A Bayesian Multilevel Framework for Uncertainty Characterization and the NASA Langley Multidisciplinary UQ Challenge

Conference Paper

Author(s):
Nagel, Joseph Benjamin; Sudret, Bruno

A Bayesian Multilevel Framework for Uncertainty Characterization and the NASA Langley Multidisciplinary UQ Challenge

Joseph B. Nagel[*] and Bruno Sudret[†]

ETH Zürich, Institute of Structural Engineering

Chair of Risk, Safety & Uncertainty Quantification

Stefano-Franscini-Platz 5, CH-8093 Zürich, Switzerland

The NASA Langley multidisciplinary uncertainty quantification challenge has raised contemporary open questions to uncertainty quantification. While originating from a specific aerospace application, an abstract and widely discipline-independent problem formulation prompts researchers and practitioners from various fields in academia and industry to devise generic problem solutions. In this contribution we will address the uncertainty characterization subproblem of the challenge posed. With responses of a given computational model the challenge is to learn about unknown model inputs that are subject to epistemic and aleatory uncertainty. We will approach the problem from a Bayesian perspective to statistical inversion and uncertainty quantification. Within a probabilistic setting Bayesian multilevel modeling allows for an elegant formulation and an efficient solution of complex inverse problems under uncertainty. The mathematical formalism that is necessary to interpret the challenge problem as calibration of a Bayesian multilevel model will be developed. Incidentally we will demonstrate how the problem could be solved in the presence of additional measurement uncertainty and how the entirety of problem unknowns, including those that are not of declared interest, could be identified. Computational key challenges posed by Bayesian inference in this context will be discussed and dedicated algorithms to overcome those will be implemented. Besides the assumptions that our approach rests upon, we will thoroughly discuss the interpretation and fidelity of the final results.

I.   Bayesian Multilevel Modeling

Due to the lack of a universally accepted terminology, we define a *multilevel* or *hierarchical model* as "an assembly of submodels at different levels of a hierarchy". The hierarchical structure can be due to stochastic dependencies and deterministic maps between quantities involved. According to that definition multilevel modeling is sort of an overarching theme in modern multidisciplinary statistics. In the last two decades it has been extensively studied from a frequentist[1,2] and a Bayesian[3,4] point of view. Bayesian multilevel modeling establishes a natural framework for solving complex inverse problems in the presence of aleatory variability and epistemic uncertainty. Prior elicitation[5,6] and posterior computation[7,8] have been discussed in the statistical literature. Data augmentation, which naturally emerges in the context of multilevel modeling, enhances Bayesian computations by introducing unobserved quantities as auxiliary variables.[9,10] Applications of multilevel modeling encompass probabilistic inversion[11] and optimal combination of information.[12,13]

Based on the established theory of Bayesian multilevel modeling we will initially formulate a framework for the solution of the NASA Langley UQ challenge[14] and for the inverse problem posed in the presence of additional measurement uncertainty.[15] While the latter is comparably straightforward within the existing frame of Bayesian multilevel modeling, we will have to establish the slightly more measure-theoretical foundation of the former. Eventually Bayesian calibration of the derived Bayesian multilevel model will be accomplished by appropriately transforming, conditioning and marginalizing probability distributions.

---

[*]PhD Candidate, D-BAUG, nagel@ibk.baug.ethz.ch

[†]Professor, D-BAUG, sudret@ibk.baug.ethz.ch

I.A. Uncertainty & Variability

A *forward model* $\mathcal{M} \colon (\boldsymbol{m}, \boldsymbol{x}, \boldsymbol{\zeta}, \boldsymbol{d}) \mapsto \tilde{\boldsymbol{y}}$ represents the system or phenomenon under consideration. It formally maps model inputs $(\boldsymbol{m}, \boldsymbol{x}, \boldsymbol{\zeta}, \boldsymbol{d}) \in \mathcal{D}_{\boldsymbol{m}} \times \mathcal{D}_{\boldsymbol{x}} \times \mathcal{D}_{\boldsymbol{\zeta}} \times \mathcal{D}_{\boldsymbol{d}}$ to outputs $\tilde{\boldsymbol{y}} = \mathcal{M}(\boldsymbol{m}, \boldsymbol{x}, \boldsymbol{\zeta}, \boldsymbol{d}) \in \mathcal{D}_{\tilde{\boldsymbol{y}}} \subset \mathbb{R}^d$. When carrying out a number of experiments the variability of measured forward model responses can be attributed to models of input uncertainty. There are fixed yet unknown model parameters $\boldsymbol{m} \in \mathcal{D}_{\boldsymbol{m}} \subset \mathbb{R}^p$, model inputs $\boldsymbol{x} \in \mathcal{D}_{\boldsymbol{x}} \subset \mathbb{R}^q$ with imperfectly known aleatory variability, input variables $\boldsymbol{\zeta} \in \mathcal{D}_{\boldsymbol{\zeta}} \subset \mathbb{R}^r$ with perfectly known aleatory variability, and explanatory variables or covariates $\boldsymbol{d} \in \mathcal{D}_{\boldsymbol{d}} \subset \mathbb{R}^s$ that are entirely known.

