.g., weather balloons, ocean buoys.



Goal:

• Assimilate data to quantify uncertain initial conditions and parameters;

• Make predictions with quantified uncertainties.



Equations of Atmospheric Physics

Conservation Relations:

$$\boxed{\frac{\partial \rho}{\partial t} + \frac{\partial \phi}{\partial x} = \text{Sources - Sinks}}$$

Mass $$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho v) = 0$$

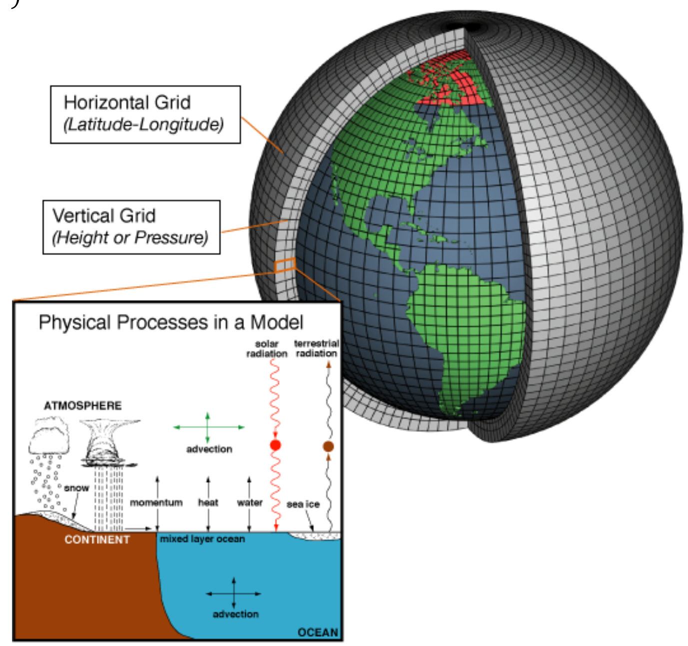
Momentum $$\frac{\partial v}{\partial t} = -v \cdot \nabla v - \frac{1}{\rho}\nabla p - g\hat{k} - 2\Omega \times v$$

Energy $$\rho c_V \frac{\partial T}{\partial t} + p\nabla \cdot v = -\nabla \cdot F + \nabla \cdot (k \nabla T) + \rho \dot{q}(T, p, \rho)$$

$$p = \rho R T$$

Water $$\frac{\partial m_j}{\partial t} = -v \cdot \nabla m_j + S_{m_j}(T, m_j, \chi_j, \rho)\ ,\ j = 1, 2, 3,$$

Aerosol $$\frac{\partial \chi_j}{\partial t} = -v \cdot \nabla \chi_j + S_{\chi_j}(T, \chi_j, \rho)\ ,\ j = 1, \cdots, J,$$



Constitutive Closure Relations: e.g.,

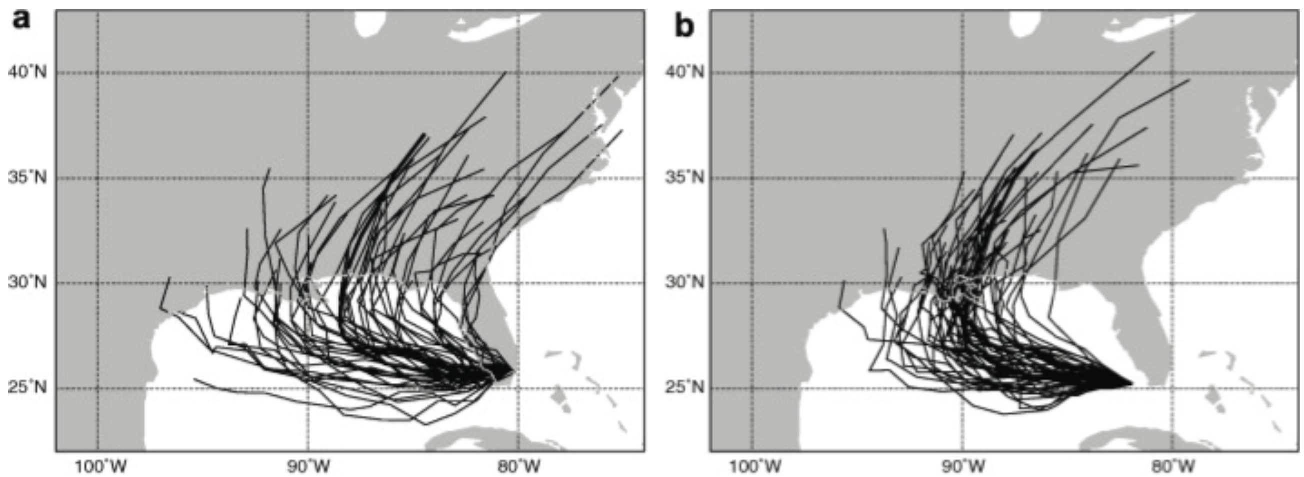
$$S_{m_2} = S_1 + S_2 + S_3 - S_4$$

where

$$S_1 = \bar{\rho}(m_2 - m_2^*)^2 \left[\underline{1.2 \times 10^{-4}} + \left(\underline{1.569 \times 10^{-12}} \frac{n_r}{d_0(m_2 - m_2^*)}\right)\right]^{-1}$$

Ensemble Predictions

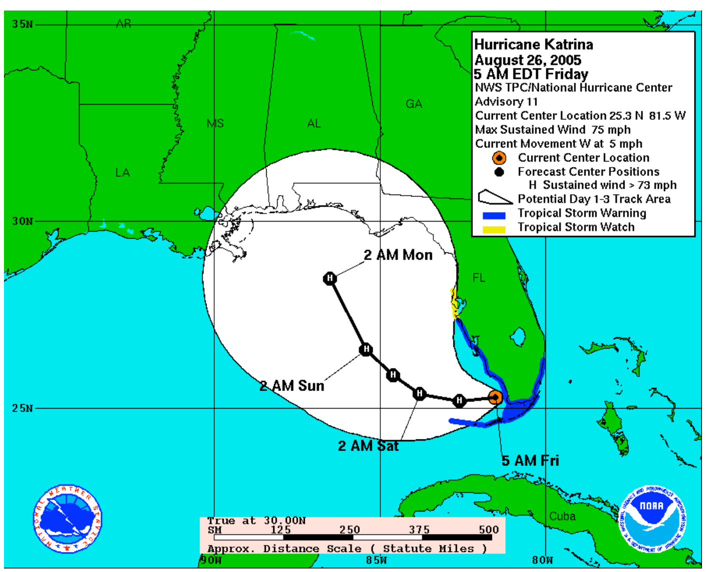
Ensemble Predictions:



00 UTC on August 26, 2005

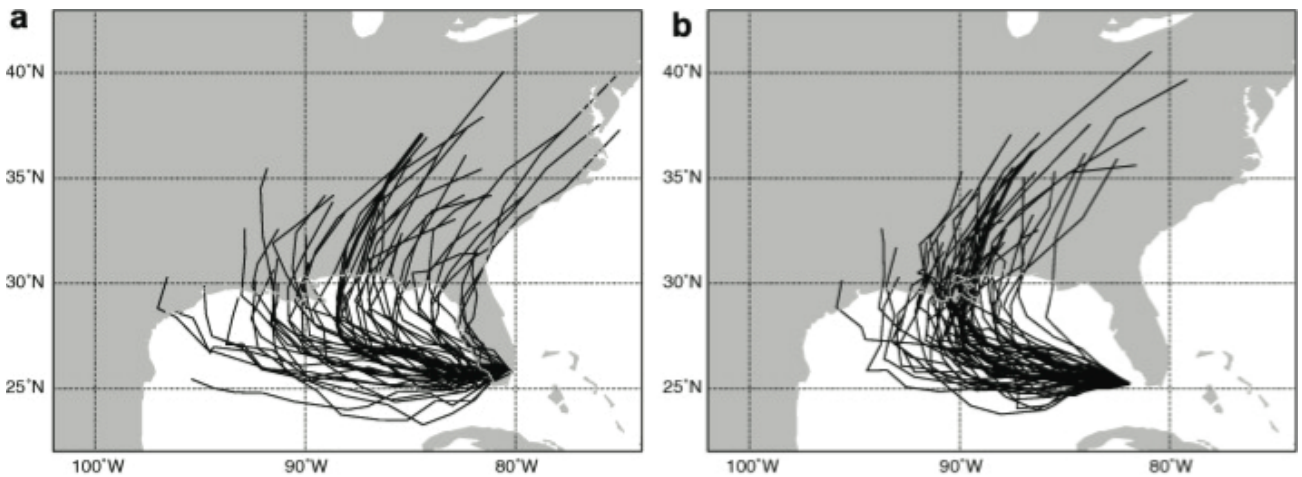
12 UTC on August 26, 2005

Cone of Uncertainty:



Ensemble Predictions

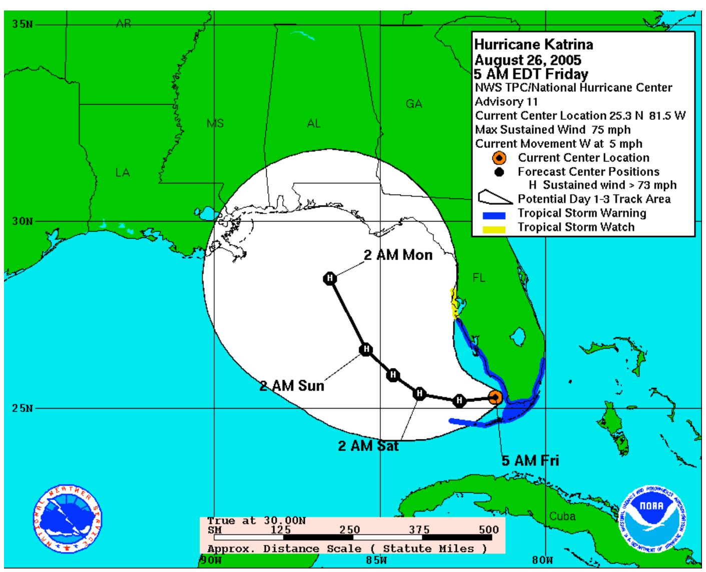
Ensemble Predictions:



00 UTC on August 26, 2005

12 UTC on August 26, 2005

Cone of Uncertainty:



General Questions:

- What is expected rainfall on July 31?
- What are average high and low temperatures?
- Note: Quantities are statistical in nature.

Example 2: HIV Model for Characterization and Control Regimes

HIV Model:

$$\dot{T}_1 = \lambda_1 - d_1 T_1 - (1-\varepsilon)k_1 V T_1$$

$$\dot{T}_2 = \lambda_2 - d_2 T_2 - (1-f\varepsilon)k_2 V T_2$$

$$\dot{T}_1^* = (1-\varepsilon)k_1 V T_1 - \delta T_1^* - m_1 E T_1^*$$

$$\dot{T}_2^* = (1-f\varepsilon)k_2 V T_2 - \delta T_2^* - m_2 E T_2^*$$

$$\dot{V} = N_T \delta (T_1^* + T_2^*) - cV - [(1-\varepsilon)\rho_1 k_1 T_1 + (1-f\varepsilon)\rho_2 k_2 T_2]V$$

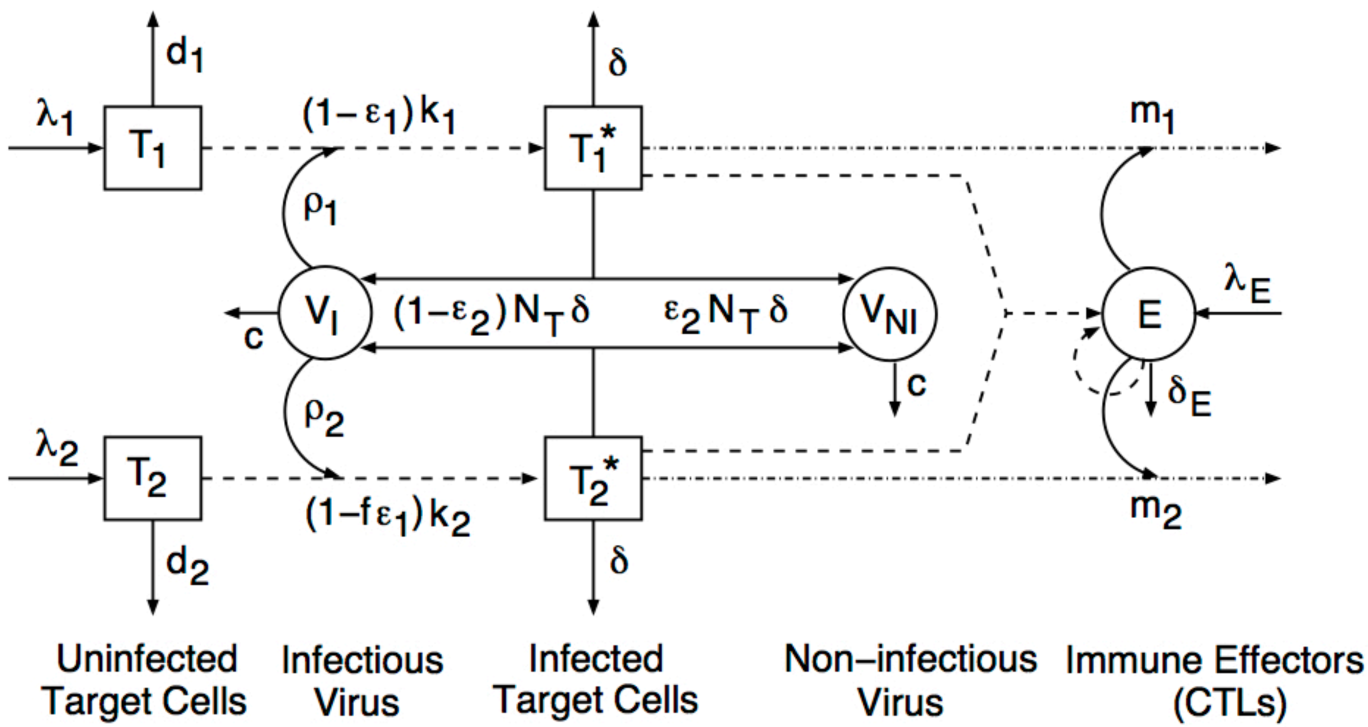
$$\dot{E} = \lambda_E + \frac{b_E(T_1^* + T_2^*)}{T_1^* + T_2^* + K_b}E - \frac{d_E(T_1^* + T_2^*)}{T_1^* + T_2^* + K_d}E - \delta_E E$$

Notes: 21 parameters

[Adams, Banks et al., 2005, 2007]

Notation: $\dot{E} \equiv \frac{dE}{dt}$

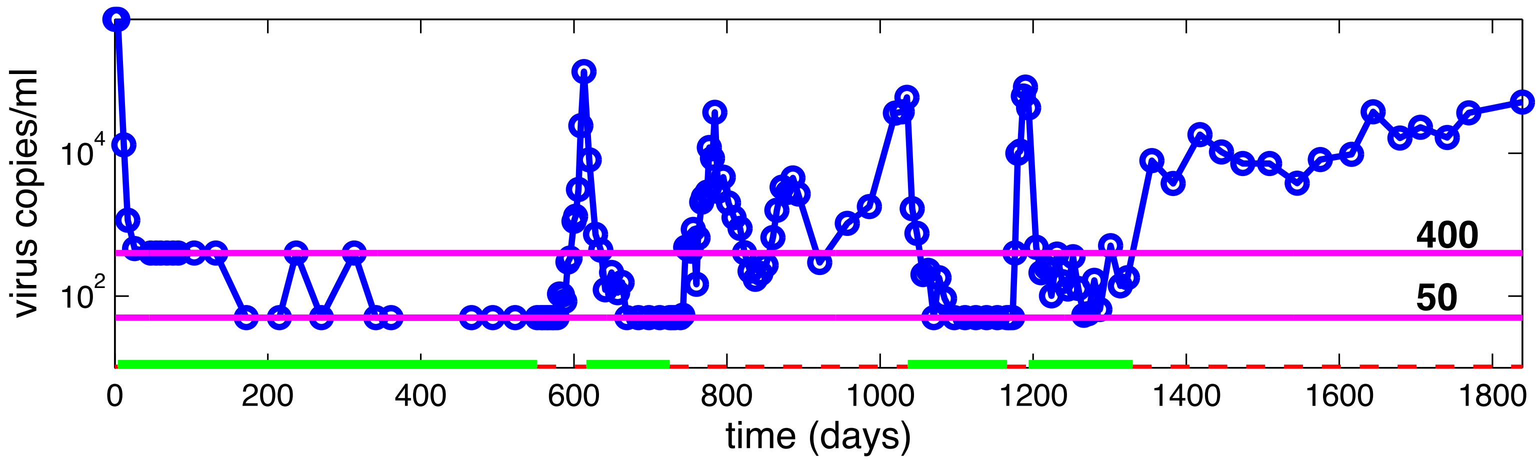
Compartments:



Example: HIV Model for Characterization and Treatment Regimes

HIV Model: Several sources of uncertainty including viral measurement techniques

Example: Upper and lower limits to assay sensitivity



UQ Questions:

• What are the uncertainties in parameters that cannot be directly measured?

• What is optimal treatment regime that is "safe" for patient?

