Bayesian Calibration of Computer Models

Marc C. Kennedy, Anthony O'Hagan

University of Sheffield, UK

Journal of the Royal Statistical Society B, 2001

Presentation: Yuanhao Zhu April 23, 2025

Overview: Computer Models and Calibration

- Computer models are widely used for complex processes where direct observation is much more expensive or infeasible.
- Computer models are generally applicable to a wide range of contexts, but in a specific content it's necessary to *calibrate* by using observed data.
- Computer models have unknown *context-specific* inputs that define a particular situation that the model being used. Calibration learns them.
- In current practice, calibration consists of *searching a set of values* of the unknown inputs that the observed data fit as closely as possible, then do 'plug-in'.
- The inputs are only estimated, the residual uncertainty should be recognized in subsequent predictions.

Bayesian Approach to Calibration

- Bayesian calibration drives the posterior distribution of unknown input θ give the observed data, improves on traditional calibration by :
  1. Explicitly quantifying uncertainty in parameter estimates.
  2. Accounting for model inadequacy (discrepancy between model predictions and real data).
- Computer codes are treated as "black boxes," avoiding complex internal mathematical analysis.
- More detailed methods may exist but would require deeper model insight and complexity.

Outline of the Paper

- Section 2: Statistical analysis of uncertainties in computer code outputs.
- Section 3: Review of related Bayesian theories and literature.
- Section 4: Bayesian calibration methodology detailed.
- Section 5-6: Practical applications and case studies.
- Section 7: Conclusions and directions for future research.

Classification of Uncertainties in Computer Models

Six primary sources of uncertainty:

1. Parameter uncertainty
2. Model inadequacy
3. Residual variability
4. Parametric variability
5. Observation error
6. Code uncertainty

Parameter Uncertainty and Model Inadequacy

Parameter uncertainty:

- Uncertainty regarding correct input parameter values across different contexts.

Model inadequacy:

- Discrepancy arises when models do not perfectly represent real processes, even if true input values are known.
- Defined as the difference between true mean values and model outputs.

Residual and Parametric Variability

Residual variability:

- Variability in real processes under identical conditions due to stochastic processes or unrecognized conditions.

Parametric variability:

- Arises when some input conditions are intentionally unspecified, leading to uncertainty propagated through parameter distributions.

Observation Error and Code Uncertainty

Observation error:

- Errors present in observed data; these must be accounted for in calibration.

Code uncertainty:

- Uncertainty in outputs because running the code with every possible configuration is impractical.
- Acknowledges that output for given inputs might remain practically unknown.

Gaussian Plume Model (Clarke, 1979)

- Predicts radioactive dispersion following an accidental release.
- Not directly observable, not available to run complex models.
- Inputs divided into atmospheric conditions (wind direction, speed, etc.) and characteristics of the release (source term, location, duration, etc.).
- Simplifications lead to substantial model inadequacy (e.g., assuming constant wind conditions).

Gaussian Plume Model (Clarke, 1979)

- Important uncertain parameters include source term and deposition velocity are very difficult to determine.
- The model is cheap so can make thousands of runs. Thus code uncertainty is not an issue.
- Measurements of deposition is not straightforward, which should be introduced at each stage of the measurement process.

Hydrological Model (Romanowicz et al., 1994)

- Models ground-water flow and contaminants' movement through soil.
- Inputs include rainfall data, evapotranspiration rate, transmissivity of soil, and subsurface drainage.
- Uncertainties:
  - Parameter uncertainty (transmissivity, drainage constant).
  - Model inadequacy due to simplifying assumptions.
  - Residual variability as true flow values are averaged due to heterogeneous flow patterns (point measurement).
  - Measurement error (rainfall data).

Statistical Methods: Interpolation Approach

- Earlier statistical analysis mainly involved interpolation for untested input configurations, when the code is large and expensive to run.
- Only accounted explicitly for code uncertainty.
- Did not address model inadequacy, residual variability, or observational errors because there is no attempt to predict the real process.

Uncertainty Analysis

Monte Carlo:

- Random sampling of inputs from probability distributions.
- Becomes impractical when the code is costly.
- Latin hypercube sampling offers efficiency gains (McKay et al., 1979; Stein, 1987; Owen, 1992).