With respect to a number of $i = 1, \ldots, n$ experiments, forward model inputs are represented as deterministic or stochastic objects within the Bayesian multilevel frame. Throughout the experiments data is acquired under known but possibly different experimental conditions $\boldsymbol{d}_i$. These model inputs $\boldsymbol{d}_i$ are therefore deterministically represented. Fixed albeit unknown model parameters $\boldsymbol{m}$ are assumed to be constant over the experiments. In Bayesian fashion they are represented as random variables $\boldsymbol{M} \sim \pi_{\boldsymbol{M}}(\boldsymbol{m})$, where the Bayesian prior distribution $\pi_{\boldsymbol{M}}(\boldsymbol{m})$ accounts for a subjective degree of belief or prior knowledge about their true values. This is the Bayesian conception of reducible *epistemic uncertainty*. Over the number of experiments varying model inputs $\boldsymbol{\zeta}$ take on unknown experiment-specific realizations $\boldsymbol{\zeta}_i$ of conditionally independent random variables $(\boldsymbol{Z}_i | \boldsymbol{\theta}_{\boldsymbol{Z}}) \sim f_{\boldsymbol{Z} | \boldsymbol{\Theta}_{\boldsymbol{Z}}}(\boldsymbol{\zeta}_i | \boldsymbol{\theta}_{\boldsymbol{Z}})$. The conditional distribution $f_{\boldsymbol{Z} | \boldsymbol{\Theta}_{\boldsymbol{Z}}}(\boldsymbol{\zeta}_i | \boldsymbol{\theta}_{\boldsymbol{Z}})$ with known hyperparameters $\boldsymbol{\theta}_{\boldsymbol{Z}}$ states a subjective degree of belief or prior knowledge about the individual realizations $\boldsymbol{\zeta}_i$. This is the Bayesian notion of irreducible *aleatory variability*. Similarly model inputs $\boldsymbol{x}$ are subject to variability and take on unknown experiment-specific realizations $\boldsymbol{x}_i$ of conditionally independent random variables $(\boldsymbol{X}_i | \boldsymbol{\theta}_{\boldsymbol{X}}) \sim f_{\boldsymbol{X} | \boldsymbol{\Theta}_{\boldsymbol{X}}}(\boldsymbol{x}_i | \boldsymbol{\theta}_{\boldsymbol{X}})$. The hyperparameters $\boldsymbol{\theta}_{\boldsymbol{X}}$ determine this variability throughout the experiments and are fixed but unknown. In turn they are modeled as random variables $\boldsymbol{\Theta}_{\boldsymbol{X}} \sim \pi_{\boldsymbol{\Theta}_{\boldsymbol{X}}}(\boldsymbol{\theta}_{\boldsymbol{X}})$, where the distribution $\pi_{\boldsymbol{\Theta}_{\boldsymbol{X}}}(\boldsymbol{\theta}_{\boldsymbol{X}})$ quantifies an a priori degree of plausibility or evidence. Random variables $(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n) \sim \int \left(\prod_{i=1}^{n} f_{\boldsymbol{X} | \boldsymbol{\Theta}_{\boldsymbol{X}}}(\boldsymbol{x}_i | \boldsymbol{\theta}_{\boldsymbol{X}}) \right) \pi_{\boldsymbol{\Theta}_{\boldsymbol{X}}}(\boldsymbol{\theta}_{\boldsymbol{X}}) \, \mathrm{d}\boldsymbol{\theta}_{\boldsymbol{X}}$ represent the prior knowledge about the experiment-specific realizations $\boldsymbol{x}_i$. This concept is commonly referred to as *exchangeability*.[16, 17]

In short, marginal distributions $\pi_{\boldsymbol{\Theta}_{\boldsymbol{X}}}(\boldsymbol{\theta}_{\boldsymbol{X}})$ and $\pi_{\boldsymbol{M}}(\boldsymbol{m})$ represent *parametric prior knowledge* about the true values of the model parameters $\boldsymbol{m}$ and the hyperparameters $\boldsymbol{\theta}_{\boldsymbol{X}}$, whereas conditional distributions $f_{\boldsymbol{X} | \boldsymbol{\Theta}_{\boldsymbol{X}}}(\boldsymbol{x}_i | \boldsymbol{\theta}_{\boldsymbol{X}})$ and $f_{\boldsymbol{Z} | \boldsymbol{\Theta}_{\boldsymbol{Z}}}(\boldsymbol{\zeta}_i | \boldsymbol{\theta}_{\boldsymbol{Z}})$ encapsulate *structural prior knowledge* about the problem, i.e. information about experiment-specific $\boldsymbol{x}_i$ and $\boldsymbol{\zeta}_i$. Probabilities are generally understood as describing subjective degrees of belief of how the data have come into being.

