Encounters with Imprecise Probabilities

Jim Berger

Duke University

ISIPTA'17 (& ECSQARU 2017)
Lugano, Switzerland
July 12, 2017

Outline

- My view on applying IP, molded by interactions with Herman Rubin and Jack Good
- Three applications of IP:
  - to correct the p-value problem in science;
  - to provide a sound analysis for any normal hierarchical model;
  - to tackle *uncertainty quantification (UQ)*, the intersection of mathematical modeling of processes and data.

My early influences: Herman Rubin and Jack Good

Herman Rubin: In *A weak system of axioms for 'rational' behavior and the non-separability of utility from prior*, he showed that incredibly weak imprecise choice axioms require compatibility with some Bayesian analysis.

An implementation: Model the (say) imprecise probabilities by the class $\mathcal{P}$ of compatible probability distributions (credal sets), and make interesting statements for the class (if possible); this should be the gold standard for IP.

Example 1 (Lenny's problem): Suppose climate science leads to a predictive probability distribution $p(y)$ for temperature y. But we assess that there is a 20% chance of the 'big surprise' i.e., we are completely wrong. This can be represented by the class of probability distributions

$$\mathcal{P} = \{0.2q(y) + 0.8p(y);\ q(y) \text{ being any distribution}\}\,.$$

If, say, $p(y)$ is $N(30, 2^2)$, one could then make the valid statement

$$Pr(Y < 34) = 0.2Pr(Y < 34 \mid q) + 0.8P(Y < 34 \mid p) \geq (0.8)(97.5) = 0.78\,.$$

Jack Good (who, while theoretical, tended to focus on practical ideas): When handling imprecise probabilities, "use probabilities of a higher type."

Example 2 (Genome-Wise Association Studies - GWAS):

- A typical GWAS study looks at, say, 20 (related) diseases and 1,000,000 genes (or SNPs), and attempts to determine which genes are associated with which diseases. (Note: 20,000,000 tests are being done here.)
- GWAS studies from 1997-2007 (about 50,000 published papers) had an extremely high rate of replication failure, because most were not adjusting enough for multiple testing, thinking that using 'strict' p-values such 10^{-3} or 10^{-4} would be enough.
- A very influential paper in Nature (2007), by the Wellcome Trust Case Control Consortium, argued for a cutoff of $p < 5 \times 10^{-7}$ for claiming discovery of an association (later shifted to 5×10^{-8} and ??? today).
- *Key step:* They did a subjective Bayesian assessment that the prior odds of a true association to false association are 1/100,000, stating this could be off by a factor of 10 either way (which they subsequently ignored).

Jack Good's solution would have been

- p, the probability of a disease/gene association should be considered a 'logical' unknown probability, to be handled at a 'higher level.'
- At the higher level, assign a prior distribution; for instance a Gamma(1,100,000) prior is compatible with the prior information of the medical geneticists.
- This would be irrelevant if there were no data (and p entered the subsequent analysis linearly) but there is lots of data.
- This fits into the class of priors framework by defining

$$\mathcal{P} = \{\text{point mass distributions at } p, \quad 0 \leq p \leq 1\}\,.$$

While dealing with $\mathcal{P}$ by placing a single prior distribution on the class logically still corresponds to just a single overall distribution, Good argued that answers are much less sensitive to such higher level distributions.

These examples outline the way I have always approached IP in practice.

- Model the imprecision through a class $\mathcal{P}$ of probability distributions, and proceed by either
  - Making interesting probability statements that are valid for any distribution in $\mathcal{P}$ (called *robust Bayesian analysis* in the old days);
  - Placing a probability distribution over $\mathcal{P}$ and proceeding (*hierarchical Bayesian analysis*).
- Robust Bayesian analysis is sometimes very effective; hierarchical Bayesian analysis is usually very effective.

An Aside: There are other versions of robust Bayesian analysis:

- Choose a 'robust' prior distribution in $\mathcal{P}$ to use.
- Choose the most 'objective' prior distribution in $\mathcal{P}$, the extreme of which, when $\mathcal{P}$ = all distributions, is *objective Bayesian analysis.*
- Choose the empirical Bayes prior distribution in $\mathcal{P}$ (almost always worse than hierarchical Bayes).

I. The p-value issue

- Significance testing using p-values, declaring a 'discovery' if $p \leq 0.05$, is by far the dominant method of testing in science.
- Its standard uncritical use is viewed by many as being a major source of the problems of reproducibility of science.
- Everyone is talking about it:
  - articles in all the major science journals;
  - changes in editorial policy (the journal *Basic and Applied Social Psychology* banned p-values);
  - the recent American Statistical Association position statement about p-values and discussion;
  - an article about to appear in *Nature Human Behavior*, with over 70 leading scientists in a variety of fields, recommending changing 'statistical significance' from $p \leq 0.05$ to $p \leq 0.005$.

The Major Problem: p-values are misinterpreted

- Few non-statisticians understand p-values, most erroneously thinking they are some type of error probability, Bayesian or frequentist; they are neither!
  - A survey 30 years ago:
    - "What would you conclude if a properly conducted, randomized clinical trial of a treatment was reported to have resulted in a beneficial response ($p < 0.05$)?
      1. Having obtained the observed response, the chances are less than 5% that the therapy is not effective.
      2. The chances are less than 5% of not having obtained the observed response if the therapy is effective.
      3. The chances are less than 5% of having obtained the observed response if the therapy is not effective.
      4. None of the above.
    - We asked this question of 24 physicians ... Half ... answered incorrectly, and all had difficulty distinguishing the subtle differences...
    - The correct answer to our test question, then, is 3."

"This isn't right. This isn't even wrong." –Wolfgang Pauli, on a submitted paper

* Actual correct answer: The chances are less than 5% of having obtained the observed response ***or any more extreme response*** if the therapy is not effective.

- But, is it fair to count 'possible data more extreme than the actual data' in the evidence against the null hypothesis?