Bayesian Approach:

- Haylock and O'Hagan (1996) introduced Bayesian Gaussian process models for uncertainty analysis.
- Consider parameter uncertainty and parametric variation, but still focused on the code output not the process itself, lack of model inadequacy, residual variation and observation error.

Sensitivity Analysis

- Explores how model outputs respond to variations in inputs.
- Sensitivity analysis typically heuristic, with limited statistical rigor.
- Rigorous approaches presented by Saltelli et al. (2000).
- As with interpolation, these statistical approaches only take account of code uncertainty.

Calibration

- Traditional way by ad hoc search contains observation errors, residual variation and model inadequacy but only implicitly through the measure of the discrepancy in fits.
- The estimated values are treated if they were known, the subsequent predictions take no account of the remaining parameter uncertainty.

Generalized Likelihood Uncertainty Estimation (GLUE)

- Romanowicz et al. (1994): effectively Bayesian approach.
- Allow for code uncertainty from only a sample of tuns.
- Allow for parametric variation from drawing from the unspecified inputs at the prediction stage.
- Does not rigorously address model inadequacy, residual variation and observational errors explicitly, still estimate the code output rather than reality.

Other Bayesian Approaches

- Craig et al. (1996, 2001): iterative Bayesian calibration, influenced by Bayes linear methods (Goldstein, Wooff).
- Cox et al. (1992): Replace and expensive code with Gaussian process interpolator as a cheaper alternative to full simulation.
- Raftery et al. (1995): Bayesian synthesis criticized by Wolpert (1995) and Schweder and Hjort (1996); alternative method, Bayesian melding (Poole & Raftery, 1998).
- **Limitations**: Does not account for remaining parameter uncertainty, neither model explicitly recognizes inadequacy and simplify code uncertainty.

Contribution of This Paper

- First explicit and comprehensive Bayesian approach addressing *all* sources of uncertainty.
- Not fully Bayesian because hyperparameters are *estimated* from posterior modes (approximate inference).
- Acknowledges possible further improvement with full hyperparameter uncertainty modeling.

Gaussian Processes

- $f(\cdot)$ has a Gaussian Process if the joint distribution of $f(\mathbf{x_1}), ..., f(\mathbf{x_n})$ is multivariate normal for all $\mathbf{x_1}, ..., \mathbf{x_n} \in \mathcal{X}, n = 1, 2, 3, ...$
- Notation: $f(\cdot) \sim N(m(\cdot), c(\cdot, \cdot))$.
- Characterized by mean function $m(\mathbf{x}) = E[f(\mathbf{x})]$ and covariance function $c(\mathbf{x}, \mathbf{x}')$.
- Widely used in modern Bayesian statistical modeling.

Gaussian Processes in Practice

- The use of GPs dates back to represent prior distributions, model a regression function in a nonparametric way (Kimeldorf and Wahba, 1970; O'Hagan, 1978).
- Historically known as "kriging" in geostatistics, widely applied in spatial modeling.

Modeling Issues with Gaussian Processes

- Practical choice for modeling unknown functions due to realism and flexibility.
- Assumption of joint normality generally reasonable, though transformations may enhance modeling.
- Widely applicable across various fields beyond geostatistics, including computer experiments.

Mean and Covariance Functions

- Gaussian processes characterized by mean $m(\cdot)$ and covariance $c(\cdot,\cdot)$ which has to be PSD.
- Hierarchical mean:

$$m(\cdot) = \mathbf{h}(\cdot)^T\boldsymbol{\beta}, \quad f(\cdot) = m(\cdot) + e(\cdot) = \mathbf{h}(\cdot)^T\boldsymbol{\beta} + e(\cdot)$$

  where $\mathbf{h}$ are known, $\boldsymbol{\beta}$ are unknown given a prior distribution. $e(\cdot)$ is a zero-mean GP with covariance $c(\cdot,\cdot)$.
- Hierarchical variance (stationary assumption):

$$c(x,x') = \sigma^2 r(x-x'), \quad r(x-x') = \exp\left\{-\sum_{j=1}^{q}\omega_j(x_j-x_j')^2\right\}$$

- Could replaces $(\cdot)^2$ by $|\cdot|^\alpha$.