I.B. Statistical Data Model

As opposed to deterministic solutions, an integral constituent of many statistical approaches to inverse problems is a residual model. Real observations $\boldsymbol{y}_i$ often deviate from model predictions $\tilde{\boldsymbol{y}}_i = \mathcal{M}(\boldsymbol{m}, \boldsymbol{x}_i, \boldsymbol{\zeta}_i, \boldsymbol{d}_i)$ even if forward model inputs were known with certainty. This discrepancy, which is due to measurement errors, numerical approximations and model inadequacies, is often accounted for by a statistical data model $\boldsymbol{y}_i = \tilde{\boldsymbol{y}}_i + \boldsymbol{\varepsilon}_i$. Residual terms $\boldsymbol{\varepsilon}_i$ are assumed to be realizations of random variables $\boldsymbol{E}_i \sim f_{\boldsymbol{E}_i}(\boldsymbol{\varepsilon}_i)$, e.g. with normal distributions $f_{\boldsymbol{E}_i}(\boldsymbol{\varepsilon}_i) = \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_i)$ and experiment-specific, symmetric and positive-semidefinite covariance matrices $\boldsymbol{\Sigma}_i$. It represents a degree of imperfection of the forward model and experimental apparatus. Hence observations are viewed as realizations $\boldsymbol{y}_i$ of random variables $(\boldsymbol{Y}_i | \boldsymbol{m}, \boldsymbol{x}_i, \boldsymbol{\zeta}_i)$ with distributions $f(\boldsymbol{y}_i | \boldsymbol{m}, \boldsymbol{x}_i, \boldsymbol{\zeta}_i) = f_{\boldsymbol{E}_i}(\boldsymbol{y}_i - \mathcal{M}(\boldsymbol{m}, \boldsymbol{x}_i, \boldsymbol{\zeta}_i, \boldsymbol{d}_i))$. The overall model formulated thus far can be summarized as

$$(\boldsymbol{Y}_i | \boldsymbol{m}, \boldsymbol{x}_i, \boldsymbol{\zeta}_i) \sim f_{\boldsymbol{E}_i}\big(\boldsymbol{y}_i - \mathcal{M}(\boldsymbol{m}, \boldsymbol{x}_i, \boldsymbol{\zeta}_i, \boldsymbol{d}_i)\big), \tag{1a}$$

$$\boldsymbol{M} \sim \pi_{\boldsymbol{M}}(\boldsymbol{m}), \tag{1b}$$

$$(\boldsymbol{X}_i | \boldsymbol{\theta}_{\boldsymbol{X}}) \sim f_{\boldsymbol{X} | \boldsymbol{\Theta}_{\boldsymbol{X}}}(\boldsymbol{x}_i | \boldsymbol{\theta}_{\boldsymbol{X}}), \tag{1c}$$

$$\boldsymbol{\Theta}_{\boldsymbol{X}} \sim \pi_{\boldsymbol{\Theta}_{\boldsymbol{X}}}(\boldsymbol{\theta}_{\boldsymbol{X}}), \tag{1d}$$

$$(\boldsymbol{Z}_i | \boldsymbol{\theta}_{\boldsymbol{Z}}) \sim f_{\boldsymbol{Z} | \boldsymbol{\Theta}_{\boldsymbol{Z}}}(\boldsymbol{\zeta}_i | \boldsymbol{\theta}_{\boldsymbol{Z}}). \tag{1e}$$

This model is composed of conditional probabilistic and deterministic relations between the quantities involved. As per our previous definition it is a generic Bayesian multilevel model. An intuitive model representation is provided by a directed acyclic graph (DAG) such as shown in Fig. 1. Unless stated or denoted otherwise random variables in Eq. (1) are assumed to be (conditionally) independent. This defines a joint overall probability density of all probabilistic quantities. By conditioning and marginalizing this overall density at one's convenience, one can derive meaningful probability densities. For inferential purposes these are certain prior and posterior distributions that we will explain in the following.
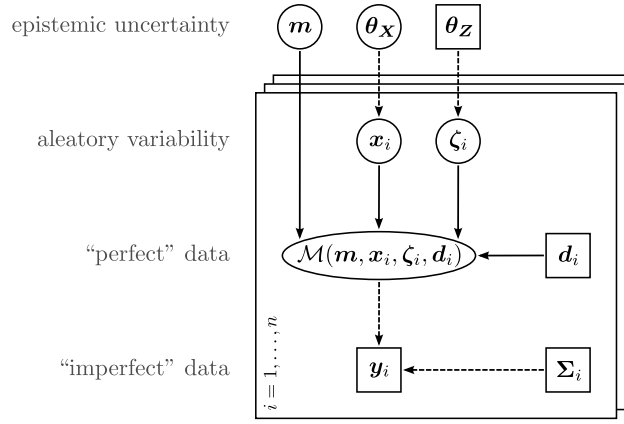
American Institute of Aeronautics and Astronautics