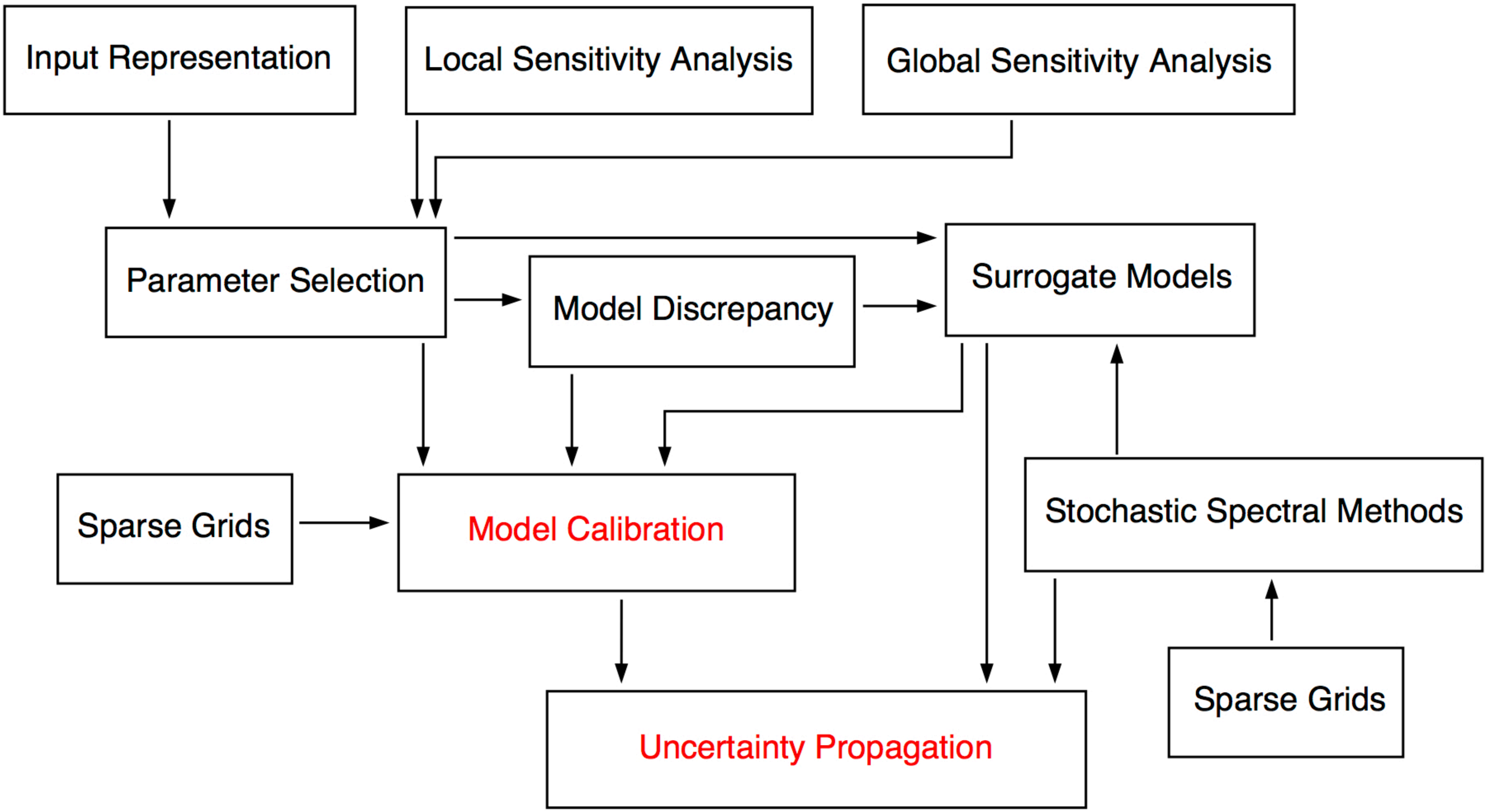
• What is expected viral load? Issue: very often requires high-dimensional integration!

• e.g., $\mathbb{E}[V(t)] = \int_{\mathbb{R}^{21}} V(t,q)\rho(q)dq$

Experimental results are believed by everyone, except for the person who ran the experiment, source anonymous, quoted by Max Gunzburger, Florida State University.

Steps in Uncertainty Quantification

Note: Uncertainty quantification requires synergy between statistics, mathematics and application area.



Model Calibration and Uncertainty Propagation

Sources of Uncertainty:

- Model
- Parameters
- Sensor measurements
- Initial conditions

Strategy:

- Quantify uncertainty in parameters
- Propagate uncertainty through model

Example: HIV model

$$\dot{T}_1 = \underline{\lambda_1} - \underline{d_1} T_1 - (1-\varepsilon) k_1 V T_1$$

$$\dot{T}_2 = \lambda_2 - d_2 T_2 - (1 - f\varepsilon) k_2 V T_2$$

$$\dot{T}_1^* = (1-\varepsilon) k_1 V T_1 - \underline{\delta} T_1^* - m_1 E T_1^*$$

$$\dot{T}_2^* = (1 - f\varepsilon) \underline{k_2} V T_2 - \underline{\delta} T_2^* - m_2 E T_2^*$$

$$\dot{V} = N_T \underline{\delta} (T_1^* + T_2^*) - cV - [(1-\varepsilon)\rho_1 k_1 T_1 + (1 - f\varepsilon)\rho_2 k_2 T_2] V$$

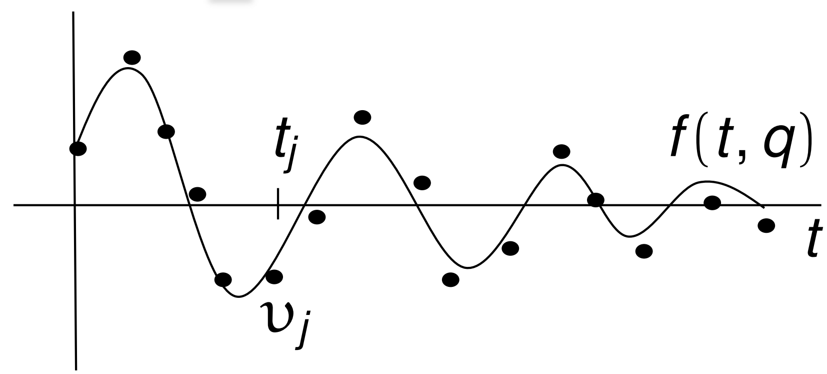
$$\dot{E} = \lambda_E + \frac{\underline{b_E}(T_1^* + T_2^*)}{T_1^* + T_2^* + \underline{K_b}} E - \frac{d_E (T_1^* + T_2^*)}{T_1^* + T_2^* + K_d} E - \delta_E E$$

Parameters: Reduced set

$$q = [b_E, \delta, d_1, k_2, \lambda_1, K_b]$$

Point Estimates: Ordinary least squares – see talk by Alun Lloyd

$$q^0 = \arg\min_q \frac{1}{2} \sum_{j=1}^{N} [\upsilon_j - f(t_j, q)]^2$$



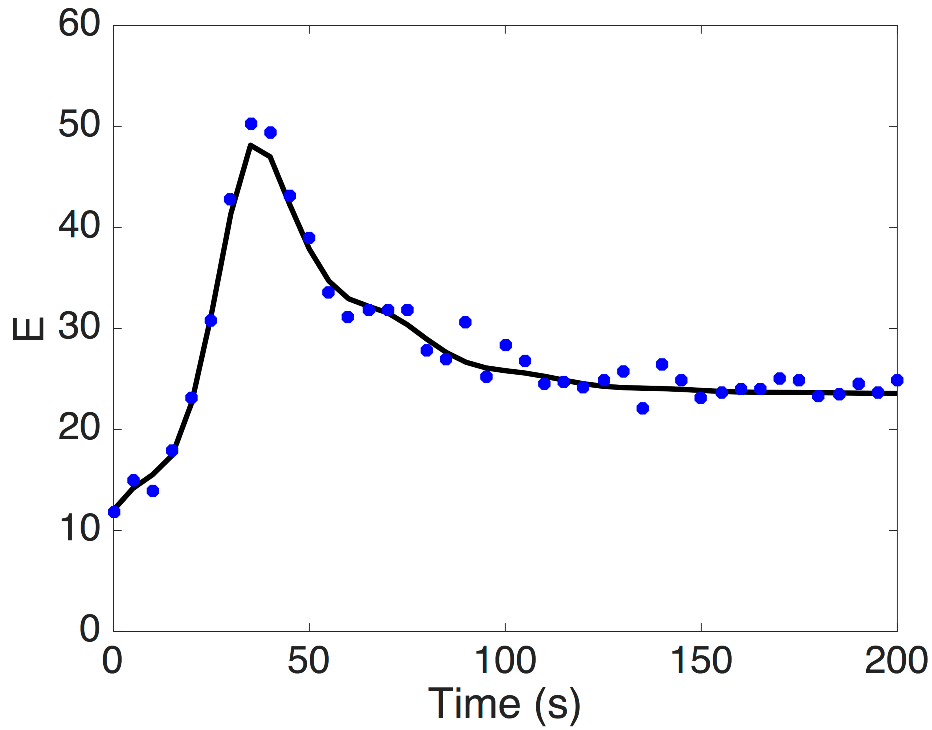
Note: Scaling critical since parameter values vary by 8 orders of magnitude.

Model Calibration and Predictions

Optimization Results:

b_E	δ	d_1	k_2	λ_1	K_b
0.30	0.68	9.1×10^{-3}	1.22×10^{-4}	9.95×10^{3}	88.5

Data and Prediction of Immune Effector Response E:



Note: Point estimates but no quantification of uncertainty in:

- Model
- Parameters
- Data

Goals:

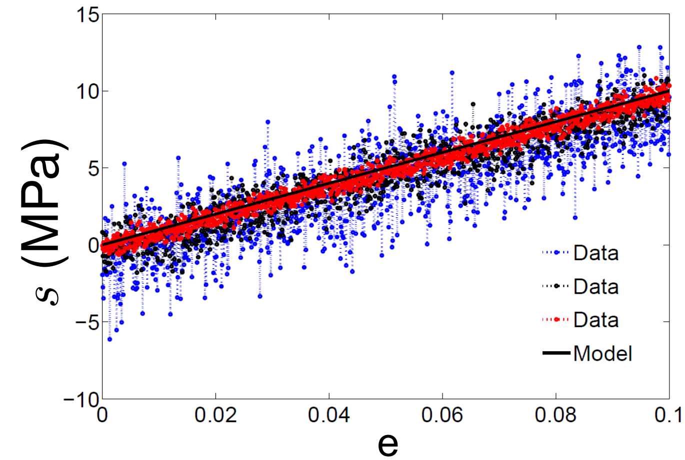
- Replace point estimates with distributions.
- Construct credible and prediction intervals.
- Natural in a Bayesian framework

Bayesian Inference: More General Model

Example: Displacement-force relation (Hooke's Law)

$$s_i = Ee_i + \varepsilon_i \ , \ i = 1, \ldots, N$$

$$\varepsilon_i \sim N(0, \sigma^2)$$



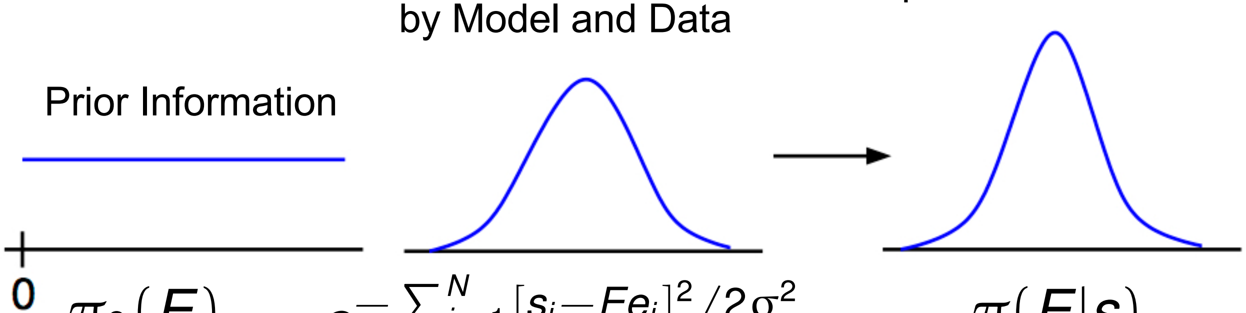
Parameter: Stiffness E

Strategy: Use model fit to data to update prior information

Information Provided by Model and Data

Updated Information

Prior Information



$\pi_0(E)$ $\quad e^{-\sum_{i=1}^N [s_i - Ee_i]^2/2\sigma^2}$ $\quad \pi(E|s)$

Data

Model

Non-normalized Bayes' Relation:

$$\pi(E|s) = e^{-\sum_{i=1}^N [s_i - Ee_i]^2/2\sigma^2} \pi_0(E)$$

Bayesian Inference

Bayes' Relation: Specifies posterior in terms of likelihood and prior

Likelihood: $e^{-\sum_{i=1}^{N}[s_i - Ee_i]^2/2\sigma^2}$, $q = E$

$\upsilon = [s_1, \dots, s_N]$

Posterior Distribution

$$\boxed{\pi(q|\upsilon) = \frac{\pi(\upsilon|q)\pi_0(q)}{\int_{\mathbb{R}^p} \pi(\upsilon|q)\pi_0(q)dq}}$$

Prior Distribution

Normalization Constant

- Prior Distribution: Quantifies prior knowledge of parameter values
- Likelihood: Probability of observing a data given set of parameter values.
- Posterior Distribution: Conditional distribution of parameters given observed data.

Problem: Can require high-dimensional integration

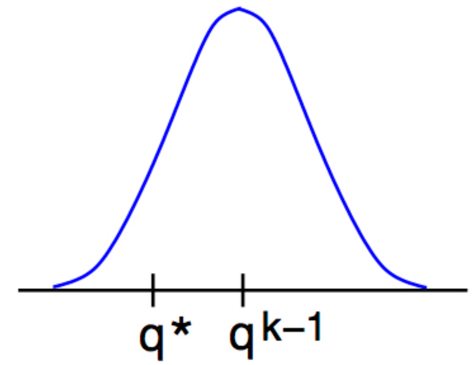
• e.g., HIV Model: p = 6 - 23!

• Solution: Sampling-based Markov Chain Monte Carlo (MCMC) algorithms.

• Metropolis algorithms first used by nuclear physicists during Manhattan Project in 1940's to understand particle movement underlying first atomic bomb.

Delayed Rejection Adaptive Metropolis (DRAM)

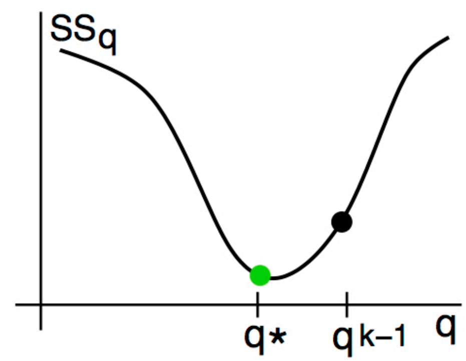
Algorithm: [Haario et al., 2006] – MATLAB, Python

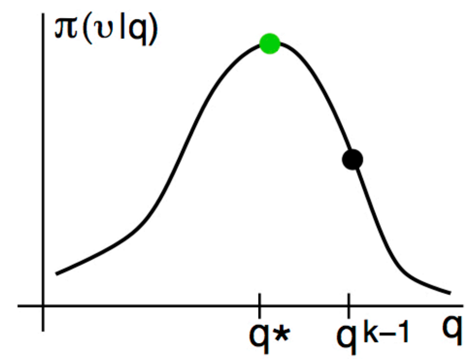


1. Determine $q^0 = \arg\min_q \sum_{i=1}^{N} [\upsilon_i - f(t_i, q)]^2$
2. For $k = 1, \cdots, M$
   (a) Construct candidate $q^* \sim N(q^{k-1}, V)$
   (b) Compute likelihood

$$SS_{q^*} = \sum_{i=1}^{N} [\upsilon_i - f(t_i, q^*)]^2$$

$$\pi(\upsilon|q) = \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-SS_q/2\sigma^2}$$

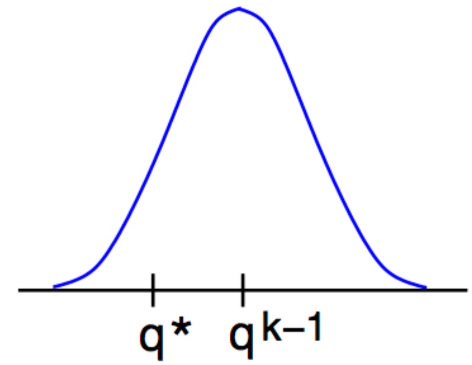




   (c) Accept q^* with probability dictated by likelihood

Delayed Rejection Adaptive Metropolis (DRAM)

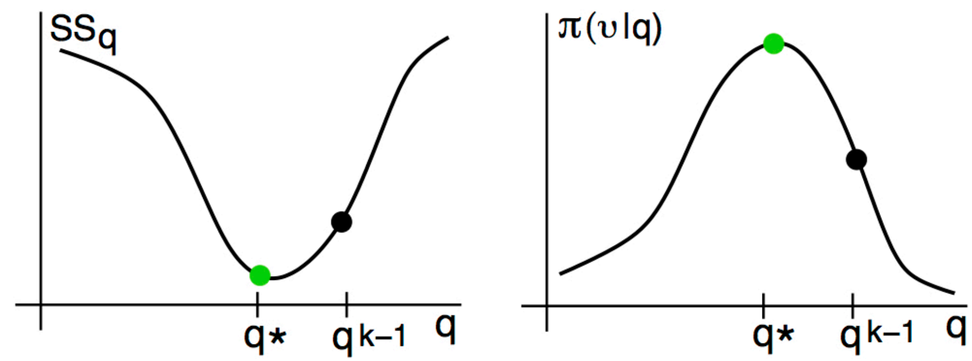
Algorithm: [Haario et al., 2006] – MATLAB, Python



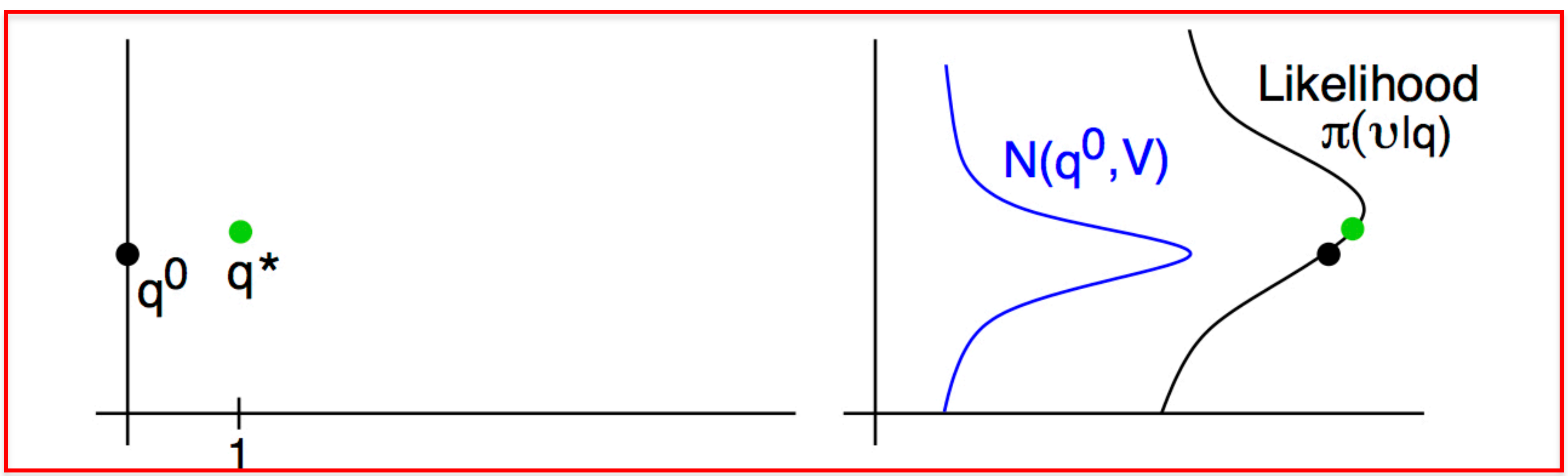
1. Determine $q^0 = \arg\min_q \sum_{i=1}^{N} [\upsilon_i - f(t_i, q)]^2$
2. For $k = 1, \cdots, M$
   (a) Construct candidate $q^* \sim N(q^{k-1}, V)$
   (b) Compute likelihood

$$SS_{q^*} = \sum_{i=1}^{N} [\upsilon_i - f(t_i, q^*)]^2$$

$$\pi(\upsilon|q) = \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-SS_q/2\sigma^2}$$