  Jeffreys (1961): "An hypothesis, that may be true, may be rejected because it has not predicted observable results that have not occurred."

- Matthews (1998): "The plain fact is that 70 years ago Ronald Fisher gave scientists a mathematical machine for turning baloney into breakthroughs, and flukes into funding."

- When testing *precise hypotheses*[a], true error probabilities (Bayesian or conditional frequentist) are much larger than p-values.

[a]For testing other types of hypotheses, such as one-sided hypotheses, the situation can be quite different.

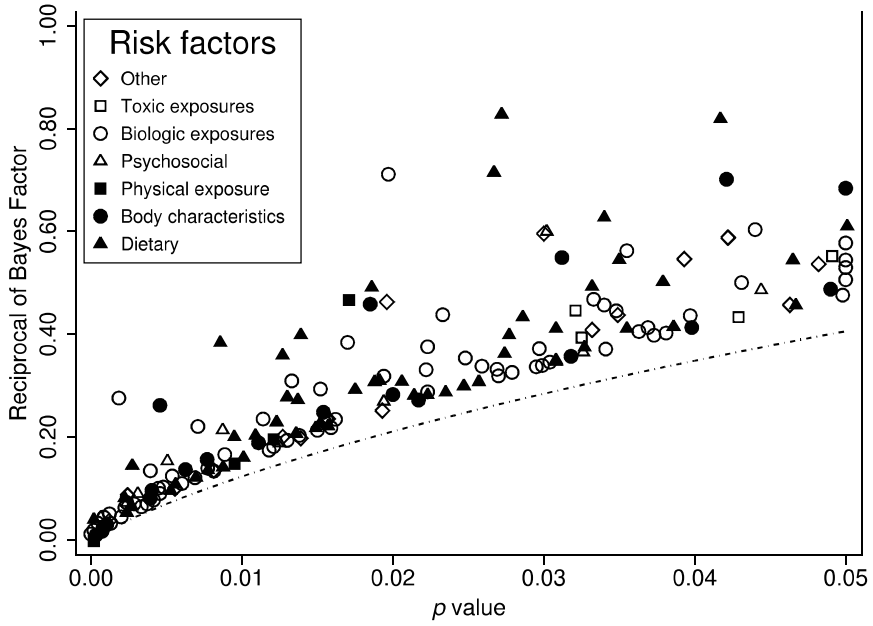
Comparison of p-values with Bayesian odds of hypotheses

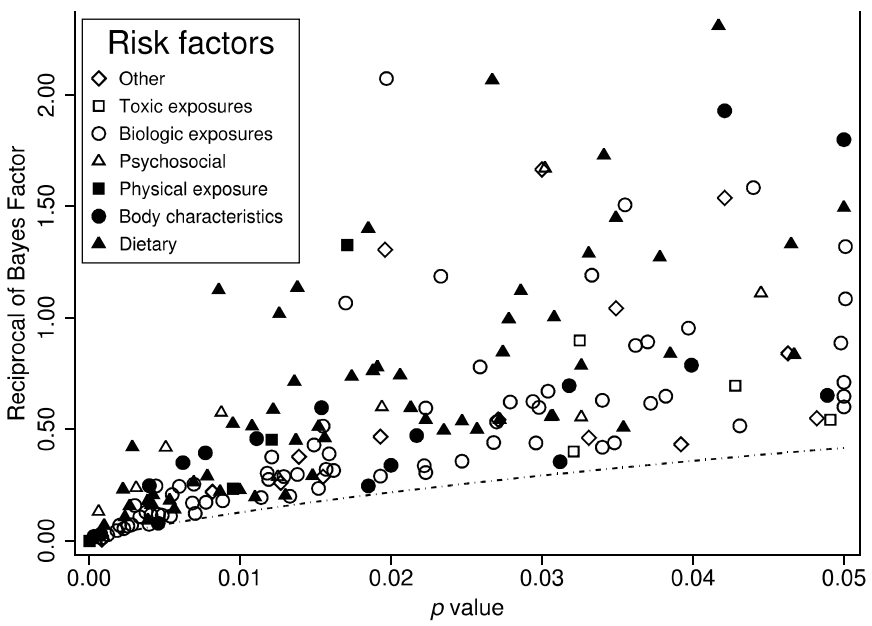
Suppose data x arises from the density $f(x \mid \theta)$, and we are interested in testing $H_0 : \theta = 0$ versus $H_1 : \theta \neq 0$ with $\pi(\theta)$ being a specified prior distribution of θ under H_1. The *Bayesian odds (Bayes factor)* of H_0 to H_1 is

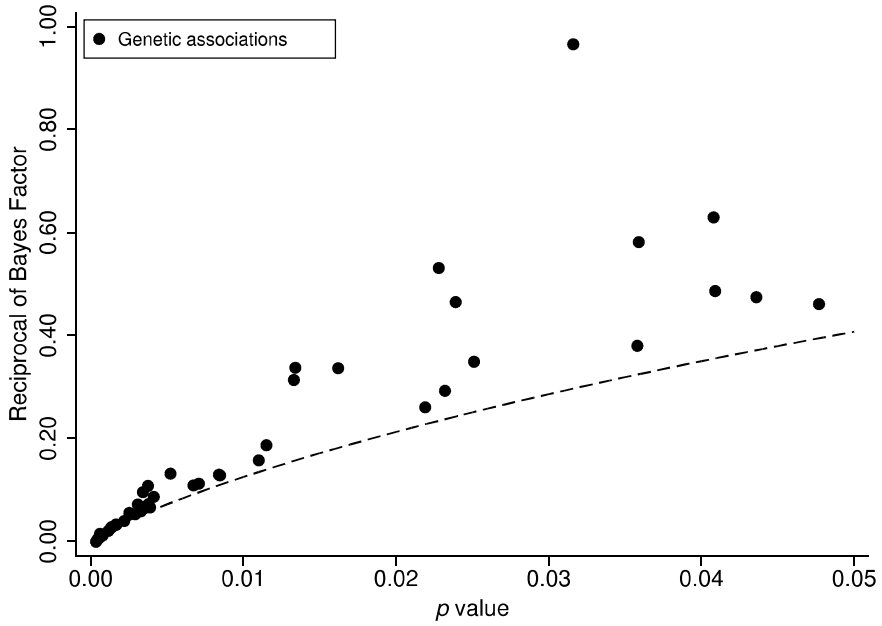
$$B_{01} = \frac{f(x \mid 0)}{\int f(x \mid \theta)\pi(\theta)\, d\theta}\,.$$