Figure 1. DAG of the generic multilevel model. Vertices symbolize unknown (○) or known (□) quantities and directed edges represent their deterministic (→) or probabilistic (⇢) relations. Model parameters $\boldsymbol{m}$ are constant over $i = 1, \ldots, n$ experiments. The variability of experiment-specific realizations $\boldsymbol{x}_i$ and $\boldsymbol{\zeta}_i$ is determined by unknown or known hyperparameters $\boldsymbol{\theta_X}$ and $\boldsymbol{\theta_Z}$, respectively. Data can be interpreted as "perfect" $\tilde{\boldsymbol{y}}_i = \mathcal{M}(\boldsymbol{m}, \boldsymbol{x}_i, \boldsymbol{\zeta}_i, \boldsymbol{d}_i)$ or "imperfect" observations $\boldsymbol{y}_i = \tilde{\boldsymbol{y}}_i + \boldsymbol{\varepsilon}_i$ with $f_{\boldsymbol{E}_i}(\boldsymbol{\varepsilon}_i) = \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_i)$.

I.C.   Inference in Multilevel Models

In what follows $\langle \boldsymbol{q}_i \rangle$ denotes an ordered n-tuple $\langle \boldsymbol{q}_i \rangle_{1 \leq i \leq n} = (\boldsymbol{q}_1, \boldsymbol{q}_2, \ldots, \boldsymbol{q}_n)$. Conditioned on the priorly known hyperparameters $\boldsymbol{\theta_Z}$, the joint prior distribution of the unknowns $(\boldsymbol{m}, \langle \boldsymbol{x}_i \rangle, \langle \boldsymbol{\zeta}_i \rangle, \boldsymbol{\theta_X})$ is given as

$$\pi\big(\boldsymbol{m}, \langle \boldsymbol{x}_i \rangle, \langle \boldsymbol{\zeta}_i \rangle, \boldsymbol{\theta_X} | \boldsymbol{\theta_Z}\big) = \left(\prod_{i=1}^{n} f_{\boldsymbol{X}|\boldsymbol{\Theta_X}}(\boldsymbol{x}_i | \boldsymbol{\theta_X}) \right) \left(\prod_{i=1}^{n} f_{\boldsymbol{Z}|\boldsymbol{\Theta_Z}}(\boldsymbol{\zeta}_i | \boldsymbol{\theta_Z}) \right) \pi_{\boldsymbol{\Theta_X}}(\boldsymbol{\theta_X}) \, \pi_{\boldsymbol{M}}(\boldsymbol{m}). \tag{2}$$

It summarizes the available parametric and structural prior knowledge. The joint posterior distribution of the unknowns $(\boldsymbol{m}, \langle \boldsymbol{x}_i \rangle, \langle \boldsymbol{\zeta}_i \rangle, \boldsymbol{\theta_X})$ is obtained by further conditioning the prior Eq. (2) on the data $\langle \boldsymbol{y}_i \rangle$. By virtue of Bayes' law this posterior is up to a scale factor found as

$$\pi\big(\boldsymbol{m}, \langle \boldsymbol{x}_i \rangle, \langle \boldsymbol{\zeta}_i \rangle, \boldsymbol{\theta_X} | \langle \boldsymbol{y}_i \rangle, \boldsymbol{\theta_Z}\big) \propto \left(\prod_{i=1}^{n} f_{\boldsymbol{E}_i}\big(\boldsymbol{y}_i - \mathcal{M}(\boldsymbol{m}, \boldsymbol{x}_i, \boldsymbol{\zeta}_i, \boldsymbol{d}_i)\big) \right) \pi\big(\boldsymbol{m}, \langle \boldsymbol{x}_i \rangle, \langle \boldsymbol{\zeta}_i \rangle, \boldsymbol{\theta_X} | \boldsymbol{\theta_Z}\big). \tag{3}$$

The posterior degree of plausibility about the *quantities of interest* (QoI) can be extracted by marginalizing the posterior Eq. (3) over parameters considered *nuisance*. Provided $(\boldsymbol{m}, \boldsymbol{\theta_X})$ are QoI and $(\langle \boldsymbol{x}_i \rangle, \langle \boldsymbol{\zeta}_i \rangle)$ are nuisance parameters, the correspondingly marginalized posterior is

$$\pi\big(\boldsymbol{m}, \boldsymbol{\theta_X} | \langle \boldsymbol{y}_i \rangle, \boldsymbol{\theta_Z}\big) = \int_{\mathcal{D}_{\boldsymbol{x}}^n} \int_{\mathcal{D}_{\boldsymbol{\zeta}}^n} \pi\big(\boldsymbol{m}, \langle \boldsymbol{x}_i \rangle, \langle \boldsymbol{\zeta}_i \rangle, \boldsymbol{\theta_X} | \langle \boldsymbol{y}_i \rangle, \boldsymbol{\theta_Z}\big) \, \mathrm{d}\langle \boldsymbol{x}_i \rangle \, \mathrm{d}\langle \boldsymbol{\zeta}_i \rangle, \tag{4}$$