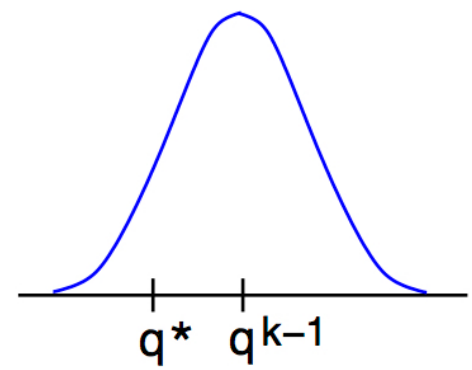


   (c) Accept q^* with probability dictated by likelihood



Delayed Rejection Adaptive Metropolis (DRAM)

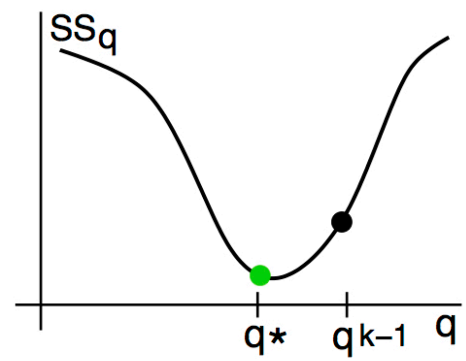
Algorithm: [Haario et al., 2006] – MATLAB, Python

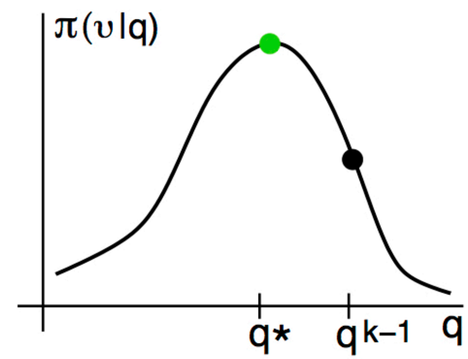


1. Determine $q^0 = \arg\min_q \sum_{i=1}^{N} [\upsilon_i - f(t_i, q)]^2$
2. For $k = 1, \cdots, M$
   (a) Construct candidate $q^* \sim N(q^{k-1}, V)$
   (b) Compute likelihood

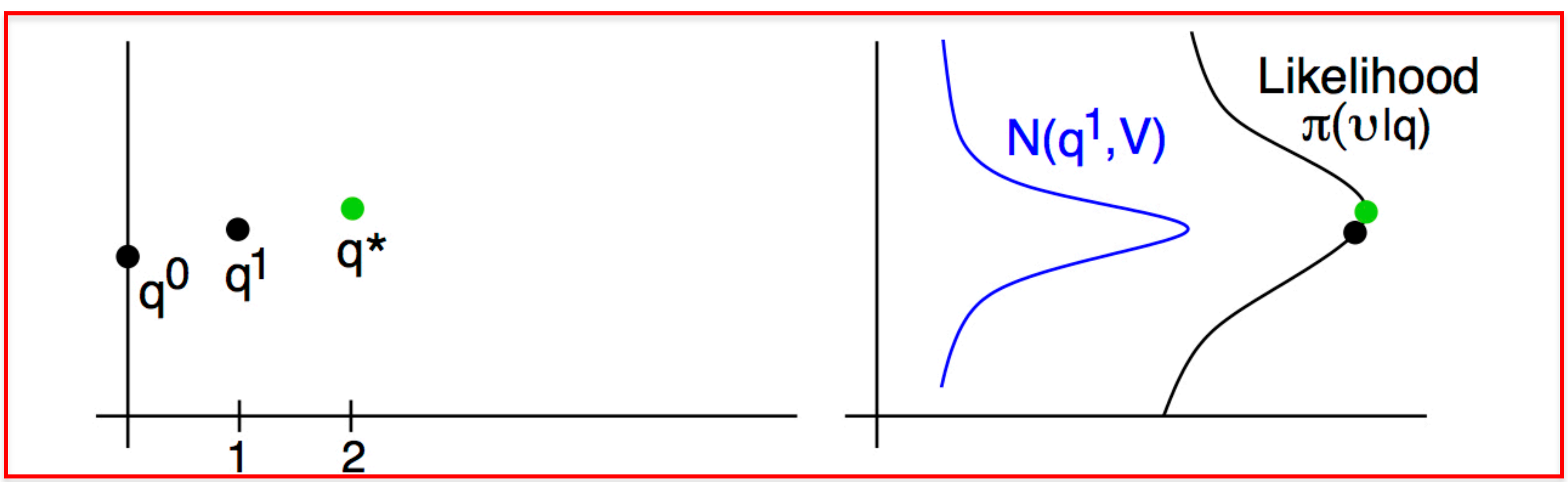
$$SS_{q^*} = \sum_{i=1}^{N} [\upsilon_i - f(t_i, q^*)]^2$$

$$\pi(\upsilon|q) = \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-SS_q/2\sigma^2}$$



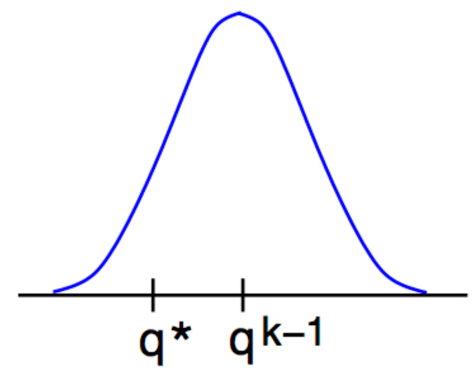


   (c) Accept q^* with probability dictated by likelihood



Delayed Rejection Adaptive Metropolis (DRAM)

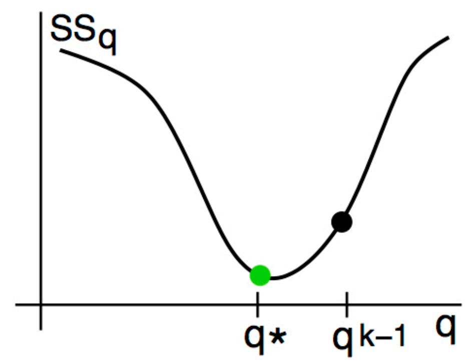
Algorithm: [Haario et al., 2006] – MATLAB, Python

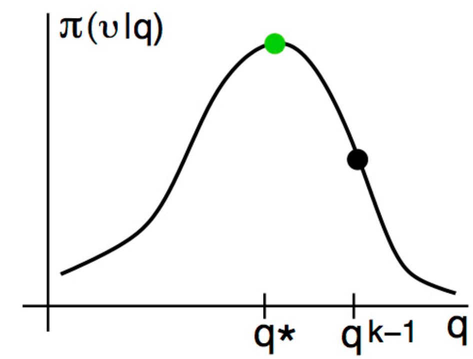


1. Determine $q^0 = \arg\min_q \sum_{i=1}^{N} [\upsilon_i - f(t_i, q)]^2$
2. For $k = 1, \cdots, M$
   (a) Construct candidate $q^* \sim N(q^{k-1}, V)$
   (b) Compute likelihood

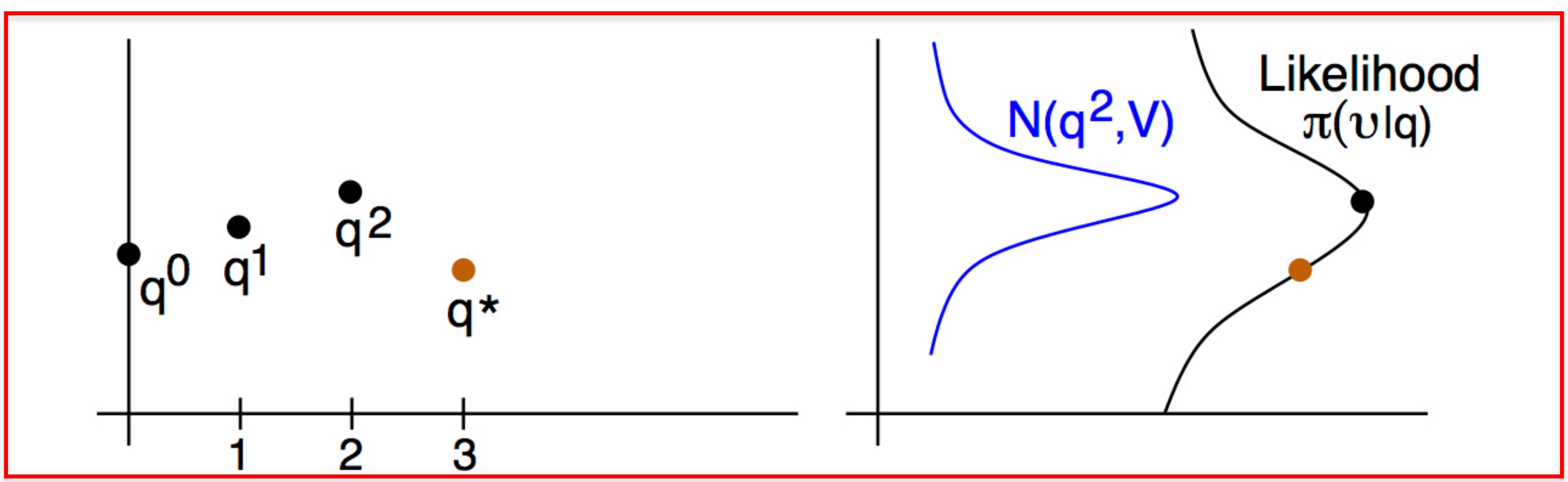
$$SS_{q^*} = \sum_{i=1}^{N} [\upsilon_i - f(t_i, q^*)]^2$$

$$\pi(\upsilon|q) = \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-SS_q/2\sigma^2}$$





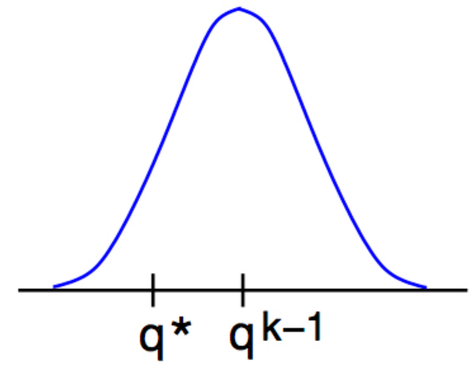
   (c) Accept q^* with probability dictated by likelihood



Delayed Rejection Adaptive Metropolis (DRAM)

Algorithm: [Haario et al., 2006] – MATLAB, Python

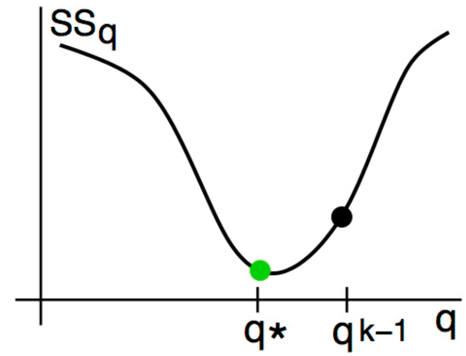
1. Determine $q^0 = \arg\min_q \sum_{i=1}^{N} [\upsilon_i - f(t_i, q)]^2$
2. For $k = 1, \cdots, M$
   (a) Construct candidate $q^* \sim N(q^{k-1}, V)$

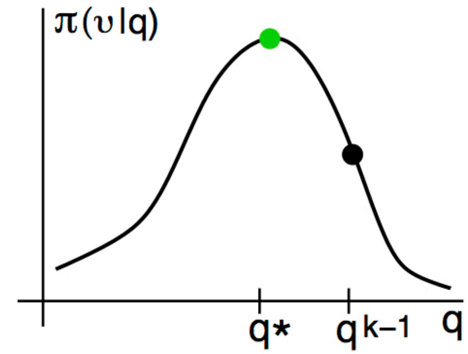


   (b) Compute likelihood

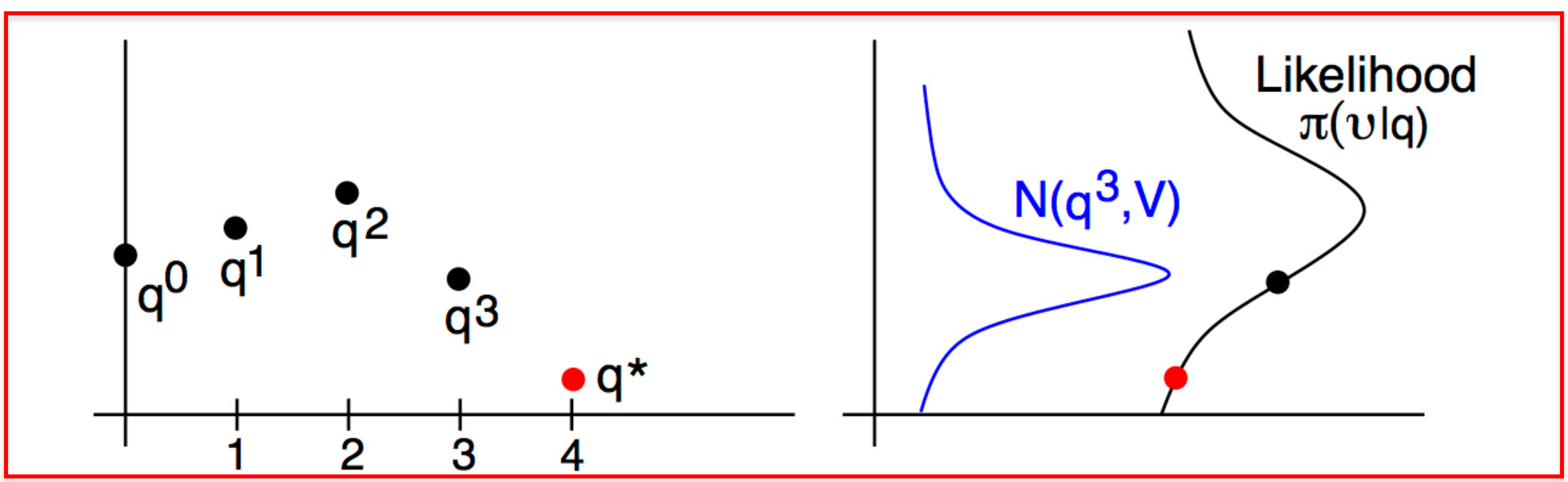
$$SS_{q^*} = \sum_{i=1}^{N} [\upsilon_i - f(t_i, q^*)]^2$$

$$\pi(\upsilon|q) = \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-SS_q/2\sigma^2}$$





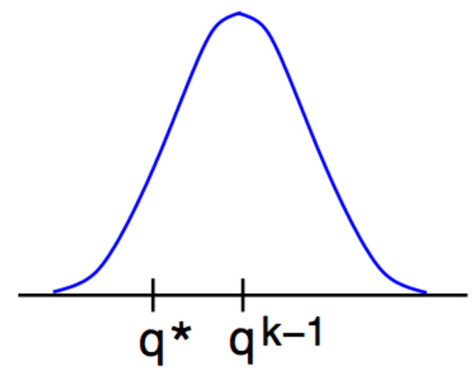
   (c) Accept q^* with probability dictated by likelihood



Delayed Rejection Adaptive Metropolis (DRAM)

Algorithm: [Haario et al., 2006] – MATLAB, Python

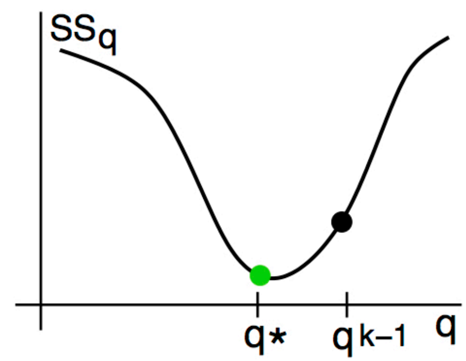
1. Determine $q^0 = \arg\min_q \sum_{i=1}^{N} [\upsilon_i - f(t_i, q)]^2$

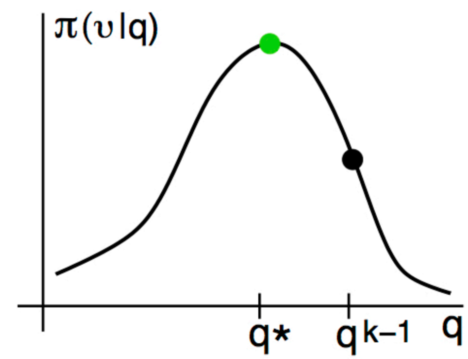


2. For $k = 1, \cdots, M$
   (a) Construct candidate $q^* \sim N(q^{k-1}, V)$
   (b) Compute likelihood

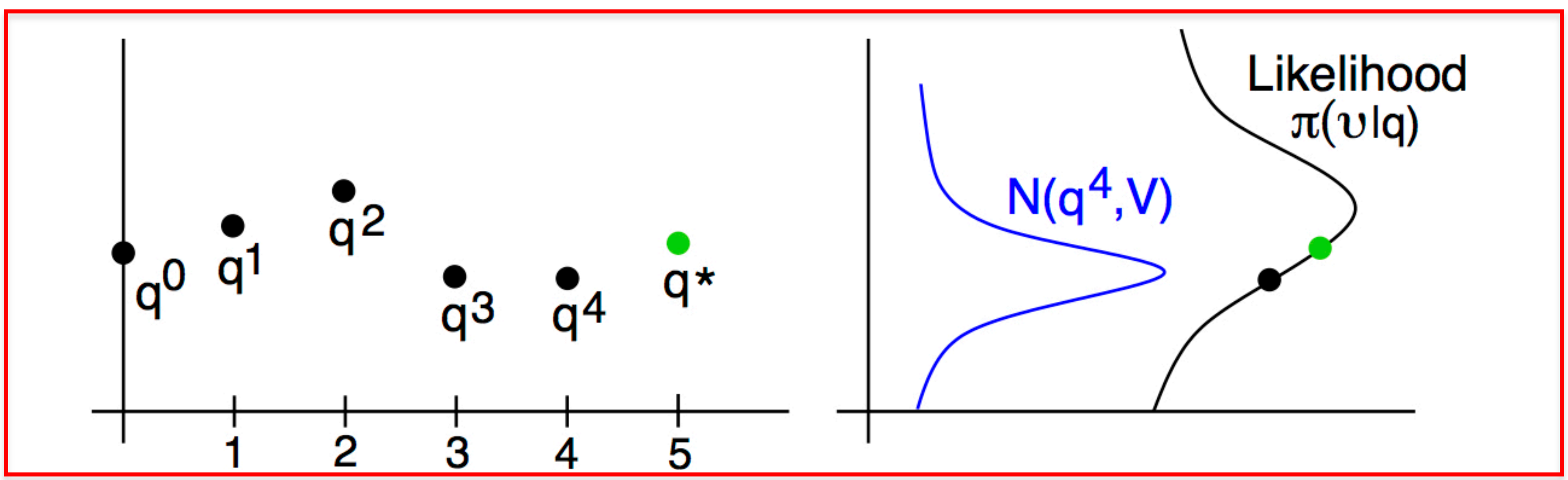
$$SS_{q^*} = \sum_{i=1}^{N} [\upsilon_i - f(t_i, q^*)]^2$$

$$\pi(\upsilon|q) = \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-SS_q/2\sigma^2}$$