Bayesian Nonparametric Alternatives

- Gaussian processes: nonparametric priors (semiparametric).
- Alternative methods:
  - Basis functions (splines, wavelets, neural networks).
  - Piecewise linear processes (Liu & Arjas, 1998).
- Connections to neural networks (Neal, 1996):
  - Infinite hidden-layer neural networks equivalent to Gaussian processes.
  - Posterior mean as weighted sum of basis functions formed by the correlation functions centered at the observations.

Calibration Inputs vs Variable Inputs

- Calibration problem involves two input types:
  1. **Calibration inputs (θ):** Unknown context-specific parameters inferred from observations.
  2. **Variable inputs (x):** Known and potentially varying parameters in practical use.
- $\mathbf{t}$ as the known calibration inputs (distinguish between θ), $\zeta(\mathbf{x})$ to be the true value of the real process when the variable inputs take values $\mathbf{x}$.
- Observations used for calibration:

$$z_i = \zeta(x_i) + e_i, \quad y_j = \eta(x_j^*, t_j)$$

where z_i real observations, y_j computer code outputs.

Model Representation

- Relation between observations, reality, and model:

$$z_i = \zeta(x_i) + e_i = \rho\,\eta(x_i, \boldsymbol{\theta}) + \delta(x_i) + e_i$$

  - $e_i \sim N(0, \lambda)$: Observation error, including residual variability.
  - $\delta(x)$: Model inadequacy term independent of η.
  - ρ: Unknown regression parameter linking model to reality.
- Separation of residual variation and observation error challenging due to lack of replicated observations (Do not imagine having replication of all the unrecognized conditions).

Model Assumptions

- Implication
$$\zeta(x) = \rho\,\eta(x, \boldsymbol{\theta}) + \delta(x)$$
- Markov assumption (O'Hagan, 1998): Predicting $\zeta(\mathbf{x}')$ is sufficient to observe $\eta(\mathbf{x}', \theta)$ of a single run at $\mathbf{x}'$.
- ρ is constant.
- $\eta(\cdot,\cdot), \delta(\cdot), \zeta(\cdot)$ are stationary processes.

Prior Distributions and Hierarchical Model

- Gaussian processes for $\eta(\cdot, \boldsymbol{\theta})$ and $\delta(\cdot)$:

$$\eta(\cdot, \boldsymbol{\theta}) \sim N[m_1(\cdot), c_1(\cdot, \cdot)], \quad \delta(\cdot) \sim N[m_2(\cdot), c_2(\cdot, \cdot)]$$

- Mean functions given by linear models:

$$m_1(x, t) = \mathbf{h}_1(x, t)^T \boldsymbol{\beta}_1, \quad m_2(x) = \mathbf{h}_2(x)^T \boldsymbol{\beta}_2$$

- Weak priors:

$$p(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) \propto 1$$

- Prior independence assumption:

$$p(\boldsymbol{\theta}, \boldsymbol{\beta}, \phi) = p(\boldsymbol{\theta})p(\phi)$$

where ϕ includes hyperparameters ρ, λ, ψ, ψ determines c.

Meaning of True Parameter Values

- Concept of *true parameter* analogous to non-linear regression:
  - θ represents "best-fitting" parameter set.
  - Defined according to error structure of residuals.
- True physical values may differ from best-fitting parameters:
  - Physical true values may lead to worse fit and poorer predictions.
  - Fixing a parameter decreases the flexibility and may lead to worse fits, even the parameter is physically true.
  - Reasonable to treat an input as unknown even the true value is known. Allowing the influential parameter deviate from the true value may produce empirically better results.