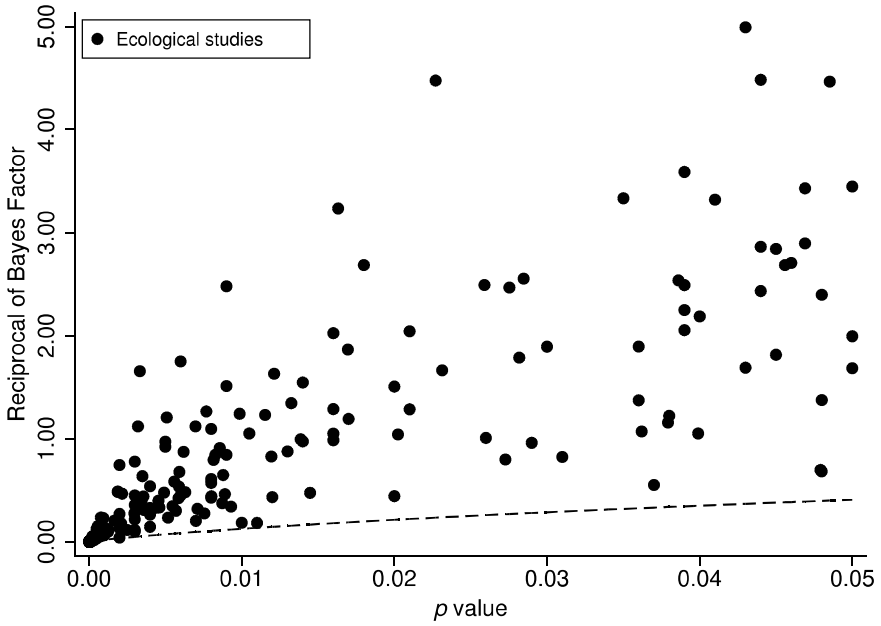
The following investigations compared the p-values from published studies with B_{01}.

- They looked at a large collections of published studies where $0 < p < 0.05$;
- computed B_{01} for each study;
- graphed B_{01} versus the corresponding p-values.
- The first two graphs are for 272 'significant' epidemiological studies with two different choices of the prior; the third for 50 'significant' meta-analyses (these three from J.P. Ioannides, Am J Epidemiology, 2008); and the last is for 314 ecological studies (reported in Elgersma and Green, 2011).


Risk factors
Other
Toxic exposures
Biologic exposures
Psychosocial
Physical exposure
Body characteristics
Dietary
Reciprocal of Bayes Factor
0.00
0.20
0.40
0.60
0.80
1.00
0.00
0.01
0.02
0.03
0.04
0.05
p value


Risk factors
Other
Toxic exposures
Biologic exposures
Psychosocial
Physical exposure
Body characteristics
Dietary
Reciprocal of Bayes Factor
0.00
0.50
1.00
1.50
2.00
0.00
0.01
0.02
0.03
0.04
0.05
p value


Genetic associations
Reciprocal of Bayes Factor
0.00
0.20
0.40
0.60
0.80
1.00
0.00
0.01
0.02
0.03
0.04
0.05
p value


Ecological studies
Reciprocal of Bayes Factor
0.00
1.00
2.00
3.00
4.00
5.00
0.00
0.01
0.02
0.03
0.04
0.05
p value

A quick IP fix of the p-value issue

Determination of the Bayesian odds of H_0 to H_1 can be a challenging problem, because of the typical need for proper priors.

But *robust Bayesian* theory can be used (Sellke, Bayarri and Berger, 2001) to give a bound on the odds of H_0 to H_1, for each given p-value:

Theorem 1 *A proper p-value satisfies* $H_0 : p(X) \sim Uniform(0, 1)$, *so test this versus* $H_1 : p \sim g(p)$, *where* $Y = -\log(p)$ *has a non-increasing failure rate (a natural non-parametric condition on* g, *defining the class* $\mathcal{P}$ *of possible prior probability distributions under* H_1*). Then*

$$B_{01} \geq \inf_{g \in \mathcal{P}} 1/g(p) = -e\, p\, \log(p) \quad \text{for } p < e^{-1}.$$

(Vovk (1993) proved this for $\mathcal{P} = \{\text{Beta}(\xi, 1),\ 0 < \xi < 1\}$.)

p	.2	.1	.05	.01	.005	.001	.0001	.00001
$-ep\log(p)$	.879	.629	.409	.123	.072	.0189	.0025	.00031

Note: This bound is the graphed dotted line in the previous figures.

II. Optimal hyperpriors for normal hierarchical models

(with Chengyuan Song and Dongchu Sun)

For $i = 1, 2, \ldots, m$,

- $\boldsymbol{X}_i = \boldsymbol{\theta}_i + \epsilon_i, \qquad \epsilon_i \sim N_k(\cdot \mid \mathbf{0}, \boldsymbol{\Sigma}_i)$,
  the $\boldsymbol{X}_i$ and $\boldsymbol{\theta}_i$ being $k \times 1$ vectors, $k \geq 2$, with the $\boldsymbol{\Sigma}_i$ known.

  Example: At hospital i,

  - $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{ik})$ is the sample averages of the costs of k different medical treatments;
  - $\boldsymbol{\theta}_i$ is the corresponding unknown vector of true mean costs of the treatments at the hospital;
  - $\boldsymbol{\Sigma}_i$ is the associated (estimated) covariance matrix.