   (c) Accept q^* with probability dictated by likelihood



Bayesian Model Calibration – HIV Example

Model:

$$\dot{T}_1 = \lambda_1 - d_1 T_1 - (1-\varepsilon) k_1 V T_1$$

$$\dot{T}_2 = \lambda_2 - d_2 T_2 - (1-f\varepsilon) k_2 V T_2$$

$$\dot{T}_1^* = (1-\varepsilon) k_1 V T_1 - \delta T_1^* - m_1 E T_1^*$$

$$\dot{T}_2^* = (1-f\varepsilon) k_2 V T_2 - \delta T_2^* - m_2 E T_2^*$$

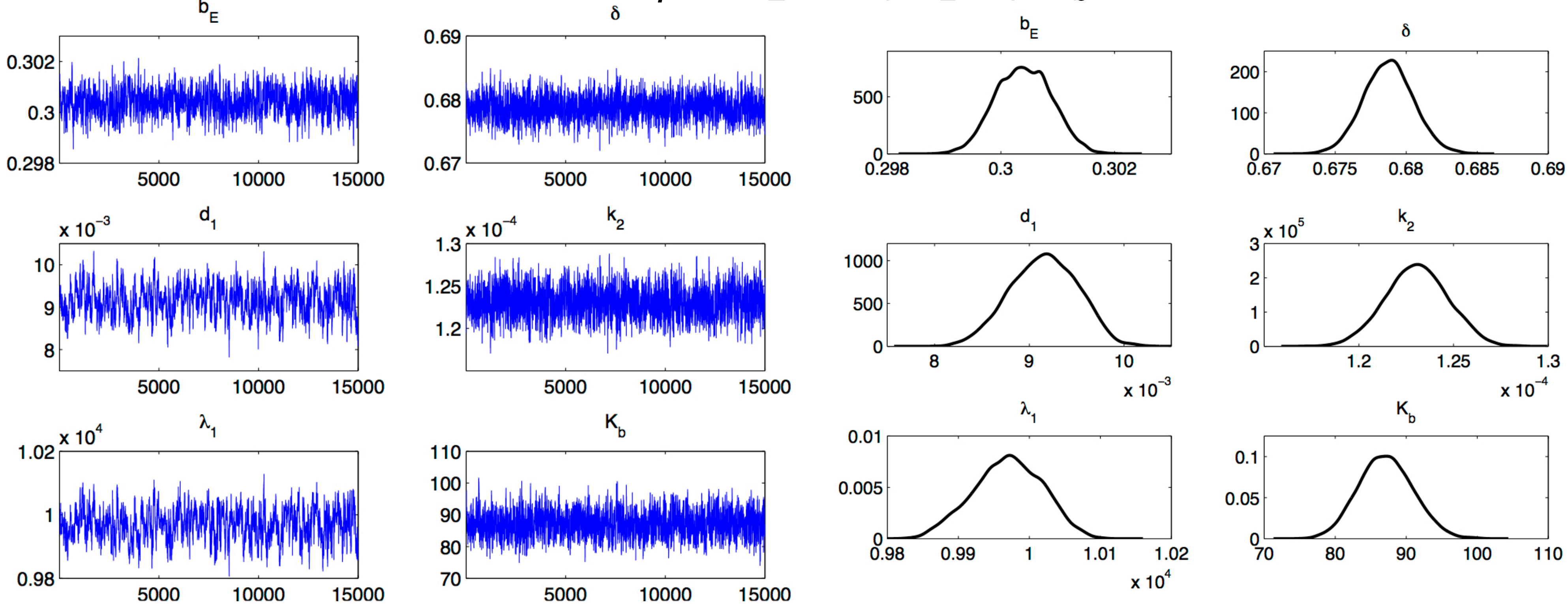
$$\dot{V} = N_T \delta (T_1^* + T_2^*) - cV - [(1-\varepsilon)\rho_1 k_1 T_1 + (1-f\varepsilon)\rho_2 k_2 T_2] V$$

$$\dot{E} = \lambda_E + \frac{b_E (T_1^* + T_2^*)}{T_1^* + T_2^* + K_b} E - \frac{d_E (T_1^* + T_2^*)}{T_1^* + T_2^* + K_d} E - \delta_E E$$

Verification: Why do we trust results??

- Compare results from different algorithms; e.g., DRAM and Gibbs

Parameter Chains and Densities: $q = [b_E, \delta, d_1, k_2, \lambda_1, K_b]$



Propagation of Uncertainty in Models – HIV Example

HIV Example:

$$\dot{T}_1 = \lambda_1 - d_1 T_1 - (1-\varepsilon)k_1 V T_1$$

$$\dot{T}_2 = \lambda_2 - d_2 T_2 - (1-f\varepsilon)k_2 V T_2$$

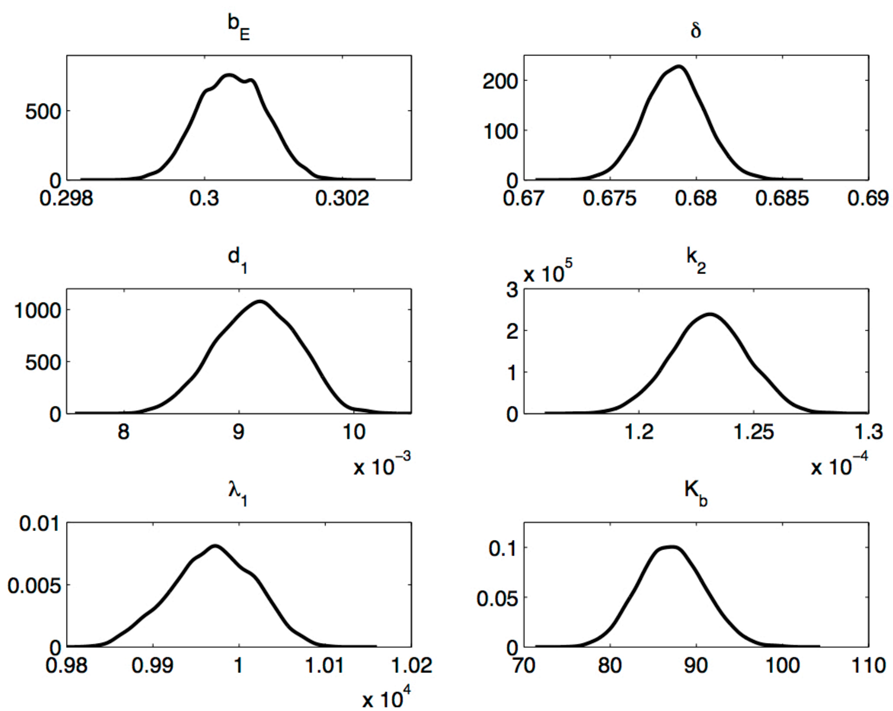
$$\dot{T}_1^* = (1-\varepsilon)k_1 V T_1 - \delta T_1^* - m_1 E T_1^*$$

$$\dot{T}_2^* = (1-f\varepsilon)k_2 V T_2 - \delta T_2^* - m_2 E T_2^*$$

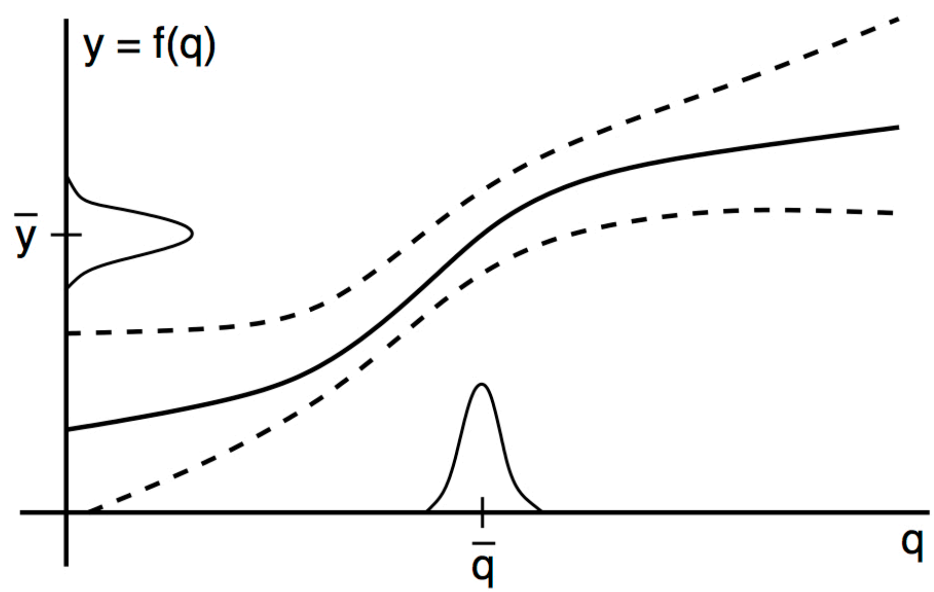
$$\dot{V} = N_T \delta(T_1^* + T_2^*) - cV - [(1-\varepsilon)\rho_1 k_1 T_1 + (1-f\varepsilon)\rho_2 k_2 T_2]V$$

$$\dot{E} = \lambda_E + \frac{b_E(T_1^* + T_2^*)}{T_1^* + T_2^* + K_b}E - \frac{d_E(T_1^* + T_2^*)}{T_1^* + T_2^* + K_d}E - \delta_E E$$

Parameter Densities:

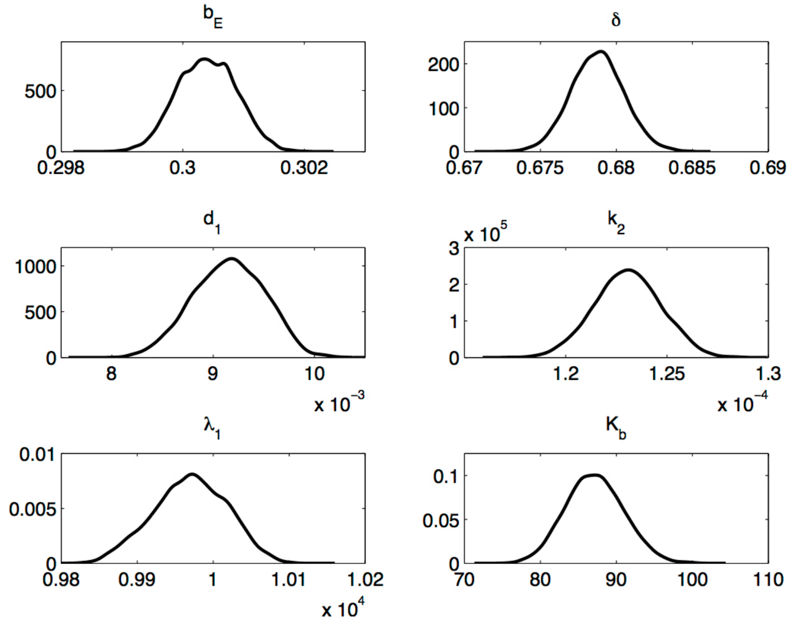


General Strategy



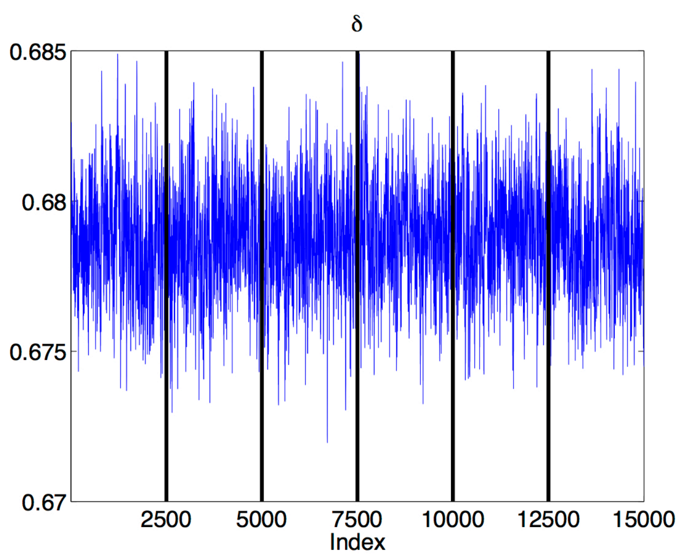
Propagation of Uncertainty in Models – HIV Example

Parameter Densities:

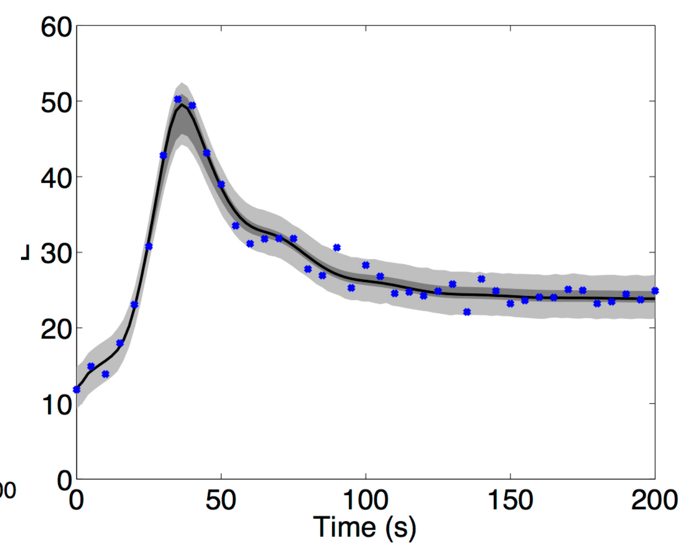


Techniques:

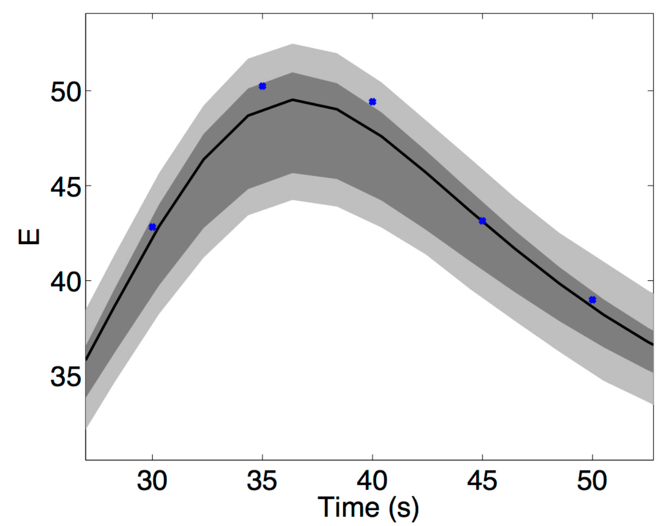
- Sample from parameter densities to construct prediction intervals for QoI.
- Slow convergence rate $\mathcal{O}(1/\sqrt{M})$
- 100-fold more evaluations required to gain additional place of accuracy.
- Significant numerical analysis used to efficiently propagate densities.



Samples from Chain



Data, Credible Intervals and Prediction Intervals



Non-Gaussian Credible and Prediction Intervals

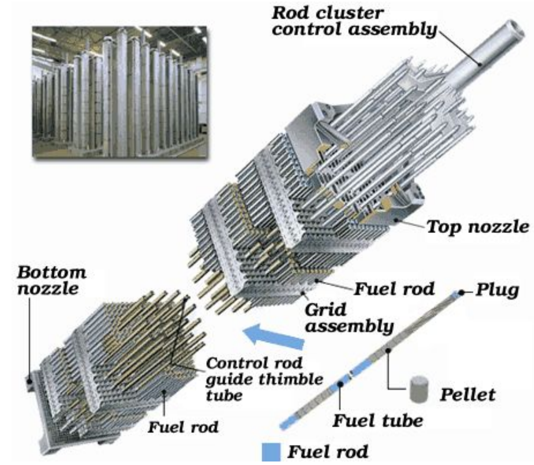
Use of Prediction Intervals: Nuclear Power Plant Design

Subchannel Code (COBRA-TF): numerous closure relations, ~70 parameters

e.g., Dittus—Boelter Relation

$$Nu = 0.023 Re^{0.8} Pr^{0.4}$$

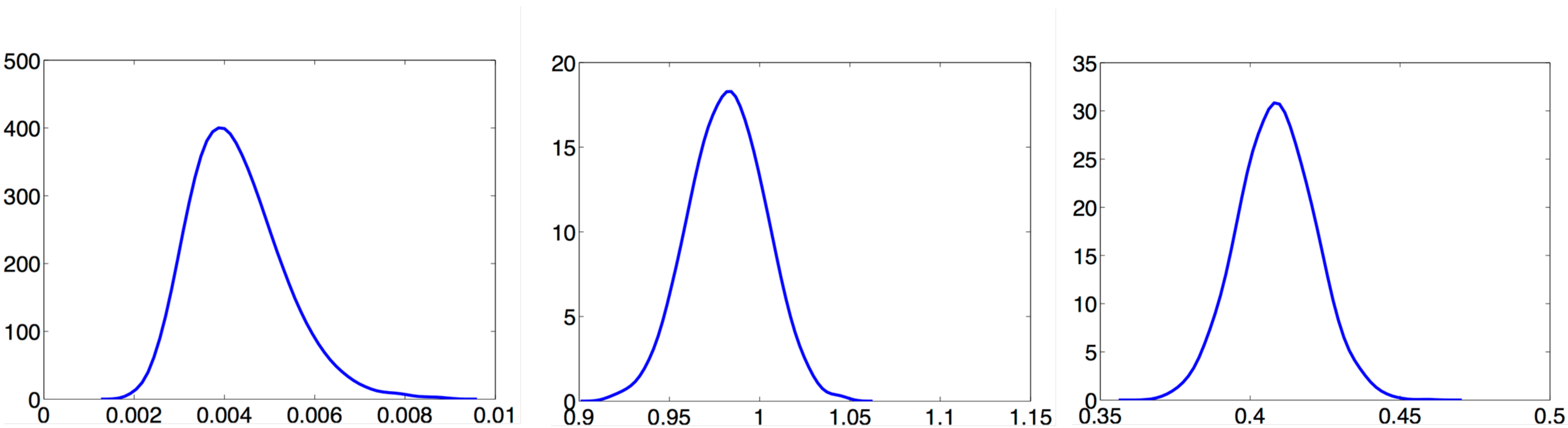
Nu: Nusselt number
Re: Reynolds number
Pr: Prandtl number



Industry Standard: Employ conservative, uniform, bounds

i.e., [0, 0.046], [0, 1.6], [0,0.8]