Posterior Distribution of Parameters

- Set $D_1 = \{(\mathbf{x}_1^*, \mathbf{t}_1), ..., (\mathbf{x}_N^*, \mathbf{t}_N)\}$, $D_2 = \{\mathbf{x}_1, ..., \mathbf{x}_N\}$.
- Define $D_2(\boldsymbol{\theta}) = \{(\mathbf{x}_1, \boldsymbol{\theta}), ..., (\mathbf{x}_N, \boldsymbol{\theta})\}$. $\mathbf{H}_1(D_1)$ denote the matrix with rows $\mathbf{h}_1(\mathbf{x}_1^*, \mathbf{t}_1)^T, ..., \mathbf{h}_1(\mathbf{x}_N^*, \mathbf{t}_N)^T$.
- Observational data vector $\mathbf{d}^T = (\mathbf{y}^T, \mathbf{z}^T)$ is normally distributed given:

$$E(\mathbf{y}) = \mathbf{H_1}(D_1)\beta_1, E(\mathbf{z}) = \rho\mathbf{H}_1\{D_2(\boldsymbol{\theta})\}\beta_1 + \mathbf{H}_2(D_2)\beta_2$$

$$E(\mathbf{d}|\boldsymbol{\theta}, \boldsymbol{\beta}, \phi) = \mathbf{m}_d(\boldsymbol{\theta}) = \mathbf{H}(\boldsymbol{\theta})\boldsymbol{\beta}$$

with

$$\mathbf{H}(\boldsymbol{\theta}) = \begin{pmatrix} \mathbf{H}_1(D_1) & \mathbf{0} \\ \rho\mathbf{H}_1\{D_2(\boldsymbol{\theta})\} & \mathbf{H}_2(D_2) \end{pmatrix}.$$

Posterior Variance Structure

Define

- $\mathbf{V}_1(D_1)_{ij} = c_1\{(\mathbf{x}_i^*, \mathbf{t}_i), (\mathbf{x}_j^*, \mathbf{t}_j)\}$
- $\mathbf{V}_1(D_2(\boldsymbol{\theta})), \mathbf{V}_2(D_2)$ similarly as $\mathbf{V}_1(D_1)$
- $\mathbf{C}_1\{D_1, D_2(\boldsymbol{\theta})\}_{ij} = c_1\{(\mathbf{x}_i^*, \mathbf{t}_i), (\mathbf{x}_j, \boldsymbol{\theta})\}$

Then

$$\text{var}(\mathbf{d}|\boldsymbol{\theta}, \boldsymbol{\beta}, \phi) = \mathbf{V}_d(\boldsymbol{\theta}) = \begin{pmatrix} \mathbf{V}_1(D_1) & \rho\,\mathbf{C}_1\{D_1, D_2(\boldsymbol{\theta})\}^\top \\ \rho\,\mathbf{C}_1\{D_1, D_2(\boldsymbol{\theta})\} & \lambda\mathbf{I}_n + \rho^2\mathbf{V}_1\{D_2(\boldsymbol{\theta})\} + \mathbf{V}_2(D_2) \end{pmatrix}$$

- Full joint posterior distribution expressed as:

$$p(\boldsymbol{\theta}, \boldsymbol{\beta}, \phi|\mathbf{d}) \propto p(\boldsymbol{\theta})p(\phi)f\{\mathbf{d}; \mathbf{m}_d(\boldsymbol{\theta}), \mathbf{V}_d(\boldsymbol{\theta})\}.$$

Estimation of Hyperparameters

- Full Bayesian computationally intensive for hyperparameters ϕ (Hard to do integrate even numerically).
- Two-stage approach proposed:
  - Stage 1: observational data $\mathbf{z}$ used to estimate hyperparameters ψ_1.
  - Stage 2: Fixψ_1, using $\mathbf{z}$, estimate ρ, λ, and remaining hyperparameters ψ_2.
- 1. Fixing $\lambda \implies$ not fully observation error and residual variation.
  2. Fixing $\rho, \psi_2 \implies$ not fully model inadequacy.
  3. Fixing $\psi_1 \implies$ not fully code uncertainty.
- Only ignore the 'second-order uncertainties', therefore still captures the major uncertainties.