  Note: If $\boldsymbol{X}_i = \boldsymbol{B}_i\boldsymbol{\theta}_i + \epsilon_i$ for given design matrix $\boldsymbol{B}_i$, transform to $\boldsymbol{X}_i^* = (\boldsymbol{B}_i^t\boldsymbol{\Sigma}_i^{-1}\boldsymbol{B}_i)^{-1}\boldsymbol{B}_i^t\boldsymbol{\Sigma}_i^{-1}\boldsymbol{X}_i$, which will be distributed as above.

- $\boldsymbol{\theta}_i = \boldsymbol{z}_i\boldsymbol{\beta} + \epsilon_i^*, \quad \epsilon_i^* \sim N_k(\cdot \mid 0, \boldsymbol{V})$,
  with the $\boldsymbol{z}_i$ being specified $k \times l$ covariate matrices.
  - $\boldsymbol{\beta}$ is an $l \times 1$ unknown 'hyper-mean' vector, $l \geq 2$;
  - $\boldsymbol{V}$ is an unknown $k \times k$ 'hyper-covariance matrix'.

  Example continued: Because all hospitals are related, the $\boldsymbol{\theta}_i$ are assigned a hierarchical prior referring to the 'population' of hospitals.

  The $\boldsymbol{z}_{ij}$ are known covariates, giving hospital i's characteristics for treatment j, such as the number of patients receiving the treatment, the average severity of the condition of the patients, the average income of the patients, etc.

We have specified a class

$$\mathcal{P} = \{\pi(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_m \mid \boldsymbol{\beta}, \boldsymbol{V}), \quad \boldsymbol{\beta} \in \mathcal{R}^k, \ \boldsymbol{V}_{k\times k} \text{ positive definite}\}.$$

Goal: Find good hyperpriors $\pi(\boldsymbol{\beta}, \boldsymbol{V}) = \pi(\boldsymbol{\beta})\pi(\boldsymbol{V})$ (independence assumed).

Recommended prior: After standardizing the covariates, use the hyperprior

$$\pi(\boldsymbol{\beta}) \propto \frac{1}{(1+||\boldsymbol{\beta}||^2)^{(p-1)/2}}, \ \boldsymbol{\beta} \in \mathbb{R}^p,$$

$$\pi(\boldsymbol{V}) \propto \frac{1}{|\boldsymbol{V}|^{1-1/(2k)} \prod_{1\leq i<j\leq k}(v_i - v_j)}, \ \boldsymbol{V} > 0\,,$$

where $v_1 > v_2 > \ldots > v_k$ are the eigenvalues of $\boldsymbol{V}$.

- These are related to *reference priors* which are highly recommended in the objective Bayesian literature.
- The priors can be efficiently implemented with MCMC algorithms.
- Using the priors will result in *admissible* frequentist shrinkage estimators of $(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_m)$, under quadratic loss, and they are the vaguest priors which do so.
- These priors can be used for any means and covariance matrices that occur in any normal hierarchical model, no matter how many levels.

The choice of the hyperprior is important. For instance, in hierarchical normal models the current standard choice of the hyperprior is $\pi(\boldsymbol{\beta}) = 1$ and $\pi(\boldsymbol{V}) = 1$, i.e., the constant prior.

- The constant prior requires more than $2k$ vector observations for posterior propriety, while the recommended prior requires only 2 vector observations.
- The constant prior yields estimates of $(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_m)$ that are much worse:

Table 3. The mean square error of estimates of $(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_m)$ for the Constant prior (C) and the Recommended prior (R) in the following eight scenarios:

$k_1 = 4, m_1 = 10$, $k_2 = 5, m_2 = 15$; $\boldsymbol{\beta}_1 = \mathbf{1}_k$, $\boldsymbol{\beta}_2 = 50\mathbf{1}_k$; $\boldsymbol{V}_1 = \boldsymbol{I}_k$, $\boldsymbol{V}_2 = diag\{8k-7, \ldots, 9, 1\}$

Prior	$k_1\beta_1V_1$	$k_1\beta_2V_1$	$k_1\beta_1V_2$	$k_1\beta_2V_2$	$k_2\beta_1V_1$	$k_2\beta_2V_1$	$k_2\beta_1V_2$	$k_2\beta_2V_2$
C	68.481	71.552	76.541	84.039	111.507	128.434	134.340	145.854
R	42.735	44.746	63.311	76.338	77.129	107.277	123.973	134.529