Bayesian Analysis: Employ conservative bounds as priors

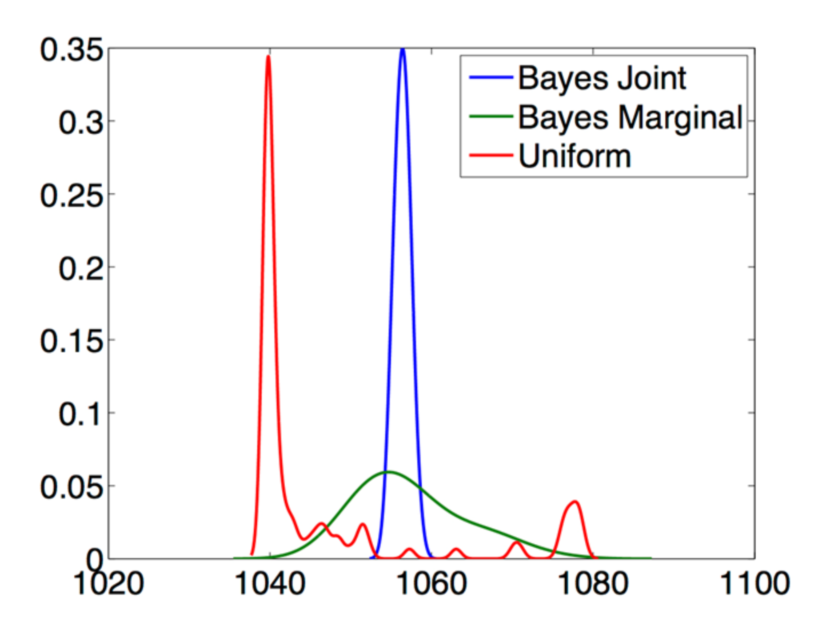


$2\sigma \approx 0.0035$ $\quad$ $2\sigma \approx 0.06$ $\quad$ $2\sigma \approx 0.03$

Note: Substantial reduction in parameter uncertainty

Use of Prediction Intervals: Nuclear Power Plant Design

Strategy: Propagate parameter uncertainties through COBRA-TF to determine uncertainty in maximum fuel temperature



Notes:

• Temperature uncertainty reduced from 40 degrees to 5 degrees

• Can run plant 20 degrees hotter, which significantly improves efficiency

Ramification: Savings of 10 billion dollars per year for US power plants

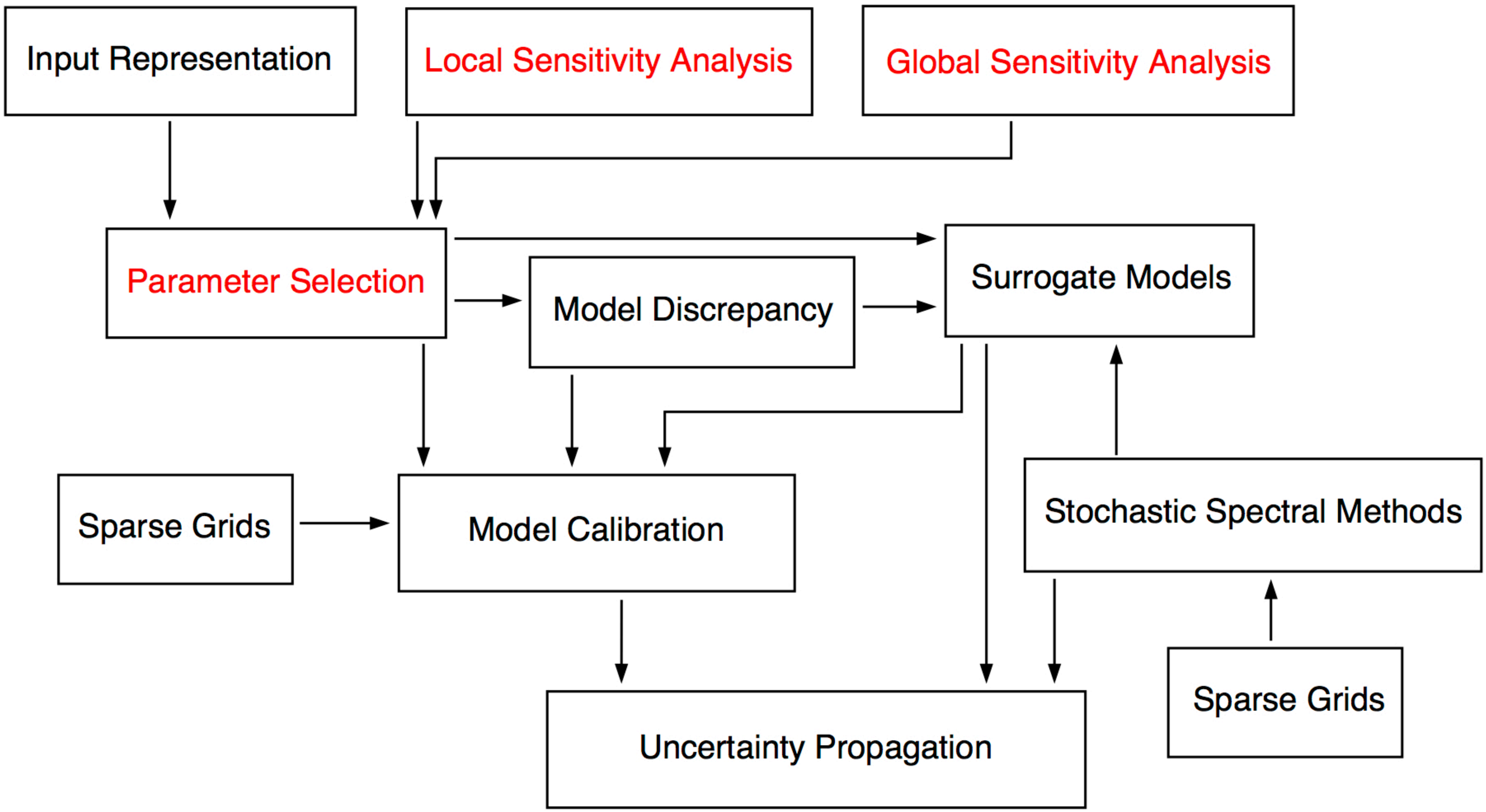
Issues:

• We considered only one of many physical relations

• Nuclear regulatory commission takes years to change requirements and codes

Good News: We are now working with Westinghouse to reduce uncertainties.

Steps in Uncertainty Quantification



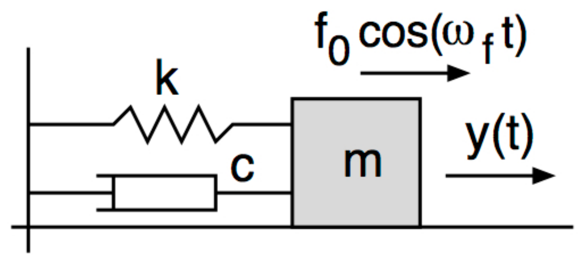
Parameter Selection: Required for models with unidentifiable or noninfluential inputs

- e.g., HIV and SIR model

Parameter Selection Techniques

First Issue: Parameters often not *identifiable* in the sense that they are uniquely determined by the data.

Example: Spring model



$$\underline{m}\frac{d^2z}{dt^2} + \underline{c}\frac{dz}{dt} + \underline{k}z = \underline{f_0}\cos(\omega_F t)$$

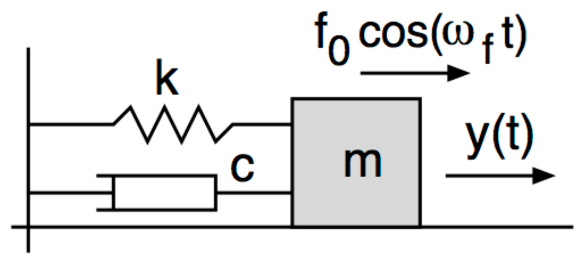
$$z(0) = z_0 \; , \; \frac{dz}{dt}(0) = z_1$$

Problem: Parameters $q = [m, c, k, f_0]$ and $q = \left[1, \frac{c}{m}, \frac{k}{m}, \frac{f_0}{m}\right]$ yield same displacements

Parameter Selection Techniques

First Issue: Parameters often *not identifiable* in the sense that they are uniquely determined by the data.

Example: Spring model



$$m\frac{d^2z}{dt^2}+c\frac{dz}{dt}+kz=f_0\cos(\omega_F t)$$

$$z(0)=z_0\ ,\ \frac{dz}{dt}(0)=z_1$$

Problem: Parameters $q=[m,c,k,f_0]$ and $q=\left[1,\frac{c}{m},\frac{k}{m},\frac{f_0}{m}\right]$ yield same displacements

Solution: Reformulate problem as

$$\frac{d^2z}{dt^2}+C\frac{dz}{dt}+Kz=F_0\cos(\omega_F t)$$

$$z(0)=z_0\ ,\ \frac{dz}{dt}(0)=z_1$$

where $C=\frac{c}{m}$, $K=\frac{k}{m}$ and $F_0=\frac{f_0}{m}$

Techniques for General Models:

- Linear algebra analysis;
  - e.g., SVD or QR algorithms
- Sensitivity analysis
- Active Subspaces

Global Sensitivity Analysis

Example: Portfolio model

$$Y = c_1 Q_1 + c_2 Q_2$$

Note:

- Q_1 and Q_2 represent hedged portfolios
- c_1 and c_2 amounts invested in each portfolio

Take

$$c_1 = 2 \ , \ c_2 = 1$$

$$Q_1 \sim N(0, 1)$$

$$Q_2 \sim N(0, 9)$$

Local Sensitivities:

$$\frac{\partial Y}{\partial Q_1} = 2 \quad , \quad \frac{\partial Y}{\partial Q_2} = 1$$

Conclusion: Investment is more sensitive to Portfolio 1 than to Portfolio 2

Limitations:

- Does not accommodate potential uncertainty in parameters.
- Sensitive to units and magnitudes of parameters.
- See talk by Pierre Gremaud.

Global Sensitivity Analysis

Example: Portfolio model

$$Y = c_1 Q_1 + c_2 Q_2$$

Note:

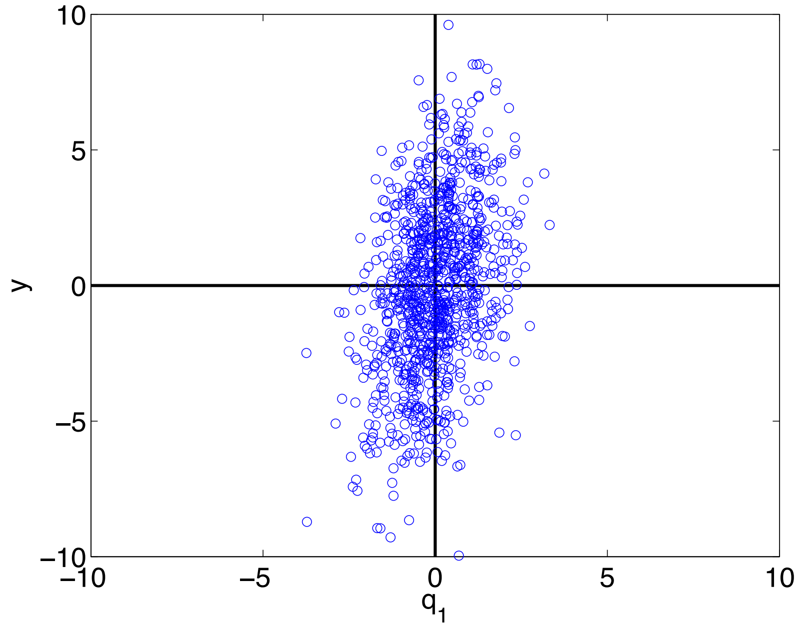
- Q_1 and Q_2 represent hedged portfolios
- c_1 and c_2 amounts invested in each portfolio

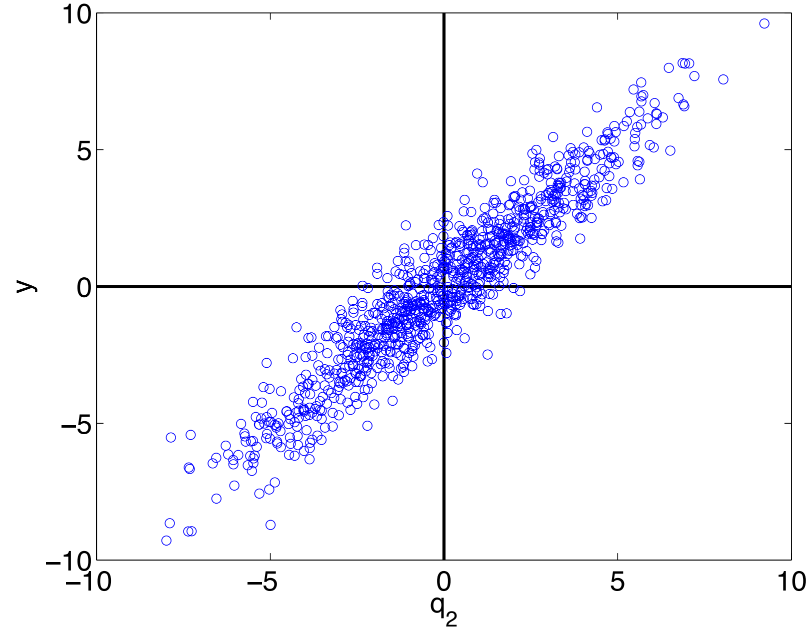
Take

$$c_1 = 2 \;,\;\; c_2 = 1$$

$$Q_1 \sim N(0, 1)$$

$$Q_2 \sim N(0, 9)$$





Local Sensitivities:

$$\frac{\partial Y}{\partial Q_1} = 2 \quad , \quad \frac{\partial Y}{\partial Q_2} = 1$$

Solutions:

- Response correlation
- Variance-based methods
- Random sampling of local sensitivities

Global Sensitivity Analysis: Variance-Based Methods

Example: Portfolio model

$$Y = c_1 Q_1 + c_2 Q_2$$

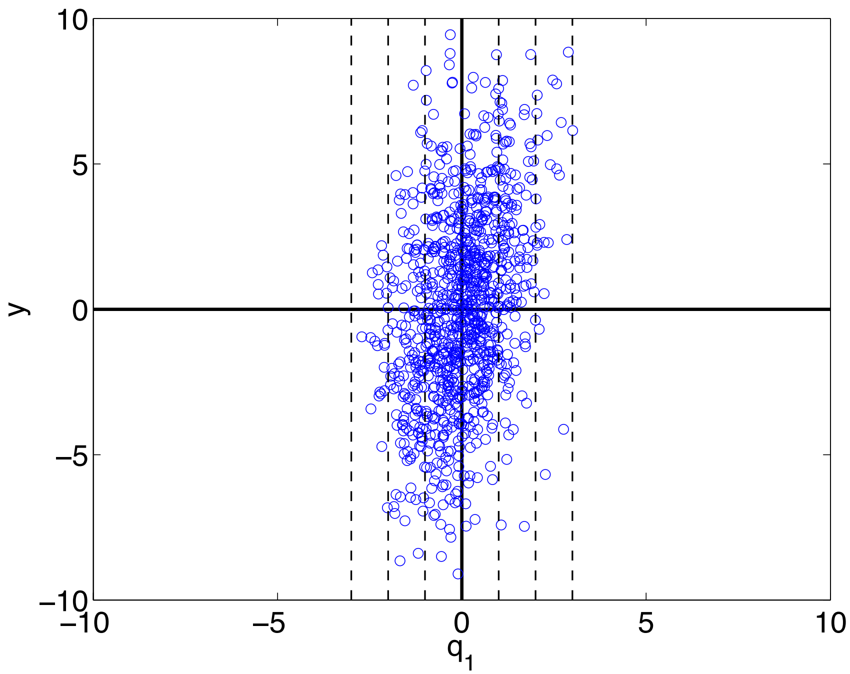
Take $c_1 = 2\,,\ c_2 = 1$

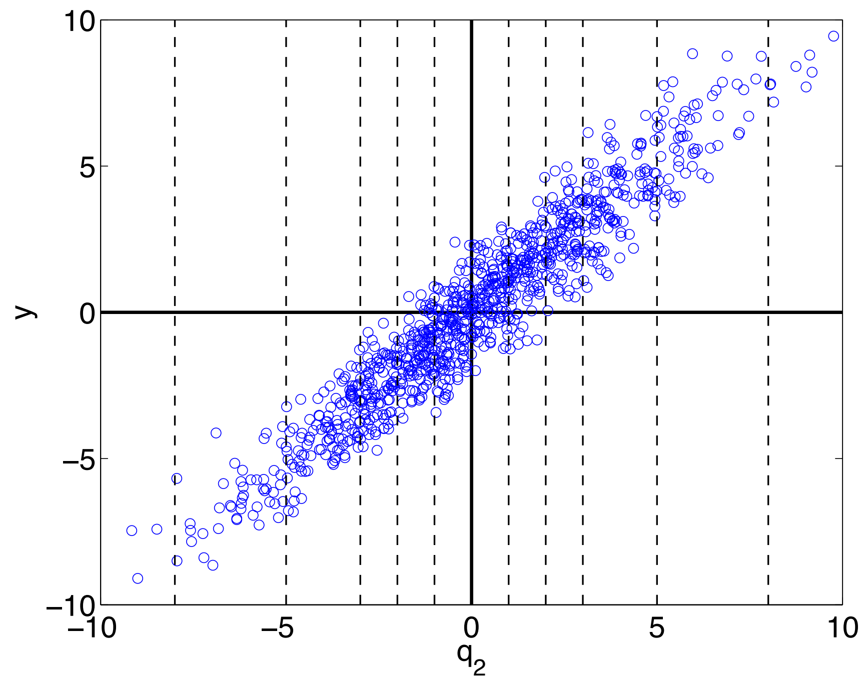
$$Q_1 \sim N(0,1)$$

$$Q_2 \sim N(0,9)$$

Statistical Motivation: Consider variability of expected values

$$D_i = \text{var}[\mathbb{E}(Y|q_i)]$$





Note: Here $D_2 > D_1$

Variance-Based Methods

Sobol Representation: For now, take $Q_i \sim \mathcal{U}(0,1)$ and $\Gamma = [0,1]^p$

Take

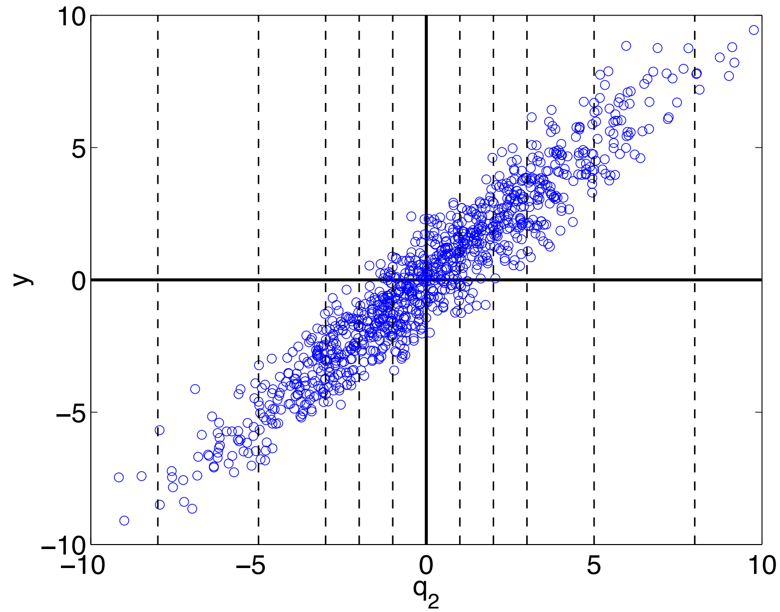
$$f(q) = f_0 + \sum_{i=1}^{p} f_i(q_i) + \sum_{1 \leqslant i < j \leqslant p} f_{ij}(q_i, q_j)$$