Calibration, Prediction, and Uncertainty Analysis

- After hyperparameter estimation,

$$p(\boldsymbol{\theta}|\boldsymbol{\phi}, \mathbf{d}) \propto p(\boldsymbol{\theta}) f\{\mathbf{d}; \mathbf{m}_d(\boldsymbol{\theta}), \mathbf{V}_d(\boldsymbol{\theta})\}.$$

- Focus on prediction rather than inference about $\boldsymbol{\theta}$ itself.
- Posterior predictive distribution of true process $\zeta(\cdot)$:

$$E[\zeta(x)|\boldsymbol{\theta}, \boldsymbol{\phi}, \mathbf{d}] = \mathbf{h}(x, \boldsymbol{\theta})^\top \hat{\boldsymbol{\beta}}(\boldsymbol{\theta}) + \mathbf{t}(x, \boldsymbol{\theta})^\top \mathbf{V}_d(\boldsymbol{\theta})^{-1}\{\mathbf{d} - \mathbf{H}(\boldsymbol{\theta})\hat{\boldsymbol{\beta}}(\boldsymbol{\theta})\},$$

where

$$\mathbf{h}(\mathbf{x}, \boldsymbol{\theta}) = \begin{pmatrix} \rho \mathbf{h}_1(\mathbf{x}, \boldsymbol{\theta}) \\ \mathbf{h}_2(\mathbf{x}) \end{pmatrix}, \mathbf{t}(\mathbf{x}, \boldsymbol{\theta}) = \begin{pmatrix} \rho \mathbf{V}_1\{(\mathbf{x}, \boldsymbol{\theta}), D_1\} \\ \rho^2 \mathbf{V}_1\{(\mathbf{x}, \boldsymbol{\theta}), D_2(\boldsymbol{\theta})\} + \mathbf{V}_2(\mathbf{x}, D_2) \end{pmatrix}$$

Calibration, Prediction, and Uncertainty Analysis

- Covariance structure of predictions explicitly defined as:

$$\begin{aligned}\mathrm{cov}\{\zeta(x),\zeta(x')|\boldsymbol{\theta},\phi,\mathbf{d}\} &= \rho^2 c_1\{(x,\boldsymbol{\theta}),(x',\boldsymbol{\theta})\} + c_2(x,x') \\ &- \mathbf{t}(x,\boldsymbol{\theta})^\top \mathbf{V}_d(\boldsymbol{\theta})^{-1}\mathbf{t}(x',\boldsymbol{\theta}) \\ &+ [\mathbf{h}(x,\boldsymbol{\theta}) - \mathbf{H}(\boldsymbol{\theta})^\top \mathbf{V}_d(\boldsymbol{\theta})^{-1}\mathbf{t}(x,\boldsymbol{\theta})]^\top \mathbf{W}(\boldsymbol{\theta}) \\ &\times [\mathbf{h}(x',\boldsymbol{\theta}) - \mathbf{H}(\boldsymbol{\theta})^\top \mathbf{V}_d(\boldsymbol{\theta})^{-1}\mathbf{t}(x',\boldsymbol{\theta})]\end{aligned}$$

  where $\mathbf{W}(\boldsymbol{\theta}) = [\mathbf{H}(\theta)^\top \mathbf{V}_d(\boldsymbol{\theta})^{-1}\mathbf{H}(\theta)]^{-1}$.
- Allows inference and uncertainty propagation using numerical integration.

Calibration, Prediction, and Uncertainty Analysis

- Uncertainty analysis is to study the extra uncertainty in model outputs induced by the parametric variability.
- Suppose now variable inputs X are random, having a distribution $G_X(x)$,to make inference about the distribution of $\zeta(X)$:

$$K = \mathbb{E}_X[\zeta(X)] = \int_X \zeta(x) dG_X(x),$$

the variance:

$$L = \mathrm{var}_X[\zeta(X)] = \int_X \zeta(x)^2 dG_X(x) - K^2,$$

Inference about the distribution of $\zeta(X)$ may be derived from the posterior of $\zeta(\cdot)$ (Kennedy & O'Hagan, 2000b).

Implementation Details: Design Issues

- Choice of observational and calibration data points critical.
- Calibration inputs should span plausible parameter ranges.
- Sequential design strategies recommended:
  - Start with prior plausible values.
  - Expand to values informed by posterior distributions.
- There should be $\mathbf{x}_i^* \in D_1$ close to $\mathbf{x_j} \in D_2$ to learn the relationship between code and reality.
- E.g., Cartesian product of D_1 and Latin hypercube designs for calibration inputs.