where $\mathrm{d}\langle \boldsymbol{x}_i \rangle = \mathrm{d}\boldsymbol{x}_1 \ldots \mathrm{d}\boldsymbol{x}_n$ and $\mathrm{d}\langle \boldsymbol{\zeta}_i \rangle = \mathrm{d}\boldsymbol{\zeta}_1 \ldots \mathrm{d}\boldsymbol{\zeta}_n$. Summarized the genuinely unique approach to Bayesian inference in multilevel models is to construct the posterior of the QoI $(\boldsymbol{m}, \boldsymbol{\theta_X})$ by conditioning on the knowns $(\langle \boldsymbol{y}_i \rangle, \boldsymbol{\theta_Z})$ and subsequently marginalizing out nuisance $(\langle \boldsymbol{x}_i \rangle, \langle \boldsymbol{\zeta}_i \rangle)$. Equivalently one could solve an inherently marginal formulation of the Bayesian multilevel calibration problem, with a marginal prior $\pi(\boldsymbol{m}, \boldsymbol{\theta_X}) = \pi_{\boldsymbol{M}}(\boldsymbol{m}) \, \pi_{\boldsymbol{\Theta_X}}(\boldsymbol{\theta_X})$ and a *marginalized* or *integrated likelihood* $\mathcal{L}(\langle \boldsymbol{y}_i \rangle | \boldsymbol{m}, \boldsymbol{\theta_X}, \boldsymbol{\theta_Z})$.[18, 19]

I.D.   "Perfect" Data Model

The Bayesian multilevel model Eq. (1) was based on probabilistic relations of quantities on one problem level conditioned on quantities located at the next "higher" level. Subject to measurement uncertainty, data were interpreted as realizations $\boldsymbol{y}_i$ of random variables $(\boldsymbol{Y}_i | \boldsymbol{m}, \boldsymbol{x}_i, \boldsymbol{\zeta}_i)$ conditioned on direct forward model inputs. We will introduce the term "imperfect" for this statistical data model in order to distinguish it from the following. Not being premised on a residual model, an alternative probability model for the data is to view them as realizations $\tilde{\boldsymbol{y}}_i$ of random variables $(\tilde{\boldsymbol{Y}}_i | \boldsymbol{m}, \boldsymbol{\theta_X}, \boldsymbol{\theta_Z})$ conditioned on the "highest-level" quantities. The proper Bayesian multilevel model involving "perfect" data can be written as

$$(\tilde{\boldsymbol{Y}}_i | \boldsymbol{m}, \boldsymbol{\theta_X}, \boldsymbol{\theta_Z}) \sim f(\tilde{\boldsymbol{y}}_i | \boldsymbol{m}, \boldsymbol{\theta_X}, \boldsymbol{\theta_Z}), \tag{5a}$$

$$(\boldsymbol{M}, \boldsymbol{\Theta_X}) \sim \pi(\boldsymbol{m}, \boldsymbol{\theta_X}) = \pi_{\boldsymbol{M}}(\boldsymbol{m}) \, \pi_{\boldsymbol{\Theta_X}}(\boldsymbol{\theta_X}). \tag{5b}$$

American Institute of Aeronautics and Astronautics

Prior knowledge about the unknowns $(\boldsymbol{m}, \boldsymbol{\theta_X})$ is embodied in Eq. (5b). As a function of the unknowns, provided the density Eq. (5a) exists and is available, one can formulate a residual-free likelihood function

$$\mathcal{L}\big(\langle \tilde{\boldsymbol{y}}_i \rangle | \boldsymbol{m}, \boldsymbol{\theta_X}, \boldsymbol{\theta_Z}\big) = \prod_{i=1}^{n} f(\tilde{\boldsymbol{y}}_i | \boldsymbol{m}, \boldsymbol{\theta_X}, \boldsymbol{\theta_Z}). \tag{6}$$

For reasons that will be discussed below, we call Eq. (6) the *transformed likelihood*. As usual Bayesian data analysis proceeds by conditioning the prior distribution on the acquired data $\langle \tilde{\boldsymbol{y}}_i \rangle$. The posterior follows as

$$\pi\big(\boldsymbol{m}, \boldsymbol{\theta_X} | \langle \tilde{\boldsymbol{y}}_i \rangle, \boldsymbol{\theta_Z}\big) \propto \mathcal{L}\big(\langle \tilde{\boldsymbol{y}}_i \rangle | \boldsymbol{m}, \boldsymbol{\theta_X}, \boldsymbol{\theta_Z}\big) \, \pi(\boldsymbol{m}, \boldsymbol{\theta_X}). \tag{7}$$

Note that the notation of Eq. (7) is reminiscent of classical Bayesian inversion. Indeed the multilevel character of the problem manifests in the likelihood function Eq. (6) that we will now formulate in more detail.