Analogy: Taylor or Fourier series

With appropriate assumptions,

$$f_0 = \int_\Gamma f(q)dq$$

$$f_i(q_i) = \int_{\Gamma^{p-1}} f(q)dq_{\sim i} - f_0$$



Variances:

$$D_i = \int_0^1 f_i^2(q_i)dq_i$$

$$D = \mathrm{var}(Y)$$

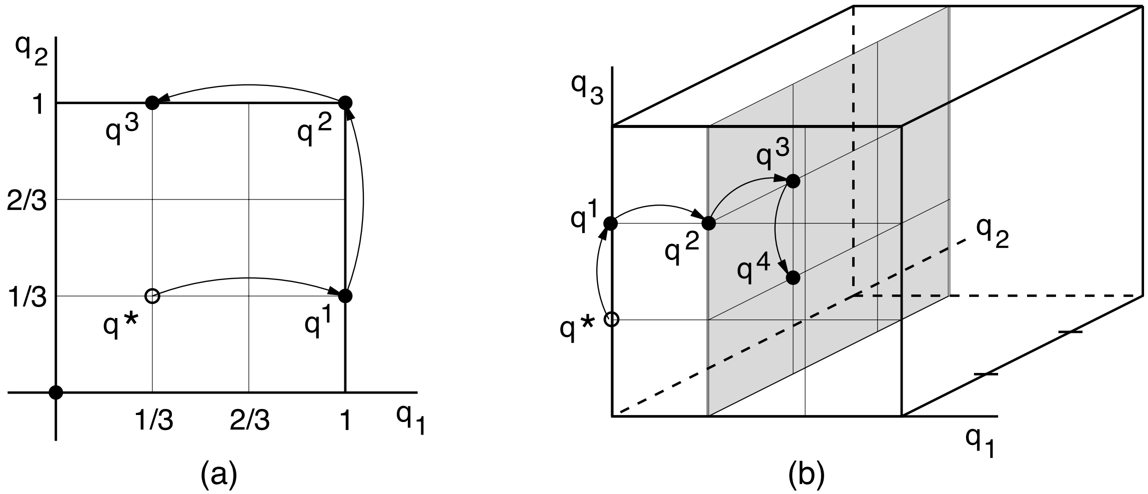
Statistical Interpretation:

$$D_i = \mathrm{var}[\mathbb{E}(Y|q_i)] \Rightarrow S_i = \frac{\mathrm{var}[\mathbb{E}(Y|q_i)]}{\mathrm{var}(Y)}$$

Sobol Indices: $S_i = \dfrac{D_i}{D}$

Morris Screening: Random Sampling of Approximated Derivatives

Example: Consider uniformly distributed parameters on $\Gamma = [0,1]^p$



Elementary Effect:

$$d_i^j = \frac{f(q^j + \Delta e_i) - F(q^j)}{\Delta}$$, i^{th} parameter , j^{th} sample

Global Sensitivity Measures: r samples

$$\mu_i^* = \frac{1}{r}\sum_{j=1}^{r} |d_i^j(q)|$$

$$\sigma_i^2 = \frac{1}{r-1}\sum_{j=1}^{r}\left(d_i^j(q) - \mu_i\right)^2 \quad , \quad \mu_i = \frac{1}{r}\sum_{j=1}^{r} d_i^j(q)$$

SIR Disease Example

SIR Model:

$$\frac{dS}{dt} = \delta N - \delta S - \gamma \underline{kIS} \quad , \quad S(0) = S_0 \qquad \text{Susceptible}$$

$$\frac{dI}{dt} = \underline{\gamma k} IS - (r + \delta) I \quad , \quad I(0) = I_0 \qquad \text{Infectious}$$

$$\frac{dR}{dt} = rI - \delta R \quad , \quad R(0) = R_0 \qquad \text{Recovered}$$

Note: Parameter set $q = [\gamma, k, r, \delta]$ is not identifiable

Assumed Parameter Distribution:

$$\gamma \sim \mathcal{U}(0,1) \; , \; k \sim Beta(\alpha, \beta) \; , \; r \sim \mathcal{U}(0,1) \; , \; \delta \sim \mathcal{U}(0,1)$$

Infection Coefficient	Interaction Coefficient	Recovery Rate	Birth/death Rate

Response:

$$y = \int_0^5 R(t,q)\,dt$$

SIR Disease Example

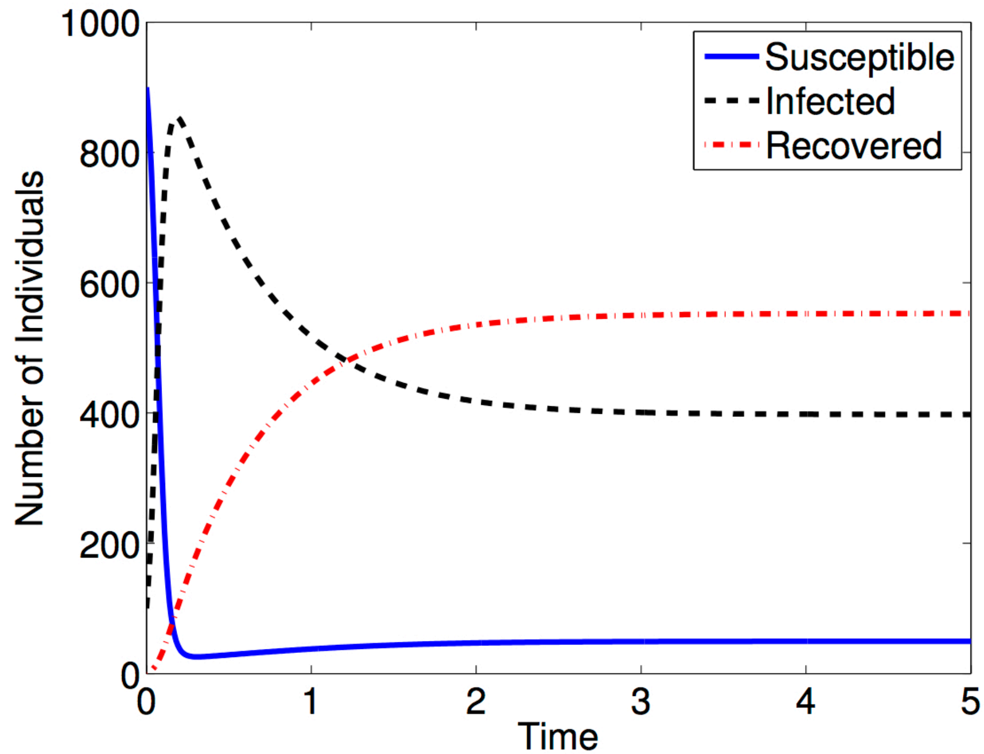
SIR Model:

$$\frac{dS}{dt} = \delta N - \delta S - \gamma k I S \quad , \quad S(0) = S_0 \qquad \text{Susceptible}$$

$$\frac{dI}{dt} = \gamma k I S - (r + \delta) I \quad , \quad I(0) = I_0 \qquad \text{Infectious}$$

$$\frac{dR}{dt} = rI - \delta R \quad , \quad R(0) = R_0 \qquad \text{Recovered}$$

Typical Realization:



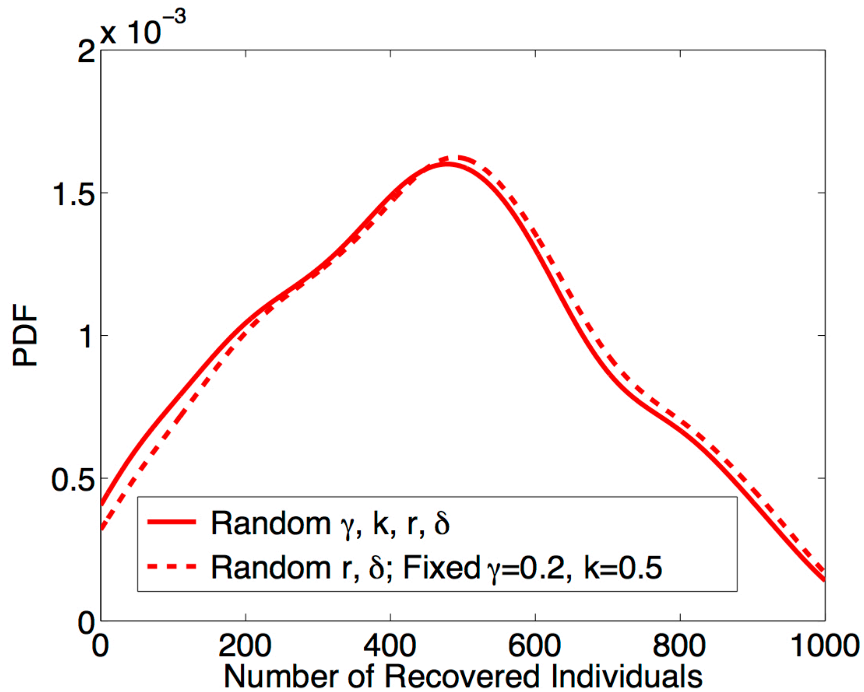
SIR Disease Example

Global Sensitivity Measures:

		γ	k	r	δ
Sobol	S_i	0.0997	0.0312	0.7901	0.1750
	S_{T_i}	-0.0637	-0.0541	0.5634	0.2029
Morris	μ_i^* $(\times 10^3)$	0.2532	0.2812	2.0184	1.2328
	σ_i $(\times 10^3)$	0.9539	1.6245	6.6748	3.9886

Influential Parameters

Result: Densities for $R(t_f)$ at $t_f = 5$



Note: Can fix non-influential parameters γ, k

Parameter Selection: Nuclear Power Plant Design

Subchannel Code (COBRA-TF): numerous closure relation and parameters

parameter	partial correlation	simple correlation	morris main	morris interaction	CPS variation
k_eta	0.07	0.03			
k_gama	-0.03	0.04			
k_sent	-0.03	-0.02			
k_sdent	-0.07	-0.01			
k_tmasv	-0.03	0.00			
k_tmasl	0.11	0.00	6.48E-05	2.28E-05	medium
k_tmasg	-0.19	-0.01			
k_tmomv	-0.12	-0.01			
k_tmome	0.02	0.00			
k_tmoml	0.02	-0.02	2.23E-04	1.30E-04	medium
k_xk	0.08	-0.02			
k_xkes	-0.05	0.00			
k_xkge	-0.07	0.01			
k_xkl	0.04	-0.01			
k_xkle	-0.03	0.00			
k_xkvls	0.11	-0.01			
k_xkwvw	-0.10	0.01			
k_xkwlw	0.14	0.01			
k_xkwew	-0.01	0.03			
k_qvapl	-0.09	-0.01			
k_tnrgv	-0.03	0.00			
k_tnrgl	-0.01	0.03	9.00E-06	9.49E-06	low
k_rodqq	0.02	-0.01			
k_qradd	-0.02	0.00			
k_qradv	-0.01	0.00			
k_qliht	-0.01	0.00			
k_sphts	-0.05	0.03			
k_cond	-0.04	0.00			
k_xkwvx	0.03	-0.02			
k_xkwlx	1.00	0.88	1.80E-01	7.07E-03	high
k_cd	1.00	0.46	9.59E-02	7.88E-03	high
k_cdfb	-0.02	-0.01			
k_wkr	0.02	0.02			

5 Identified Active Inputs:

k_cd: Pressure loss coefficient of space in sub-channel

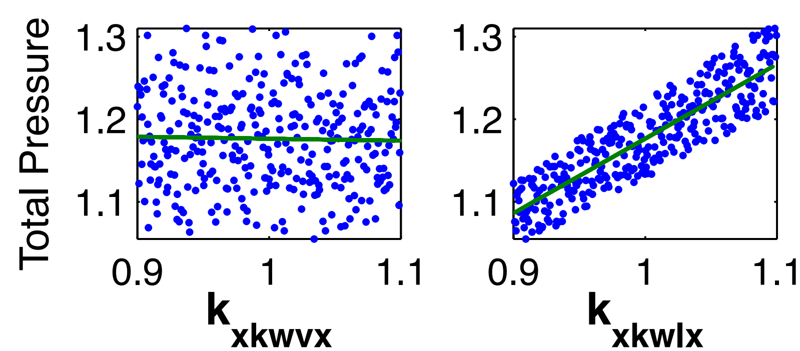
k_xkwlx: Vertical liquid wall drag coefficient

k_tmasl: Loss of liquid mass due to mixing and void drift

k_tmoml: Loss of liquid momentum due to mixing and void drift

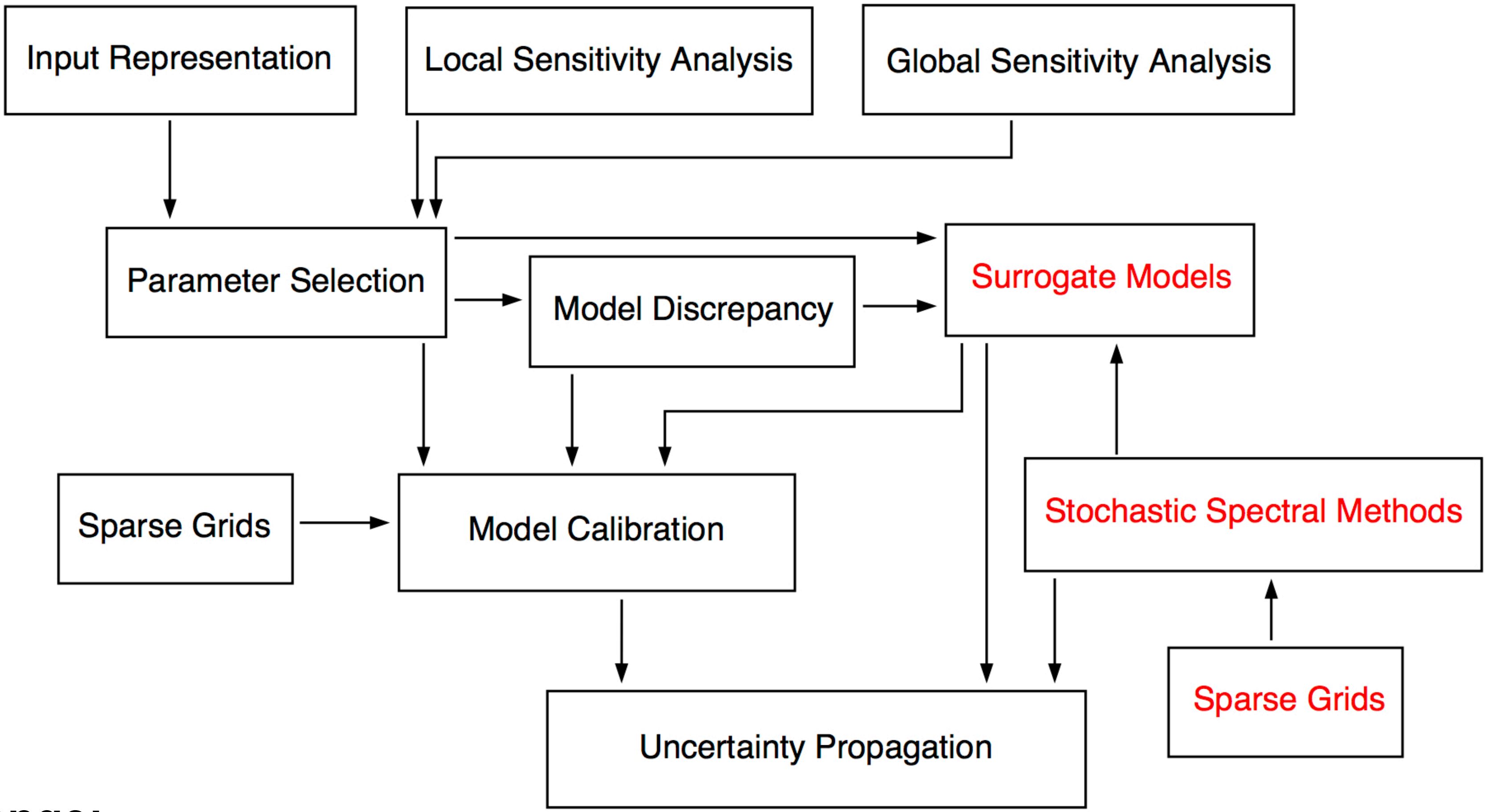
k_tnrgl: Loss of liquid enthalpy due to mixing and void drift

Partial Correlation:



Note: 33 initial VUQ parameters reduced to 5 via sensitivity analysis

Steps in Uncertainty Quantification

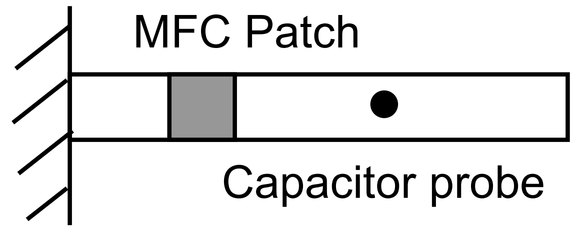


Challenge:

- How do we do uncertainty quantification for computationally expensive models?
- Example:
  - We have a computational budget of 5000 model evaluations.
  - Bayesian inference and uncertainty propagation require 120,000 evaluations.

Uncertainty Quantification Challenges

Example: MFC model – Fourth-order PDE

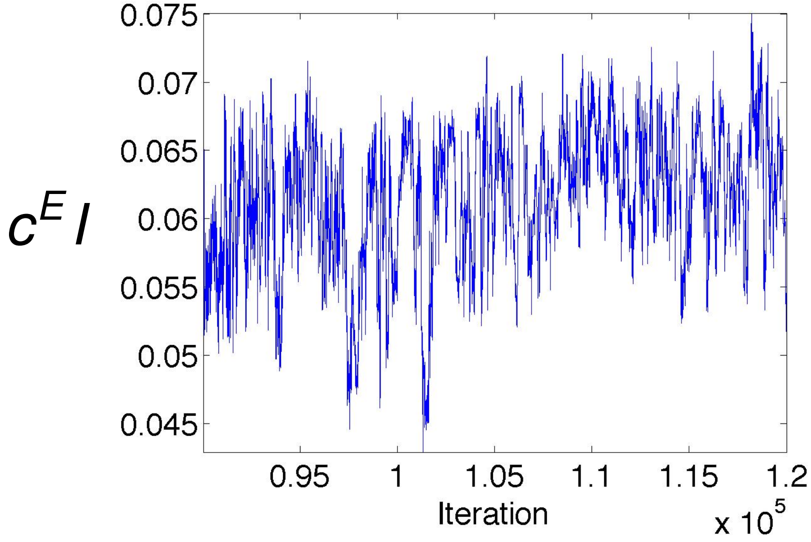


$$\rho \frac{\partial^2 w}{\partial t^2} - \frac{\partial^2 M}{\partial x^2} = f$$

$$M = -\underline{c^E I} \frac{\partial^2 w}{\partial x^2} - c_D I \frac{\partial^3 w}{\partial x^2 \partial t} - \left[k_1 e(E, \sigma_0) E + k_2 \varepsilon_{irr}(E, \sigma_0)\right] \chi_{MFC}(x)$$

Macro-Fiber Composite

Bayesian Inference: Took 6 days!