Modelling Choices

- Gaussian process structure requires careful specification:
  - Contrast with parametric regression, GP will do better if spurious regressor not included.
  - Generally chosen parsimoniously with defaults:

$$\mathbf{h}_1(\mathbf{x}, t) = 1, \quad \mathbf{h}_2(\mathbf{x}) = 1$$

- Covariance functions typically Gaussian form:

$$\begin{aligned} c_1\{(\mathbf{x}, t), (\mathbf{x}', t')\} &= \sigma_1^2 \exp[-(\mathbf{x} - \mathbf{x}')^\top \mathbf{\Omega}_x (\mathbf{x} - \mathbf{x}')] \\ &\quad \times \exp[-(t - t')^\top \mathbf{\Omega}_t (t - t')] \\ c_2(\mathbf{x}, \mathbf{x}') &= \sigma_2^2 \exp[-(\mathbf{x} - \mathbf{x}')^\top \mathbf{\Omega}_x^* (\mathbf{x} - \mathbf{x}')] \end{aligned}$$

Assumptions

Some assumptions used up to now

- Gaussian forms for the covariance imply
  - Differentiability of $\eta(\cdot,\cdot)$ and $\delta(\cdot)$ implies the same beliefts about $\zeta(\cdot)$.
  - Separability between calibration and variable inputs in the covariance.
- The underlying assumption of stationary.

Using these assumptions allow analytical results that save computation times, essentially for convenience and simplicity.

Computational Considerations

- Computational cost mainly due to numerical integration and inversion of covariance matrix $\mathbf{V}_d(\boldsymbol{\theta})$.
- Computational savings achieved by the code design D_1 such as Cartesian product form.
- Gauss-Hermite quadrature method when dimension of $\boldsymbol{\theta}$ is low.
- For high dimensional $\boldsymbol{\theta}$, may use simulation methods of integration that the authors didn't try.

Tomsk Data Example

- Case study: radioactive deposition from Tomsk-7 chemical plant (1993).
- Data: 695 observations of ^{106}Ru deposition.
- Gaussian plume model for log deposition ($\zeta(\mathbf{x})$):
  - Logarithmic transformation to meet normality assumptions.
  - Calibration parameters : source term, deposition velocity (log).
  - Inputs: distances downwind and plume centre line (coordinates).
- Normal prior distributions obtained from National Radiological Protection Board.
- Calibration and uncertainty analysis based on subsets (10, 15, 20, 25 points) reasonably close to the source and relatively dispersed.