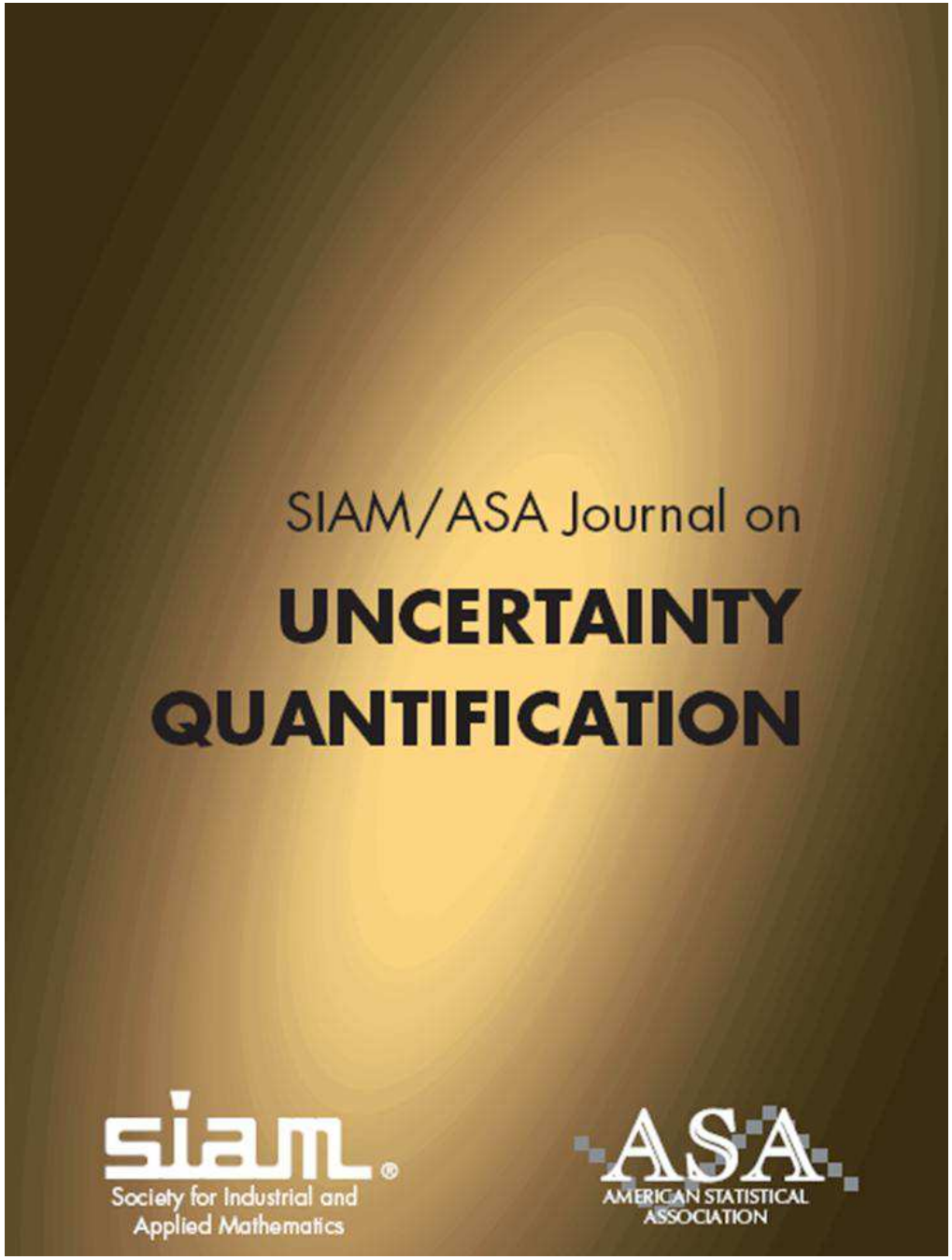
III. Uncertainty Quantification (UQ): Dealing with uncertainties involved in math modeling of processes

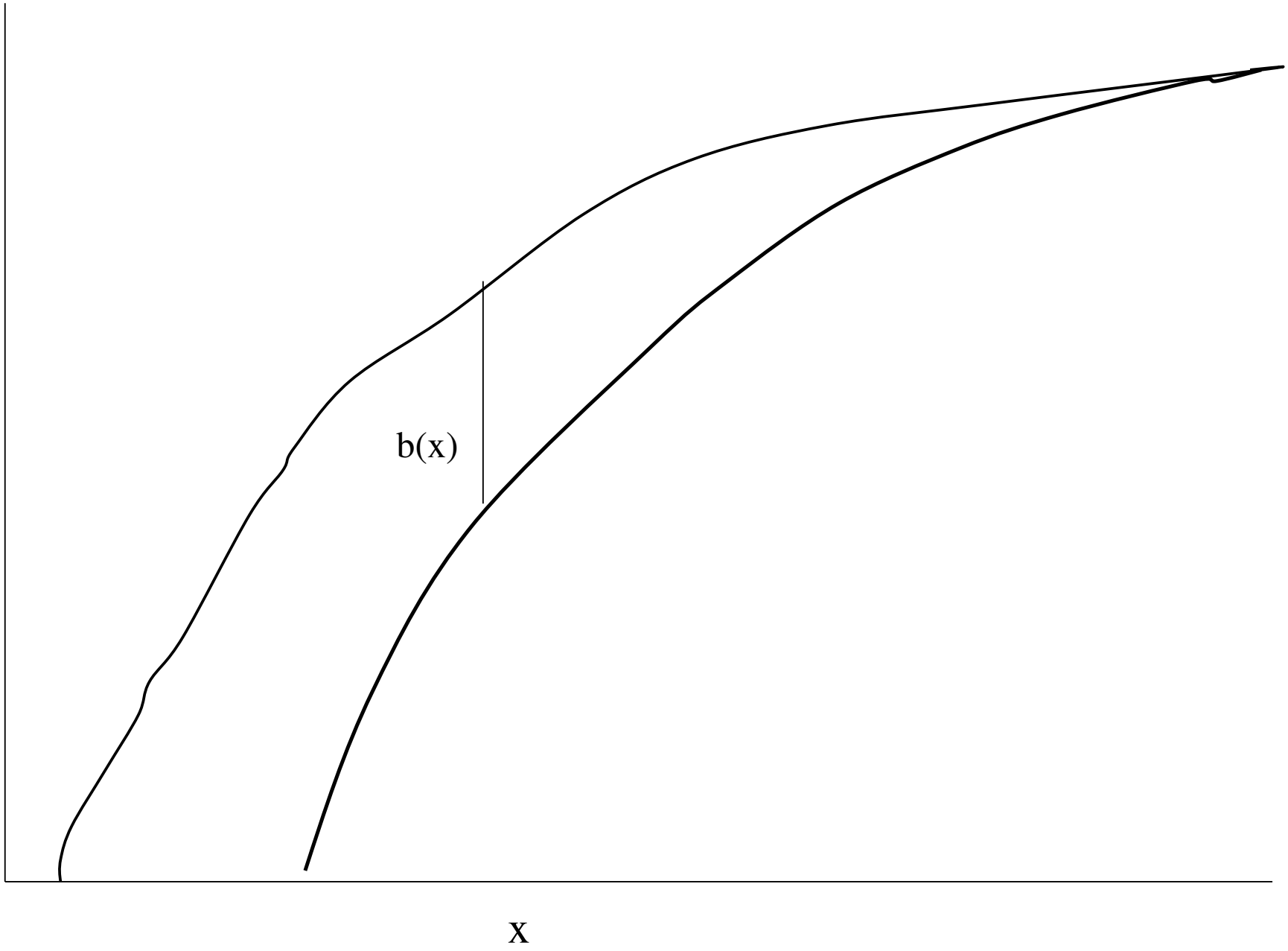
- *Real process:* $y^R(\mathbf{x})$, with input $\mathbf{x}$.
  - Example: $y^R(\mathbf{x})$ is future temperature under carbon forcing $\mathbf{x}$.
- *Computer model output:* $y^M(\mathbf{x}, \mathbf{u})$, with unknown parameters $\mathbf{u}$.
  - Example: y^M is prediction of future temperature from a climate model; $\mathbf{u}$ is the hundreds of unknown parameters in the model.
- *Observations of the real process:* $y^O(\mathbf{x})$
  - Example: Alas, we only have one (partial) observation of climate.
- *Classical formulation:* $y^O(\mathbf{x}) = y^M(\mathbf{x}, \mathbf{u}) + \epsilon$, with ϵ being random error.