I.E. Push-Forward Measure

Let $(\mathcal{D}_{\boldsymbol{x}, \boldsymbol{\zeta}}, \mathcal{F}_{\boldsymbol{x}, \boldsymbol{\zeta}}, \mathcal{P}_{\boldsymbol{\theta_X}, \boldsymbol{\theta_Z}})$ be the probability space with the sample space $\mathcal{D}_{\boldsymbol{x}, \boldsymbol{\zeta}} = \mathcal{D}_{\boldsymbol{x}} \times \mathcal{D}_{\boldsymbol{\zeta}}$, the Borel σ-field $\mathcal{F}_{\boldsymbol{x}, \boldsymbol{\zeta}} = \mathcal{B}(\mathcal{D}_{\boldsymbol{x}, \boldsymbol{\zeta}})$ and the probability measure $\mathcal{P}_{\boldsymbol{\theta_X}, \boldsymbol{\theta_Z}}$ that is assumed to have independent marginal densities $f_{\boldsymbol{X} | \boldsymbol{\Theta_X}}(\boldsymbol{x} | \boldsymbol{\theta_X})$ and $f_{\boldsymbol{Z} | \boldsymbol{\Theta_Z}}(\boldsymbol{\zeta} | \boldsymbol{\theta_Z})$ for given hyperparameters $(\boldsymbol{\theta_X}, \boldsymbol{\theta_Z})$. For fixed $(\boldsymbol{m}, \boldsymbol{d}_i)$ the mapping $\mathcal{M}_{\boldsymbol{m}, \boldsymbol{d}_i} : (\boldsymbol{x}, \boldsymbol{\zeta}) \mapsto \tilde{\boldsymbol{y}}_i = \mathcal{M}(\boldsymbol{m}, \boldsymbol{x}, \boldsymbol{\zeta}, \boldsymbol{d}_i)$ induces a *push-forward* probability measure $\mathcal{Q}_{\boldsymbol{\theta_X}, \boldsymbol{\theta_Z}}$ on $(\mathcal{D}_{\tilde{\boldsymbol{y}}}, \mathcal{F}_{\tilde{\boldsymbol{y}}}, \mathcal{Q}_{\boldsymbol{\theta_X}, \boldsymbol{\theta_Z}})$ with the Borel σ-algebra $\mathcal{F}_{\tilde{\boldsymbol{y}}} = \mathcal{B}(\mathcal{D}_{\tilde{\boldsymbol{y}}})$. For $\tilde{\mathcal{Y}} \in \mathcal{F}_{\tilde{\boldsymbol{y}}}$ it can be written as

$$\mathcal{Q}_{\boldsymbol{\theta_X}, \boldsymbol{\theta_Z}}(\tilde{\mathcal{Y}}) = \mathcal{P}_{\boldsymbol{\theta_X}, \boldsymbol{\theta_Z}}\big(\mathcal{M}_{\boldsymbol{m}, \boldsymbol{d}_i}^{-1}(\tilde{\mathcal{Y}})\big) = \int_{\mathcal{D}_{\boldsymbol{x}}} \int_{\mathcal{D}_{\boldsymbol{\zeta}}} \mathcal{I}\big(\mathcal{M}_{\boldsymbol{m}, \boldsymbol{d}_i}(\boldsymbol{x}, \boldsymbol{\zeta}) \in \tilde{\mathcal{Y}}\big) f_{\boldsymbol{X} | \boldsymbol{\Theta_X}}(\boldsymbol{x} | \boldsymbol{\theta_X}) f_{\boldsymbol{Z} | \boldsymbol{\Theta_Z}}(\boldsymbol{\zeta} | \boldsymbol{\theta_Z}) \, \mathrm{d}\boldsymbol{x} \, \mathrm{d}\boldsymbol{\zeta}, \tag{8}$$

where $\mathcal{M}_{\boldsymbol{m}, \boldsymbol{d}_i}^{-1}$ is a generalized inverse and $\mathcal{I}$ is the indicator function. The underlying assumption is the measurability of the function $\mathcal{M}_{\boldsymbol{m}, \boldsymbol{d}_i} : (\mathcal{D}_{\boldsymbol{x}, \boldsymbol{\zeta}}, \mathcal{F}_{\boldsymbol{x}, \boldsymbol{\zeta}}) \to (\mathcal{D}_{\tilde{\boldsymbol{y}}}, \mathcal{F}_{\tilde{\boldsymbol{y}}})$, between measurable spaces $(\mathcal{D}_{\boldsymbol{x}, \boldsymbol{\zeta}}, \mathcal{F}_{\boldsymbol{x}, \boldsymbol{\zeta}})$ and $(\mathcal{D}_{\tilde{\boldsymbol{y}}}, \mathcal{F}_{\tilde{\boldsymbol{y}}})$. Assuming the existence of a corresponding density function, for fixed $(\boldsymbol{m}, \boldsymbol{d}_i, \boldsymbol{\theta_X}, \boldsymbol{\theta_Z})$ it reads