Tomsk Data Example

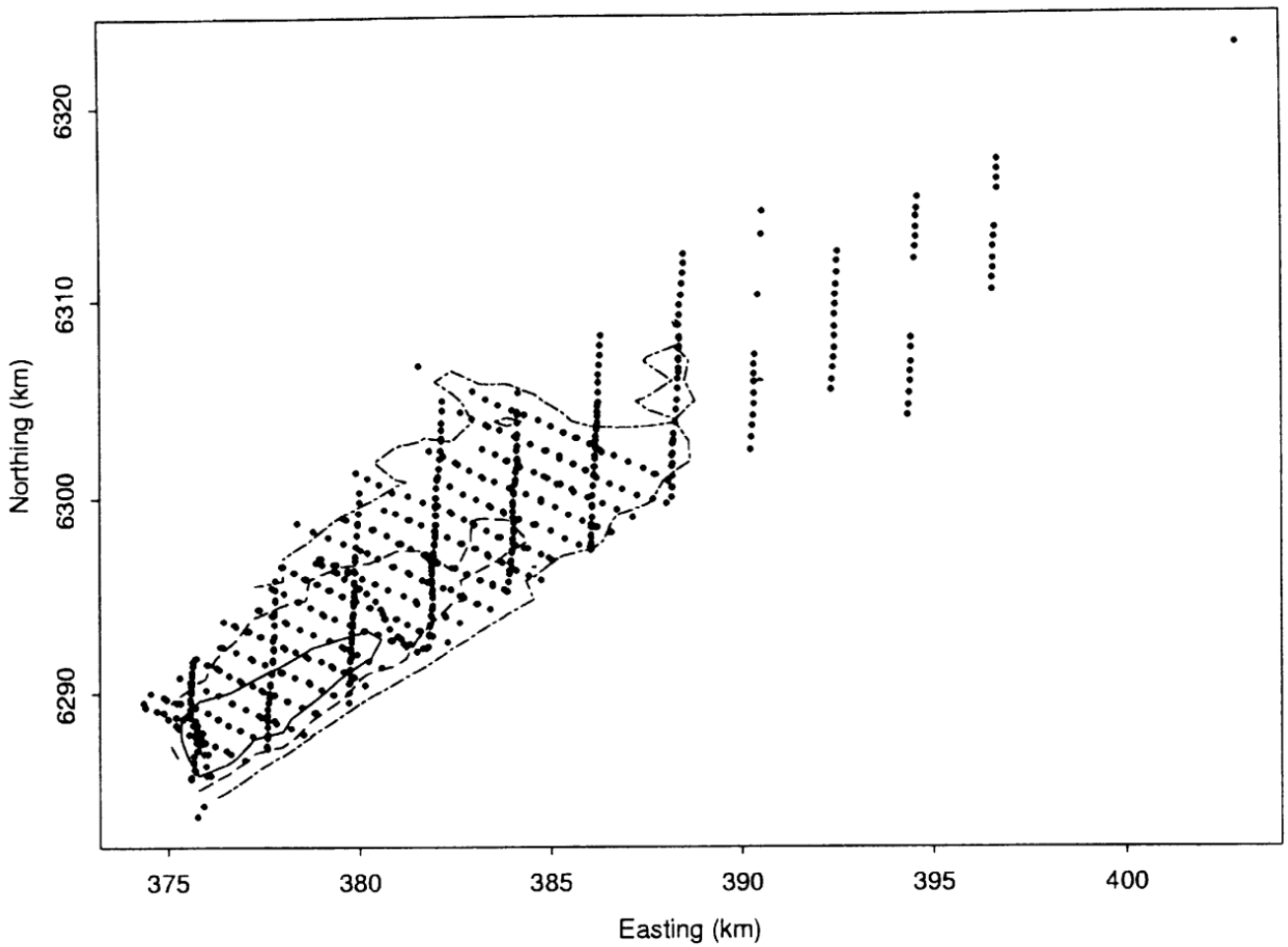

Fig. 1. Tomsk aerial survey of 695 ^{106}Ru deposition measurements, with contours at heights of 11 (———), 10 (– – –) and 9 (— · —)

Comparison of Calibration Strategies

- Three calibration strategies compared:
  1. GP interpolation of physical observations only.
  2. Bayesian calibration with model inadequacy.
  3. Gaussian plume with plug-in input parameters that min SSE.
- RMSE (0.84 using the input parameters fixed ar prior mean):

Strategy	$n = 10$	$n = 15$	$n = 20$	$n = 25$
1	0.75	0.76	0.86	0.79
2	0.42	0.41	0.37	0.36
3	0.82	0.79	0.76	0.66

- Bayesian calibration (strategy 2) significantly reduces prediction error, demonstrating practical improvement.