Twenty years ago Tony O'Hagan said, "no, the math model is virtually always *imprecise*, so the correct formulation is

$$y^O(\mathbf{x}) = y^M(\mathbf{x}, \mathbf{u}) + \mathbf{b}(\mathbf{x}, \mathbf{u}) + \epsilon\,,$$

where $b(\mathbf{x}, \mathbf{u}) = y^R(\mathbf{x}) - y^M(\mathbf{x}, \mathbf{u})$ is *model discrepancy (bias)*."


SIAM/ASA Journal on
UNCERTAINTY
QUANTIFICATION
siam
Society for Industrial and
Applied Mathematics
ASA
AMERICAN STATISTICAL
ASSOCIATION


b(x)
x

The Jeffreys-Lindley Paradox

Suppose we have n i.i.d. observations from density $f(x \mid \theta)$ and wish to test $H_0 : \theta = 0$ versus $H_1 : \theta \neq 0$.

A Bayesian has fixed (nonzero) prior probabilities on the hypotheses and a fixed bounded prior on θ under H_1.

Let $p(x_1, \ldots, x_n)$ be the p-value for the test, corresponding (say) to the likelihood ratio test.

The 'Paradox': For any fixed p (e.g., $p = 10^{-10}$), $Pr(H_0 \mid x_1, \ldots, x_n) \to 1$ as $n \to \infty$.

Model bias: Rarely is $f(x \mid \theta)$ exactly true. Suppose, instead, that $x_i - b$ has density $f(\cdot \mid \theta)$, where b reflects a model bias (alternatively, suppose that H_0 is really $H_0 : |\theta| < b$.) This bias won't be known, but its prior distribution will typically lie in the class

$\mathcal{P} = \{\text{all prior distributions that do not have a postive probability mass at } 0\}$.

Result: For any $\pi \in \mathcal{P}$ and fixed p, $Pr(H_0 \mid x_1, \ldots, x_n, \pi) \to c < 1$ as $n \to \infty$.

Example: UQ for a computer model of road-load dynamics

(with M.J. Bayarri, J. Cafeo, G. Garcia-Donato, F. Liu, J. Palomo, R.J. Parthasarathy, R. Paulo, J. Sacks, and D. Walsh: *AOS, 2007*)

Consider a vehicle being driven over a road with two major potholes.

- $\boldsymbol{x} = (x_1, \ldots, x_7)$ is the vector of key vehicle characteristics, unknown because of manufacturing variability.
- $y^R(\boldsymbol{x}; t)$ is the time-history curve of resulting forces.

A finite element PDE computer model of the vehicle being driven over the road

- depends on $\boldsymbol{x} = (x_1, \ldots, x_7)$ and unknown calibration parameters $\boldsymbol{u} = (u_1, u_2)$;
- yields time-history force curve $y^M(\boldsymbol{x}, \boldsymbol{u}; t)$.