Problem:

1.2×10^5 PDE solutions

Solution: Highly efficient surrogate models

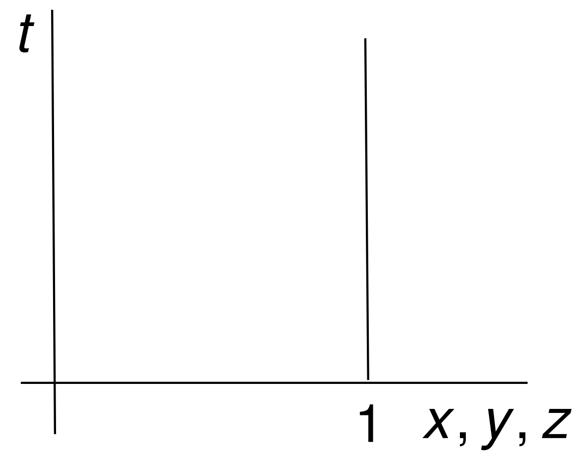
Surrogate Models: Motivation

Example: Consider the heat equation

$$\frac{\partial u}{\partial t} + \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2} = f(q)$$

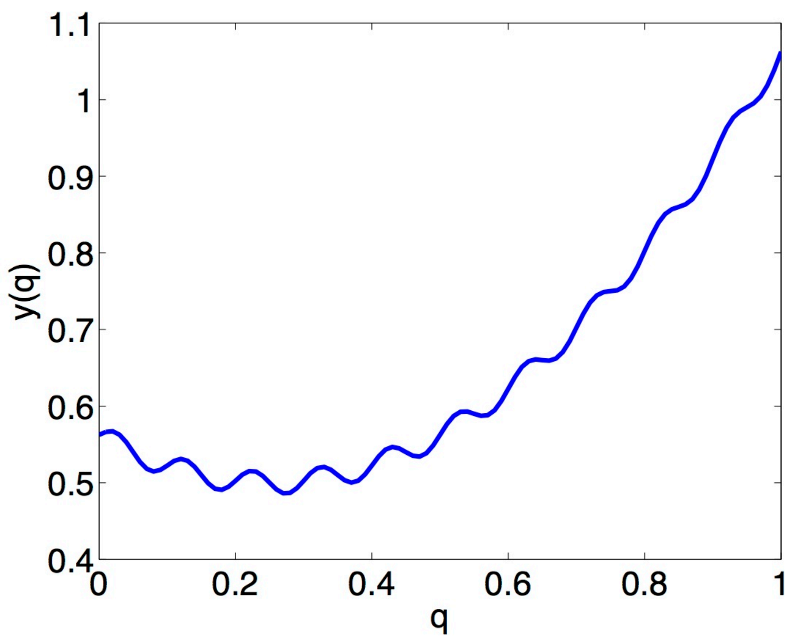
Boundary Conditions

Initial Conditions



with the response

$$y(q) = \int_0^1 \int_0^1 \int_0^1 \int_0^1 u(t,x,y,z)\,dxdydzdt$$



Notes:

- Requires approximation of PDE in 3-D
- What would be a simple surrogate?

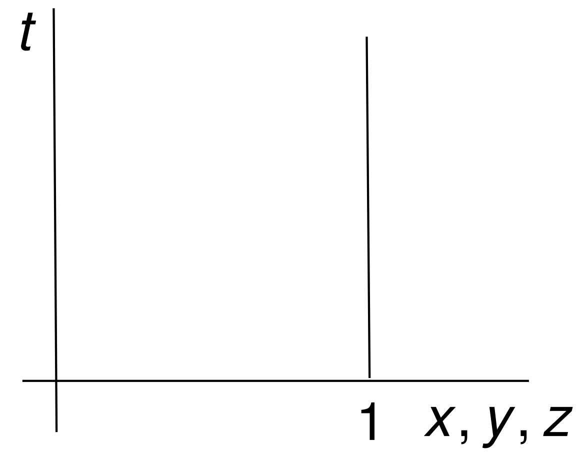
Surrogate Models: Motivation

Example: Consider the heat equation

$$\frac{\partial u}{\partial t} + \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2} = f(q)$$

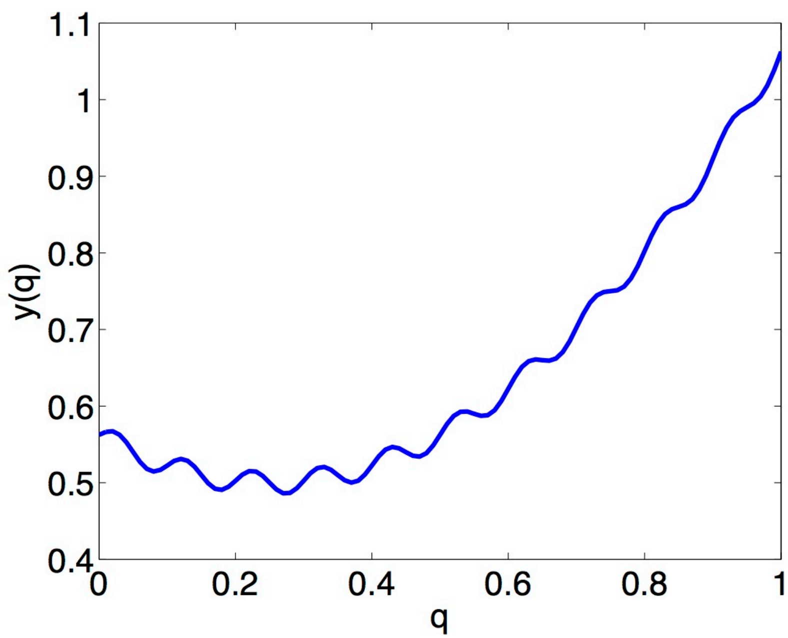
Boundary Conditions

Initial Conditions



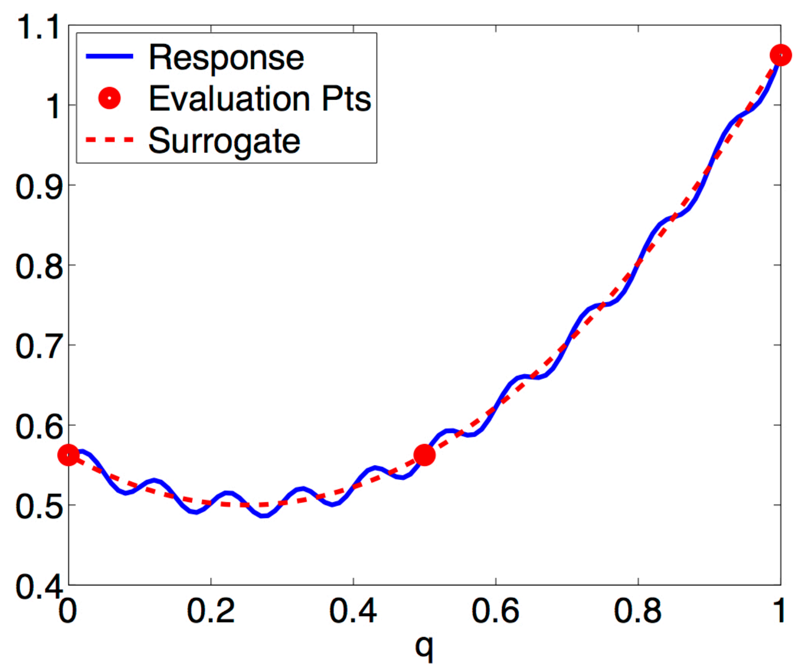
with the response

$$y(q) = \int_0^1 \int_0^1 \int_0^1 \int_0^1 u(t, x, y, z)\, dxdydzdt$$



Surrogate: Quadratic

$$y_s(q) = (q - 0.25)^2 + 0.5$$



Data-Fit Models

Notes:

• Often termed response surface models, surrogates, emulators, meta-models.

• Rely on interpolation or regression.

• Data can consist of high-fidelity simulations or experiments.

• Common techniques: polynomial models, Gaussian process (GP), orthogonal polynomials ('analytic' moment relations)

Strategy: Consider high fidelity model

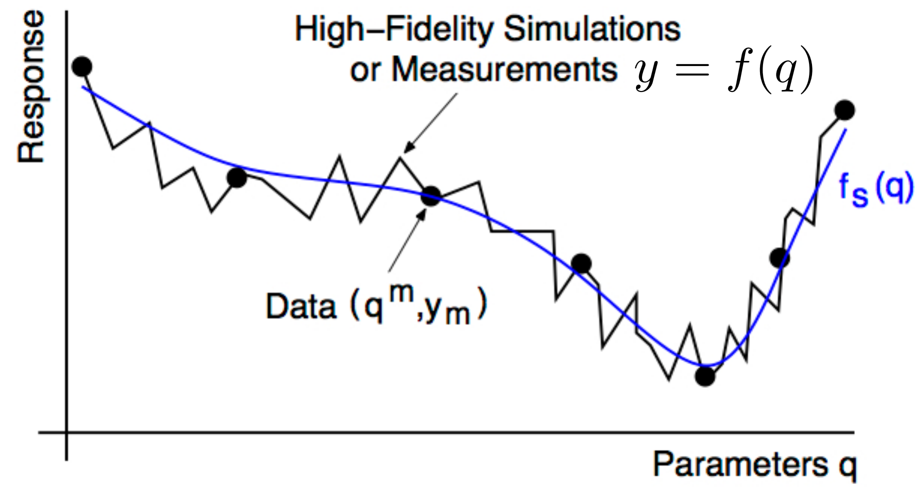
$$y = f(q)$$

with M model evaluations

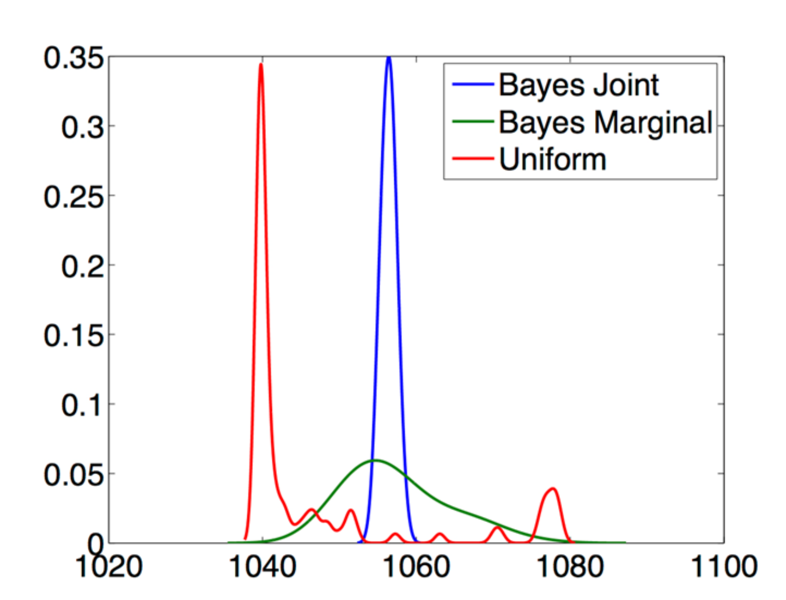
$$y_m = f(q^m) \ , \ \ m = 1, \dots, M$$

Statistical Model: $f_s(q)$: Surrogate for $f(q)$

$$y_m = f_s(q^m) + \varepsilon_m \ , \ \ m = 1, \dots, M$$



Note: Employed GP surrogate of CTF to compute uncertainty in max fuel temperatures



Gaussian Process (GP) Emulators

Strategy:

- Model simulator outputs $y_m(q)$ as being generated by a Gaussian process; i.e.,

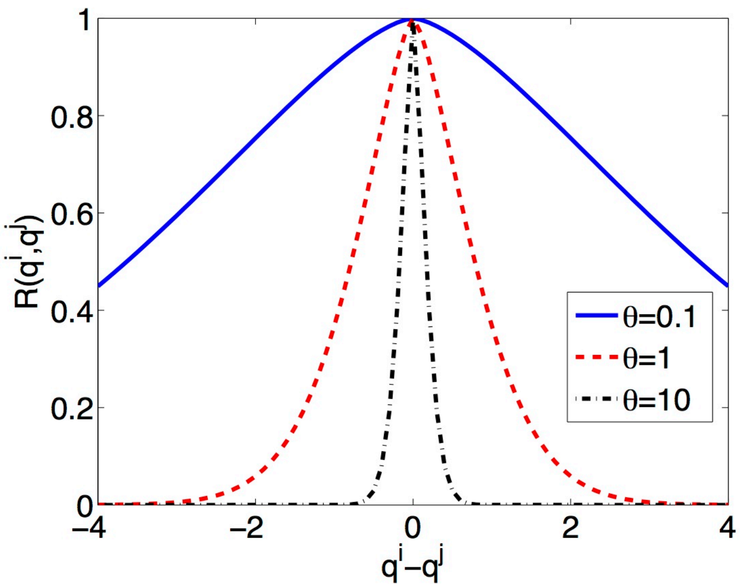
$$[y_m(q^1), \ldots, y_m(q^M)] \sim MVN((\mu(q^1), \ldots, \mu(q^M), \sigma^2 R)$$

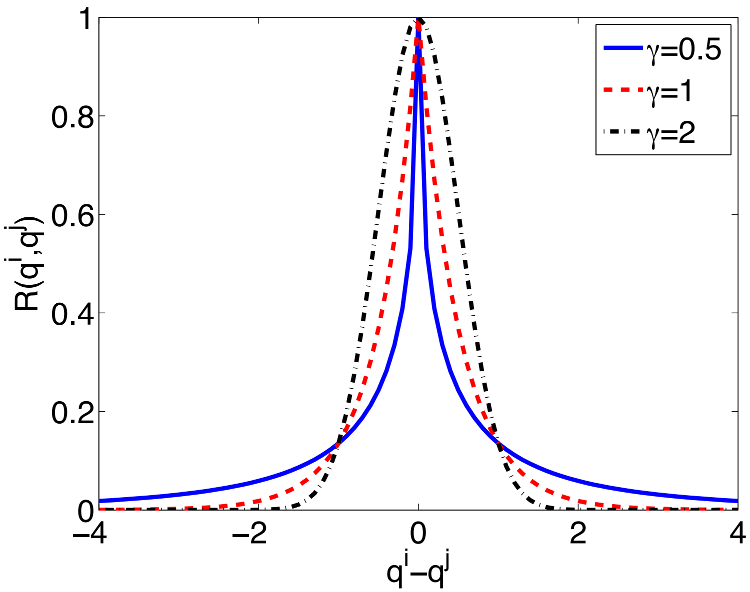
Here R is constructed to model the correlation structure of simulator outputs

Example:

$$R(q^i, q^j) = \exp\left(-\sum_{k=1}^{p} \left|\theta_k(q_k^i - q_k^j)\right|^{\gamma_k}\right)$$

Note: Hyper-parameters θ_k, γ_k tuned to achieve varying degrees of correlation





Gaussian Process (GP) Emulators

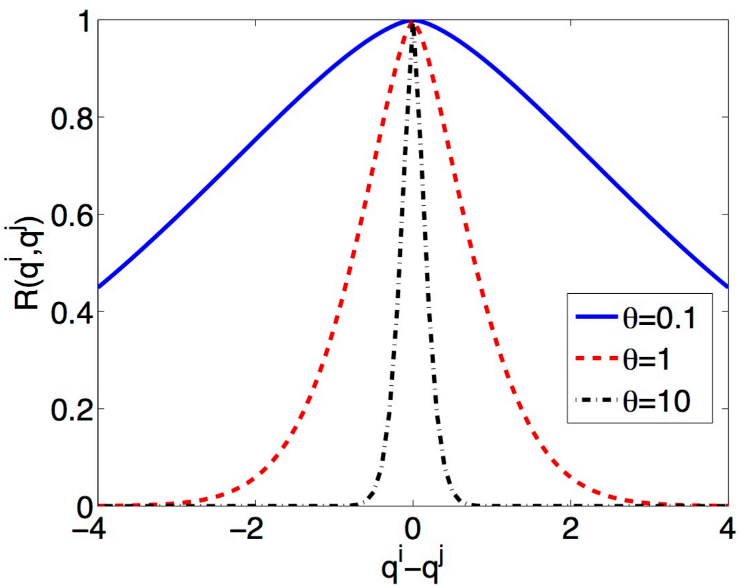
Strategy:

- Model simulator outputs $y_m(q)$ as being generated by a Gaussian process; i.e.,

$$[y_m(q^1), \ldots, y_m(q^M)] \sim MVN((\mu(q^1), \ldots, \mu(q^M), \sigma^2 R)$$

Here R is constructed to model the correlation structure of simulator outputs

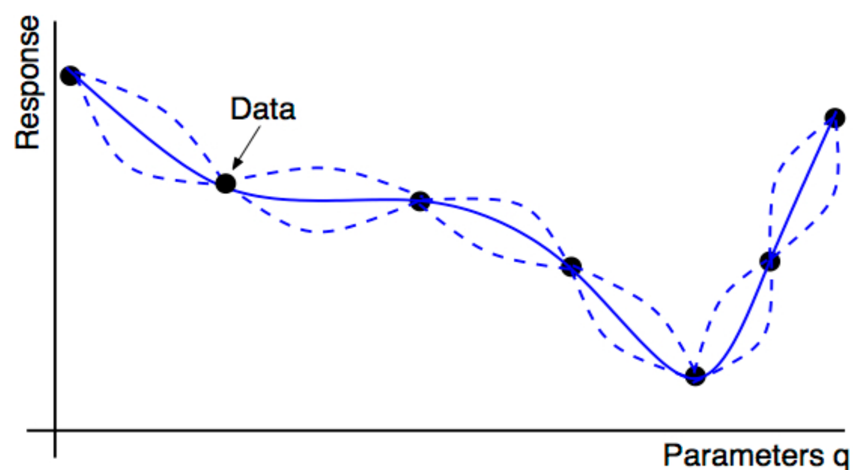
Example:



Note: In absence of observation noise, GP surrogate interpolates; i.e.,

$$y_m = f_s(q^m, \beta)\ ,\ \ m = 1, \ldots, M$$

Uncertainty Bounds:



Example: Modeling of Volcanic Pyroclastic Flows

Authors: Bayarri, Berger, Calder Dalbey, Lunagomez, Patra, Pitman, Spiller, Wolpert; *Technometrics,* 51(4), 2009; Gu and Berger, *The Annals of Applied Statistics*, 2016.