Comparison of Calibration Strategies

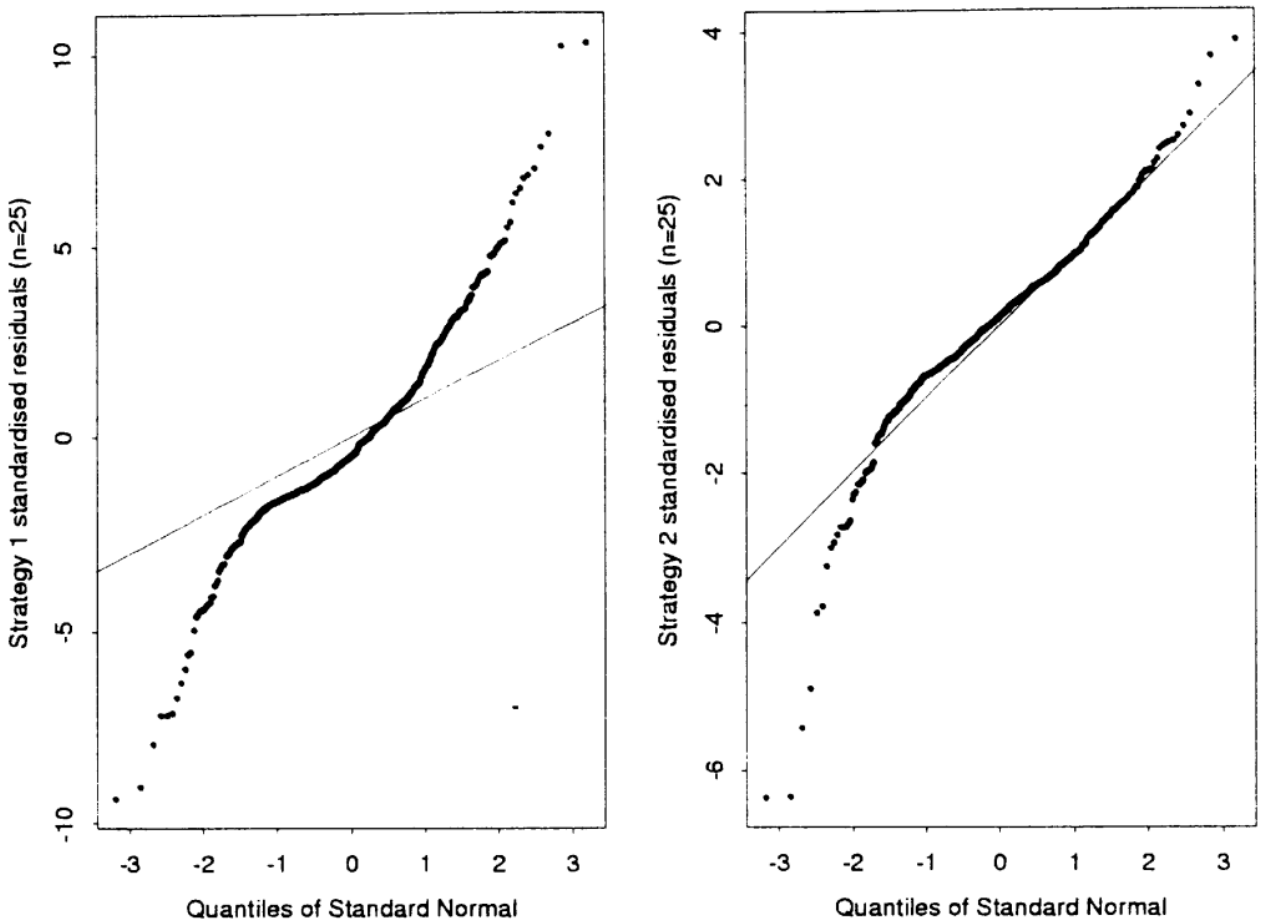

Fig. 2. Quantile–quantile plots for strategies 1 and 2 with $n = 25$

Residual Analysis

- Quantile–quantile plots of standardized residuals ($n = 25$):
  - Strategy 1 shows poorer fit with heavy-tailed residuals.
  - Strategy 2 offers improved fit, still slightly heavy-tailed.
- Residual analysis supports the Bayesian calibration approach, particularly in sparse-data contexts.

Sensitivity to Modelling Assumptions

- Examined robustness to alternative modelling assumptions:
  - Model M1: Strategy 2.
  - Model M2: Integration over roughness parameters.
  - Model M3: Isotropic Matérn correlation.
  - Model M4: Isotropic Gaussian correlation.
- Findings:
  (a) Integration vs. maximization of hyperparameters had minor effect.
  (b) Alternative covariance structures also showed minor effects.
- Predictive distributions retained some heavy-tail characteristics, suggesting local inadequacies in covariance modelling.

Conclusions and Further Work

- Bayesian calibration framework effectively integrates uncertainty from multiple sources by observation from real process, subsequent prediction and uncertainty analysis.
- Methodology generalizable to arbitrarily complex code as "black boxes".
- Key open issues:
  - Optimal choice of calibration and observational design points.
  - Efficient integration in higher-dimensional parameter spaces (e.g., MCMC methods) but the distribution is complex.
  - Extension to multivariate outputs and combined code-observation analyses.
  - Alternative covariance structures to improve local accuracy.
  - Opening up the black box, dimension reduction, ...