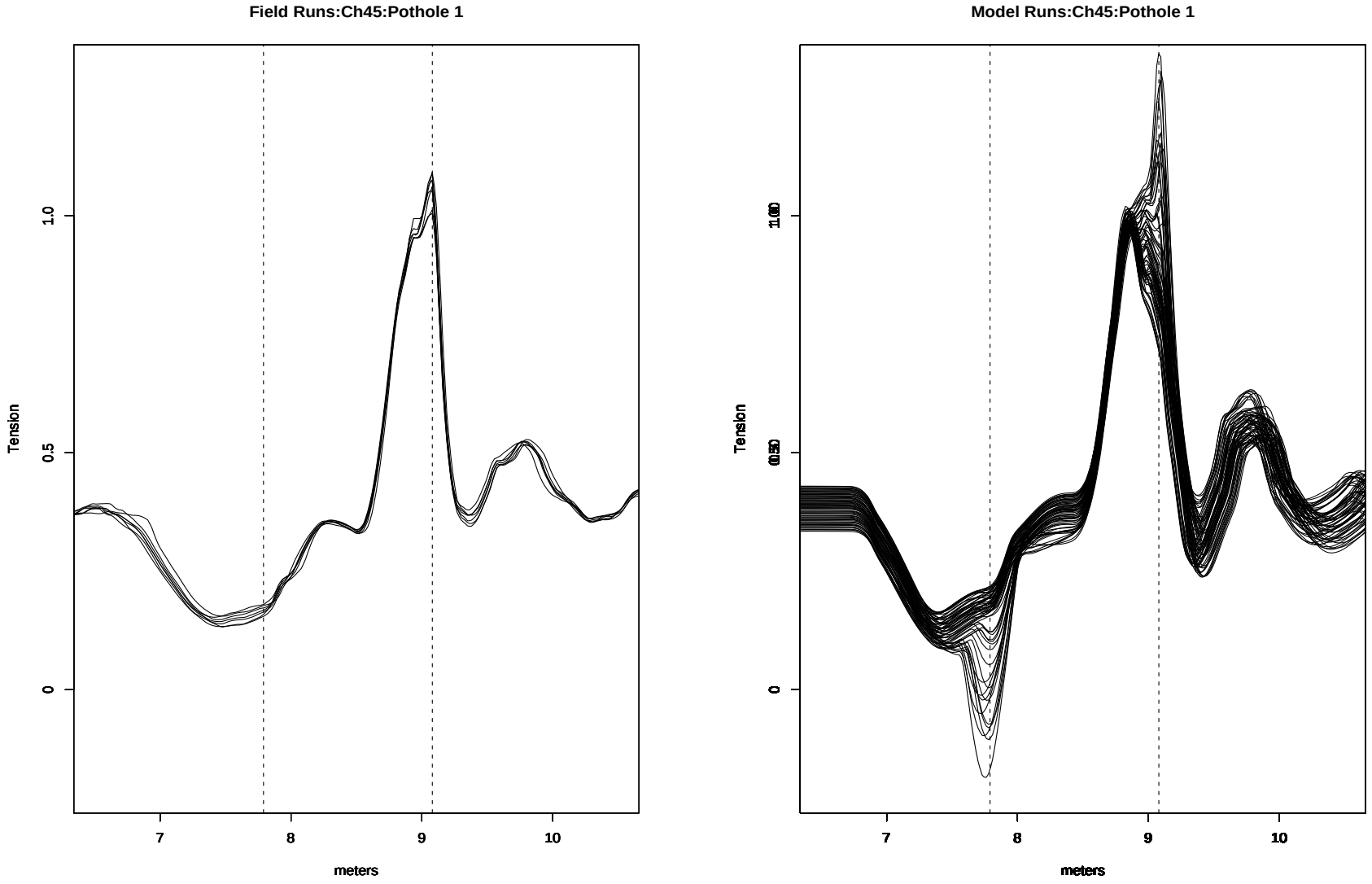


Figure 1: Force curves from 7 field runs (left) and 65 computer model runs (right) for one of the potholes.

Analysis proceeded by

- registration (aligning) of field and model force curves;
- employing a wavelet representation of all force curves, so that

$$y^M(\boldsymbol{x},\boldsymbol{u};t) \approx \sum_{i=1}^{289} w_i^M(\boldsymbol{x},\boldsymbol{u})\psi_i(t), \quad y_r^F(\boldsymbol{x}^*;t) \approx \sum_{i=1}^{289} w_{ir}^F(\boldsymbol{x}^*)\psi_i(t)\,,$$

  where the $w_i^M(\boldsymbol{x},\boldsymbol{u})$ and $w_{ir}^F(\boldsymbol{x}^*)$ are the coefficients computed through the wavelet decomposition and the $\psi_i(t)$ are known basis functions;
- introducing model bias having a zero mean Gaussian process prior;
- for computational reasons and to allow inference at new inputs, replacing each $w_i^M(\boldsymbol{x},\boldsymbol{u})$ by an emulator (a Gaussian process approximation to that part of the computer model);
- assigning priors (a mix of subjective and objective) to all unknowns;
- employing modularized MCMC to determine the posterior distribution of all unknowns and to make future predictions.

Parameter	**Type**	**Uncertainty Range**
$Damping_1$	Calibration	[0.125, 0.875]
$Damping_2$	Calibration	[0.125, 0.875]
x_1	Nominal+Variation	[0.1667, 0.8333]
x_2	Nominal+Variation	[0.1667, 0.8333]
x_3	Nominal+Variation	[0.2083, 0.7917]
x_4	Nominal+Variation	[0.1923, 0.8077]
x_5	Nominal+Variation	[0.3529, 0.6471]
x_6	Nominal+Variation	[0.1471, 0.8529]
x_7	Nominal+Variation	[0.1923, 0.8077]

Table 1: Uncertainty ranges for calibration parameters and parameters subject to manufacturing variation.