$$f(\tilde{\boldsymbol{y}}_i | \boldsymbol{m}, \boldsymbol{\theta_X}, \boldsymbol{\theta_Z}) = \int_{\mathcal{D}_{\boldsymbol{x}}} \int_{\mathcal{D}_{\boldsymbol{\zeta}}} \delta\big(\tilde{\boldsymbol{y}}_i - \mathcal{M}_{\boldsymbol{m}, \boldsymbol{d}_i}(\boldsymbol{x}, \boldsymbol{\zeta})\big) f_{\boldsymbol{X} | \boldsymbol{\Theta_X}}(\boldsymbol{x} | \boldsymbol{\theta_X}) f_{\boldsymbol{Z} | \boldsymbol{\Theta_Z}}(\boldsymbol{\zeta} | \boldsymbol{\theta_Z}) \, \mathrm{d}\boldsymbol{x} \, \mathrm{d}\boldsymbol{\zeta}, \tag{9}$$

where δ denotes the Dirac delta distribution. Basically Eq. (9) is the solution to forward propagation of the aleatory input uncertainties $(\boldsymbol{X} | \boldsymbol{\theta_X}) \sim f_{\boldsymbol{X} | \boldsymbol{\Theta_X}}(\boldsymbol{x} | \boldsymbol{\theta_X})$ and $(\boldsymbol{Z} | \boldsymbol{\theta_Z}) \sim f_{\boldsymbol{Z} | \boldsymbol{\Theta_Z}}(\boldsymbol{\zeta} | \boldsymbol{\theta_Z})$ through the function $\mathcal{M}_{\boldsymbol{m}, \boldsymbol{d}_i}(\boldsymbol{x}, \boldsymbol{\zeta})$ into an output uncertainty $(\tilde{\boldsymbol{Y}}_i | \boldsymbol{m}, \boldsymbol{\theta_X}, \boldsymbol{\theta_Z}) \sim f(\tilde{\boldsymbol{y}}_i | \boldsymbol{m}, \boldsymbol{\theta_X}, \boldsymbol{\theta_Z})$. At this point it is to be noted that epistemic uncertainties are not propagated through the forward model.

I.F. Kernel Density Estimation

Since the transformed likelihood Eq. (6) is rarely available in analytical form, one has to rely on some numerical approximation. A possible approach is to simulate the response density Eq. (9) by Monte Carlo (MC) sampling and kernel density estimation (KDE).[20] In the d-variate case, given a sample $(\tilde{\boldsymbol{y}}^{(1)}, \ldots, \tilde{\boldsymbol{y}}^{(K)})$ from some distribution with an unknown density $f(\tilde{\boldsymbol{y}})$, a kernel smoothing (KS) estimate of this density is given as $\hat{f}(\tilde{\boldsymbol{y}}) = K^{-1} \sum_{k=1}^{K} \mathcal{K}_{\boldsymbol{H}}(\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{y}}^{(k)})$. The scaled kernel $\mathcal{K}_{\boldsymbol{H}}(\tilde{\boldsymbol{y}}) = |\boldsymbol{H}|^{-1/2} \mathcal{K}(\boldsymbol{H}^{-1/2} \tilde{\boldsymbol{y}})$ is defined by a kernel function $\mathcal{K}$ and a symmetric and positive-definite bandwidth matrix $\boldsymbol{H}$. Common types of bandwidth matrices are multiples of the identity matrix $\boldsymbol{H} = h^2 \boldsymbol{1}_d$ for $h > 0$ or diagonal matrices $\boldsymbol{H} = \mathrm{diag}(h_1^2, \ldots, h_d^2)$ with $h_1, \ldots, h_d > 0$. According to certain criteria and assumptions, "optimal" bandwidths are commonly selected to prevent over- and undersmoothing, i.e. a classical bias-variance trade-off. The KDE of a univariate density similar to a Gaussian with kernels of the same type, can be based on Silverman's normal reference rule $h = (4/3K)^{1/5} \hat{\sigma}$, where $\hat{\sigma}$ is the sample standard deviation. Based on MC and KDE techniques the transformed likelihood can be statistically estimated as

$$\hat{\mathcal{L}}_{\mathrm{KS}}\big(\langle \tilde{\boldsymbol{y}}_i \rangle | \boldsymbol{m}, \boldsymbol{\theta_X}, \boldsymbol{\theta_Z}\big) = \prod_{i=1}^{n} \left(\frac{1}{K} \sum_{k=1}^{K} \mathcal{K}_{\boldsymbol{H}}\left(\tilde{\boldsymbol{y}}_i - \tilde{\boldsymbol{y}}_i^{(k)}\right) \right), \quad \text{with} \quad \begin{cases} \boldsymbol{x}^{(k)} \sim f_{\boldsymbol{X} | \boldsymbol{\Theta_X}}(\boldsymbol{x}^{(k)} | \boldsymbol{\theta_X}), \\ \boldsymbol{\zeta}^{(k)} \sim f_{\boldsymbol{Z} | \boldsymbol{\Theta_Z}}(\boldsymbol{\zeta}^{(k)} | \boldsymbol{\theta_Z}), \\ \tilde{\boldsymbol{y}}_i^{(k)} = \mathcal{M}(\boldsymbol{m}, \boldsymbol{x}^{(k)}, \boldsymbol{\zeta}^{(k)}, \boldsymbol{d}_i). \end{cases} \tag{10}$$