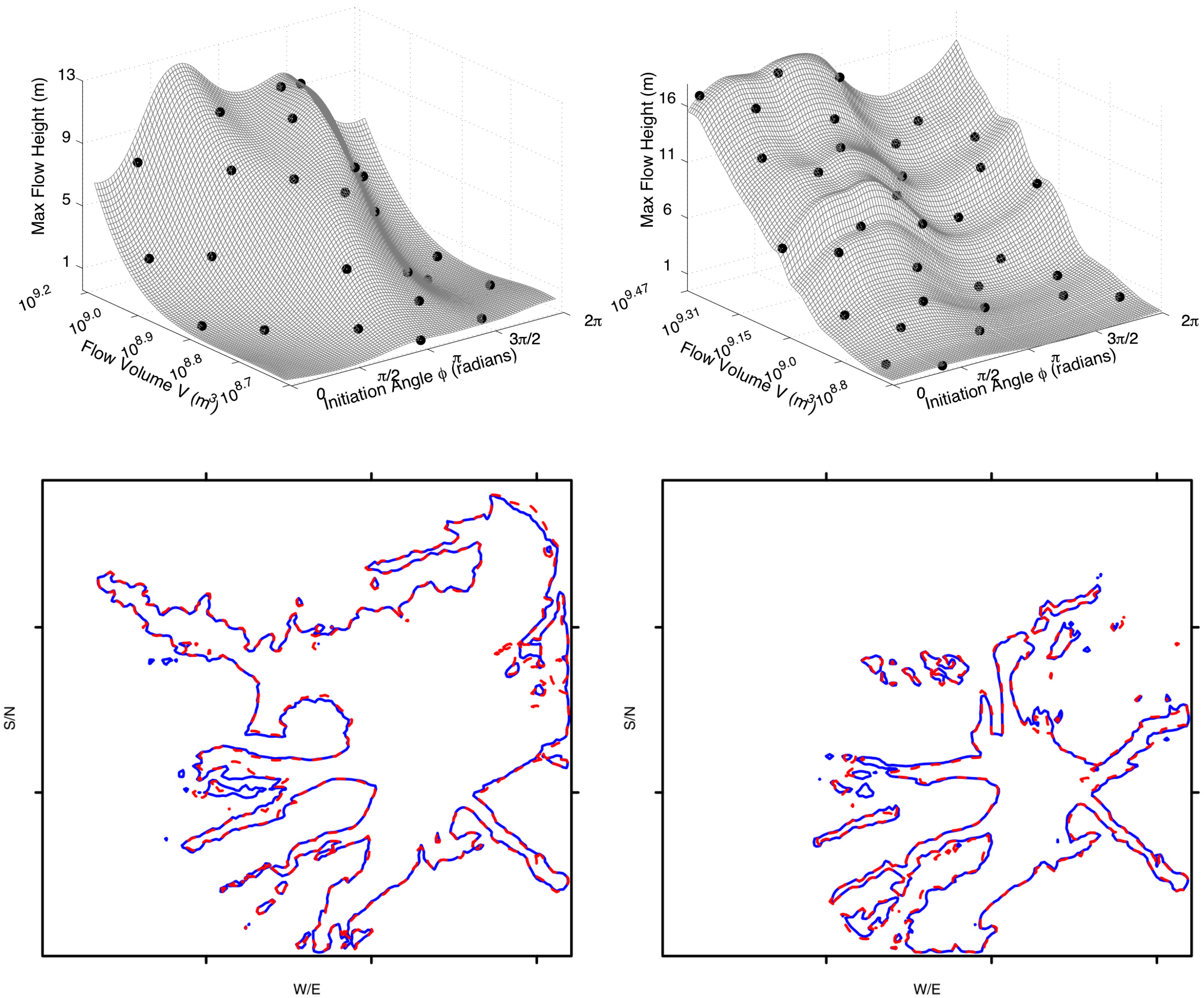
Objectives:

- Employ simulation models and surrogates to assess risk of *rare* catastrophic events; e.g., volcanic eruption.
- Employed TITAN2D to simulate flows.
- Test Case: Soufrière Hills Volcano on Island of Montserrat.
- Use emulator to identify threshold inputs – e.g., critical flow depth – that define catastrophic event.
- Compared GP and mathematical surrogates; GP advantageous for this application.

Example: Modeling of Volcanic Pyroclastic Flows

Objectives:

- Use emulator to identify threshold inputs – e.g., critical flow depth – that define catastrophic event. Employed TITAN2D and GP surrogates.



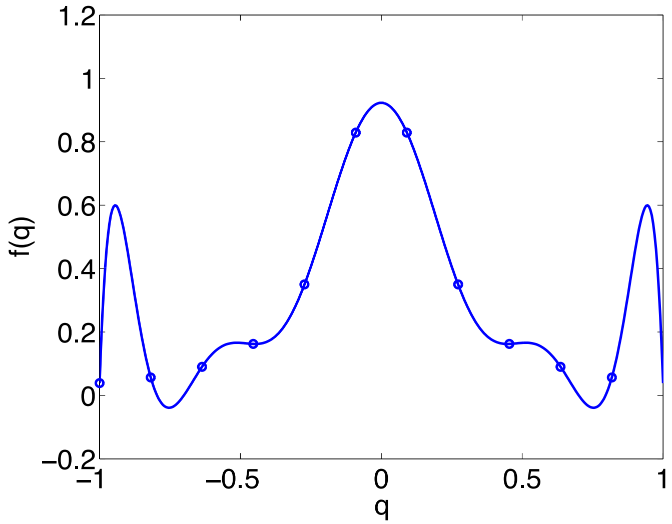
TITAN2D

GP

Surrogate Models – Grid Choice

Example: Consider the Runge function $f(q) = \frac{1}{1+25q^2}$ with points

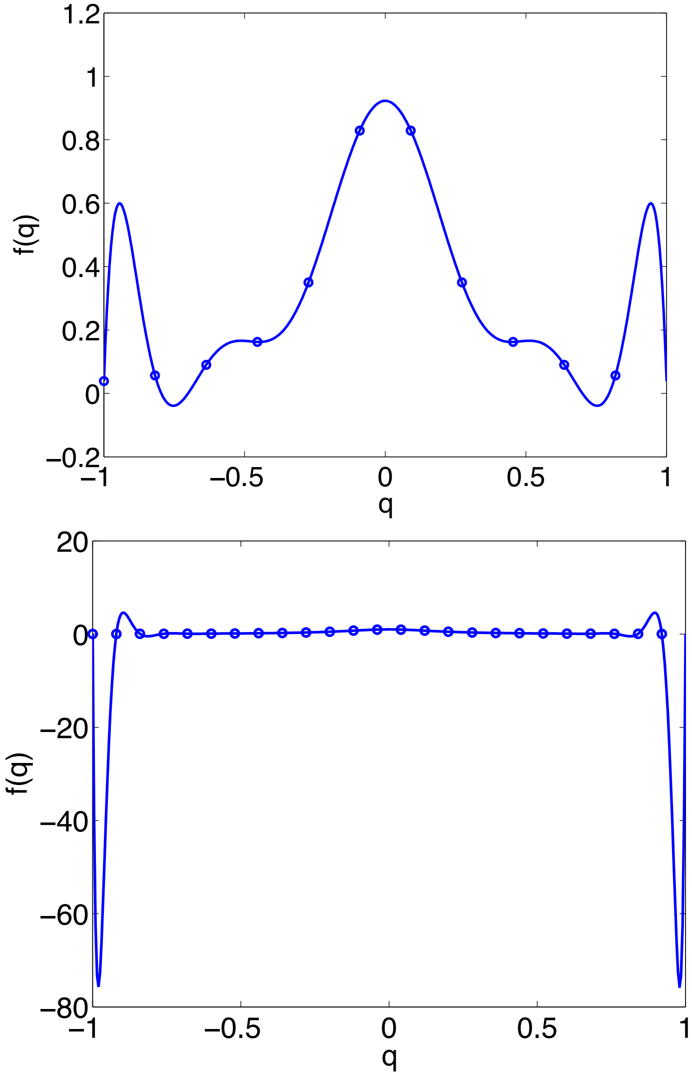
$$q^j = -1 + (j-1)\frac{2}{M}\ ,\ j = 1, \ldots, M$$



Surrogate Models – Grid Choice

Example: Consider the Runge function $f(q) = \frac{1}{1+25q^2}$ with points

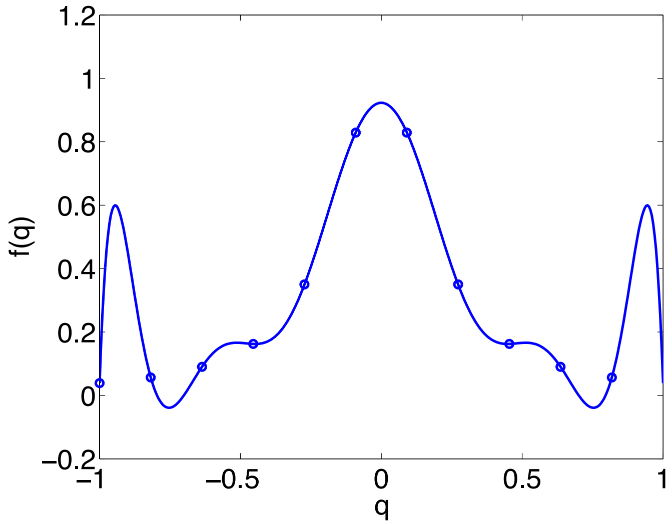
$$q^j = -1 + (j-1)\frac{2}{M} \; , \; j = 1, \ldots, M$$

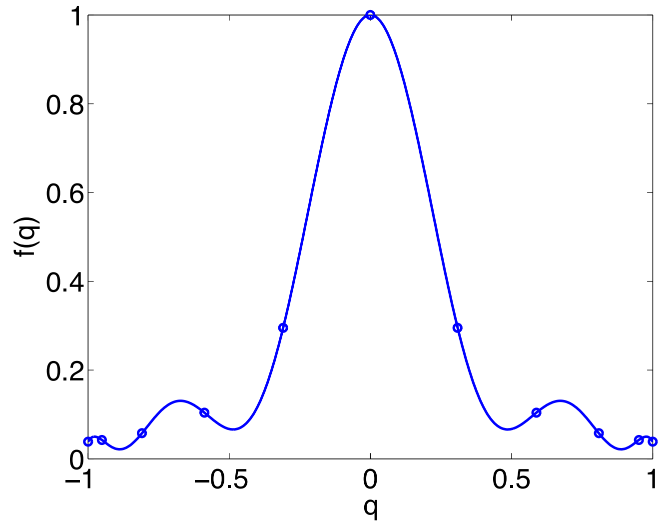


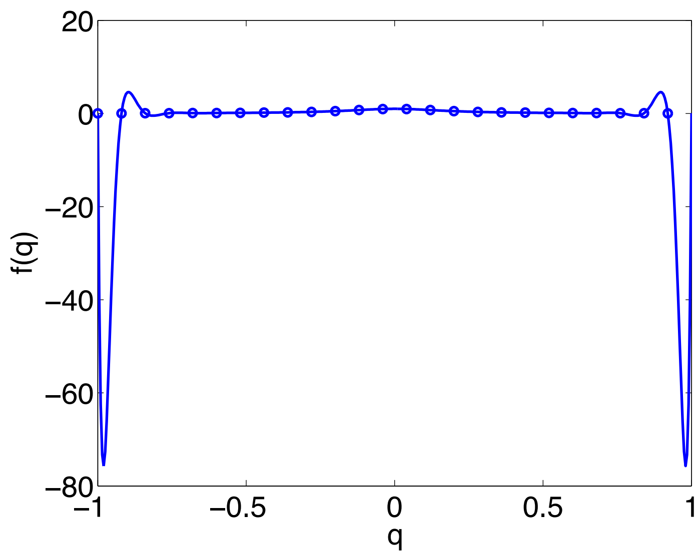
Surrogate Models – Grid Choice

Example: Consider the Runge function $f(q) = \frac{1}{1+25q^2}$ with points

$$q^j = -1 + (j-1)\frac{2}{M} \; , \; j = 1, \ldots, M \qquad q^j = -\cos\frac{\pi(j-1)}{M-1} \; , \; j = 1, \ldots, M$$



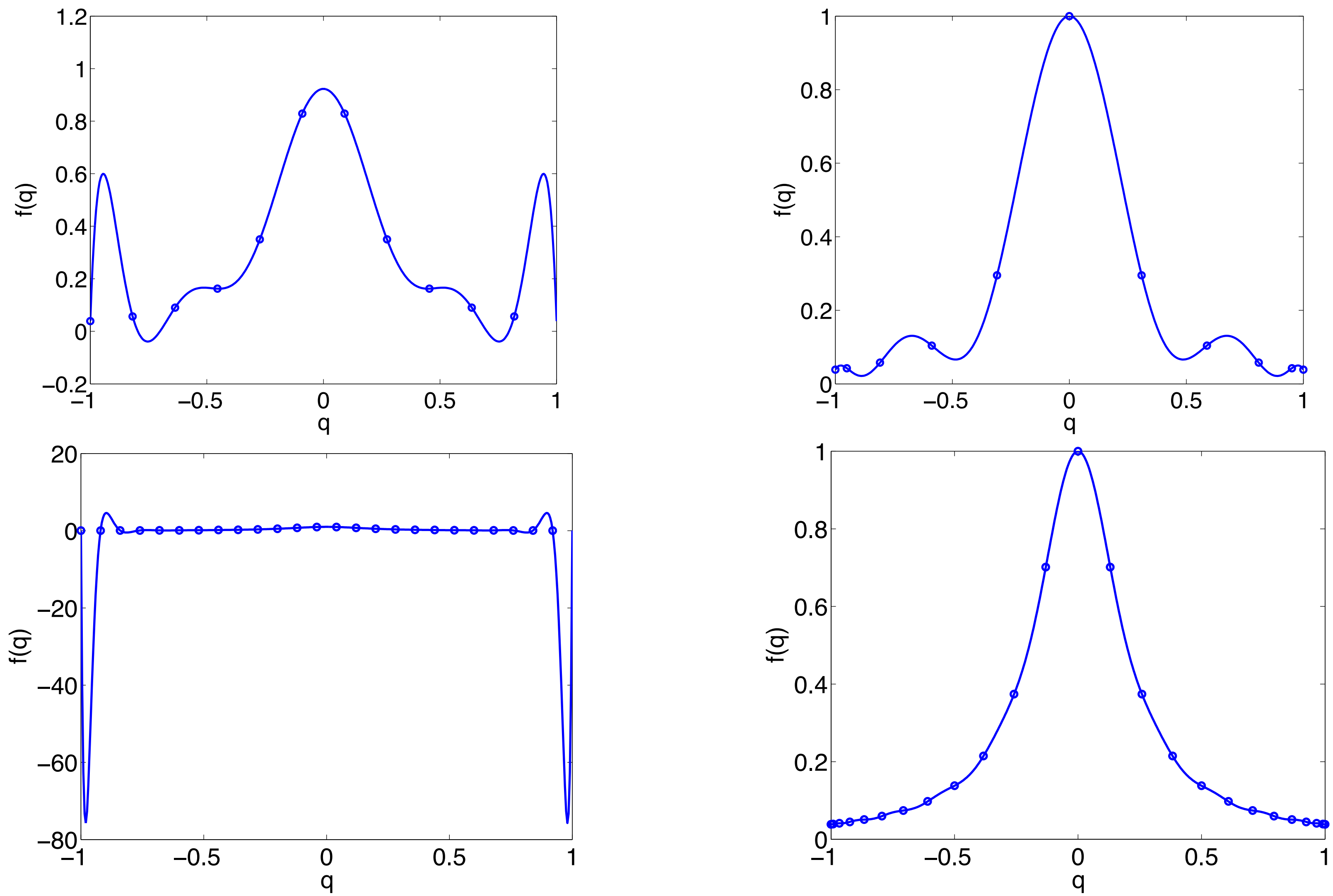




Surrogate Models – Grid Choice

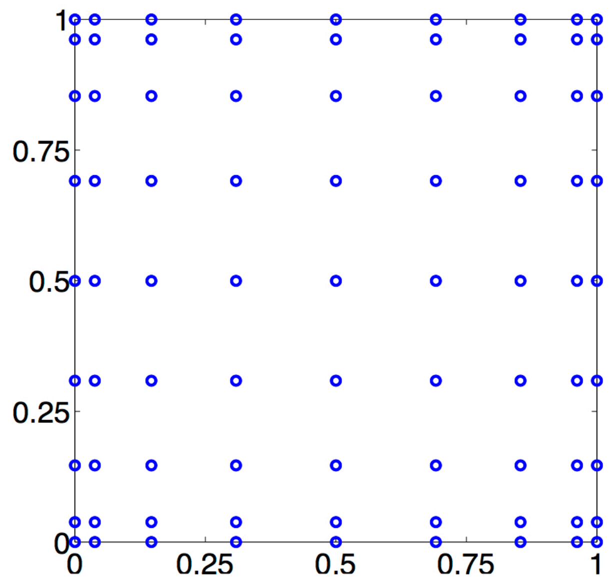
Example: Consider the Runge function $f(q) = \frac{1}{1+25q^2}$ with points

$$q^j = -1 + (j-1)\frac{2}{M}\ ,\ j = 1, \ldots, M \qquad q^j = -\cos\frac{\pi(j-1)}{M-1}\ ,\ j = 1, \ldots, M$$

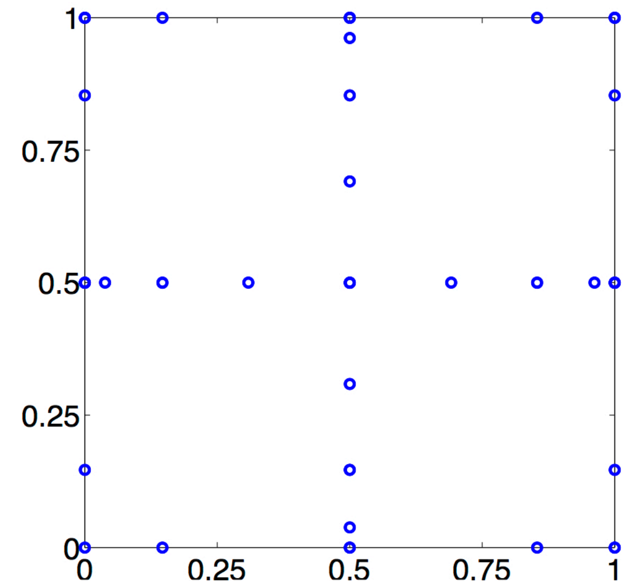


Sparse Grid Techniques

Tensored Grids: Exponential growth



Sparse Grids: Same accuracy



p	R_ℓ	Sparse Grid $\mathcal{R}$	Tensored Grid $R = (R_\ell)^p$
2	9	29	81
5	9	241	59,049
10	9	1581	$> 3 \times 10^9$
50	9	171,901	$> 5 \times 10^{47}$
100	9	1,353,801	$> 2 \times 10^{95}$

Concluding Remarks

Notes:

• UQ requires a synergy between engineering, statistics, and applied mathematics.

• Model calibration, model selection, uncertainty propagation and experimental design are natural in a Bayesian framework.

• Goal is to predict model responses with quantified and reduced uncertainties.

• Parameter selection is critical to isolate identifiable and influential parameters.

• Surrogate models critical for computationally intensive simulation codes.

• Codes and packages: Sandia Dakota, R, MATLAB, Python, nanoHUB.

• *Prediction is very difficult, especially if it's about the future*, Niels Bohr.