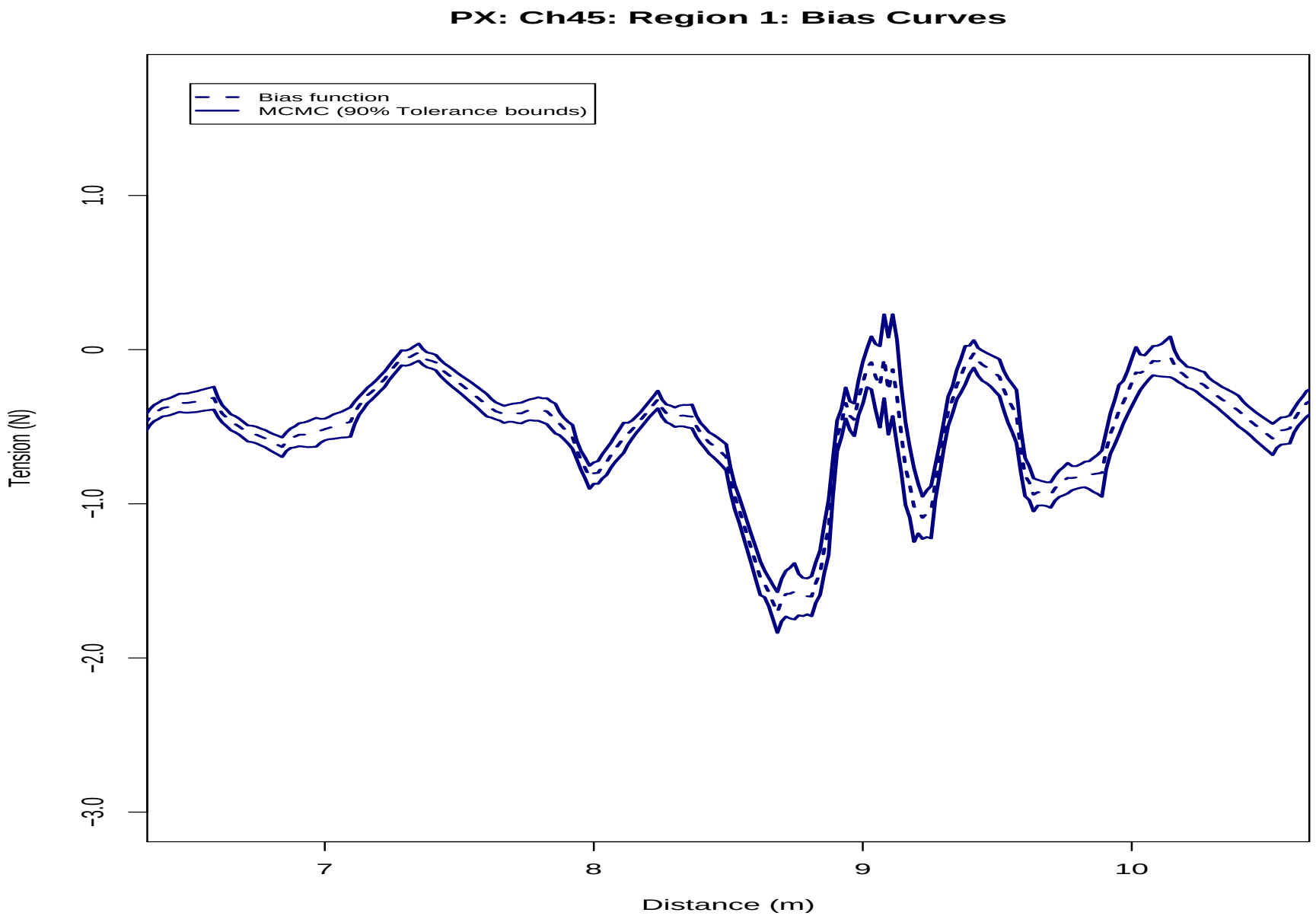


Figure 2: Posterior bias curve estimate and 90% tolerance bands.

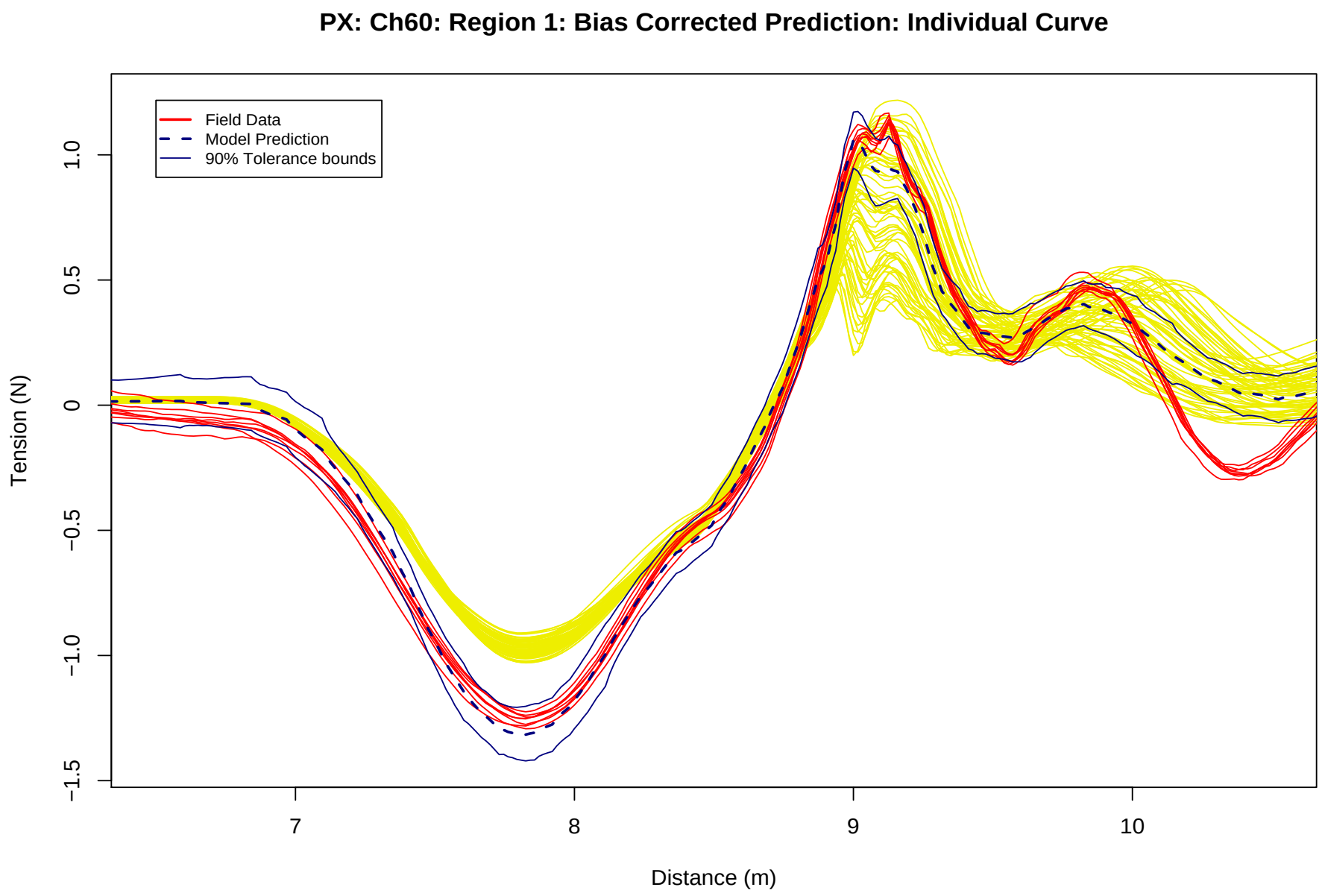


Figure 3: Extrapolation of bias to predict the force curve for Vehicle B.

Thanks!