For $k = 1, \ldots, K$ forward model responses $\tilde{\boldsymbol{y}}_i^{(k)} = \mathcal{M}(\boldsymbol{m}, \boldsymbol{x}^{(k)}, \boldsymbol{\zeta}^{(k)}, \boldsymbol{d}_i)$ are computed for inputs $\boldsymbol{x}^{(k)}$ and $\boldsymbol{\zeta}^{(k)}$ that are randomly sampled from their parent distributions $f_{\boldsymbol{X} | \boldsymbol{\Theta_X}}(\boldsymbol{x}^{(k)} | \boldsymbol{\theta_X})$ and $f_{\boldsymbol{Z} | \boldsymbol{\Theta_Z}}(\boldsymbol{\zeta}^{(k)} | \boldsymbol{\theta_Z})$, respectively. As it will be discussed, the application of the abovementioned classical criteria to select kernel bandwidths is questionable in the present context.

II. Bayesian Computations

More often than not Bayesian posterior densities do not have analytic closed-form solutions. Nevertheless one can explore posteriors through Markov chain Monte Carlo (MCMC) sampling techniques.[21] The principle of MCMC is to construct a Markov chain over the posterior support whose long-run and steady-state distribution equals the posterior. Prototypical MCMC techniques encompass the Metropolis-Hastings (MH) algorithm[22, 23] and the Gibbs sampler,[24] respectively. Below we will review the MH algorithm and reflect about multilevel-immanent MCMC key issues. Posterior fidelity will be introduced as a major concept.

II.A. The Metropolis-Hastings Algorithm

Let $\pi_0(\boldsymbol{q})$ the prior distribution and $\pi_1(\boldsymbol{q}) \propto \mathcal{L}(\boldsymbol{q})\,\pi_0(\boldsymbol{q})$ the unscaled posterior distribution of an unknown QoI $\boldsymbol{q}$. Initialized at $\boldsymbol{q}^{(0)}$ the MH algorithm generates a Markov chain with equilibrium distribution $\pi_1(\boldsymbol{q})$ by iteratively applying the Markov chain transition kernel as follows. For the current state $\boldsymbol{q}^{(t)}$ a candidate state $\boldsymbol{q}^{(\star)} \sim P(\boldsymbol{q}^{(\star)}|\boldsymbol{q}^{(t)})$ is sampled from a proposal distribution $P(\boldsymbol{q}^{(\star)}|\boldsymbol{q}^{(t)})$. The proposal state becomes accepted, i.e. $\boldsymbol{q}^{(t+1)} = \boldsymbol{q}^{(\star)}$, with probability

$$\alpha\left(\boldsymbol{q}^{(\star)}|\boldsymbol{q}^{(t)}\right) \;=\; \min\left(1, \frac{\pi_1(\boldsymbol{q}^{(\star)})\,P(\boldsymbol{q}^{(t)}|\boldsymbol{q}^{(\star)})}{\pi_1(\boldsymbol{q}^{(t)})\,P(\boldsymbol{q}^{(\star)}|\boldsymbol{q}^{(t)})}\right). \tag{11}$$

Otherwise the proposal will be rejected, i.e. $\boldsymbol{q}^{(t+1)} = \boldsymbol{q}^{(t)}$. Note that the MH correction Eq. (11) requires the computation of posterior ratios, hence only unscaled posterior densities have to be evaluated. Classical random walk Metropolis sampling is based on local proposals, e.g. sampling candidate states from a Gaussian $\boldsymbol{q}^{(\star)} \sim \mathcal{N}(\boldsymbol{q}^{(t)}, \boldsymbol{\Sigma_q})$ with covariance matrix $\boldsymbol{\Sigma_q}$. Independence MH samplers are based on nonlocal proposals, e.g. sampling candidate states from the prior distribution $\boldsymbol{q}^{(\star)} \sim \pi_0(\boldsymbol{q}^{(\star)})$ or from some suitable approximation of the posterior distribution $\boldsymbol{q}^{(\star)} \sim \hat{\pi}_1(\boldsymbol{q}^{(\star)})$.

II.B. Key Challenges

Typically MCMC sampling calls for a high number of forward model runs for likelihood evaluations in Eq. (11). Besides that, the degree as to which MCMC samples will be autocorrelated governs their quality as posterior representatives. The design and efficient tuning of MCMC algorithms therefore aims at optimizing the mixing properties, i.e. the speed of convergence of the Markov chain towards its equilibrium distribution. This is a challenging and highly problem-dependent task that can only be accomplished by employing and combining the most suitable sampling schemes. MCMC methods demand careful convergence diagnostics, i.e. the assessment of when the Markov chain has reached its target distribution and has lost the dependency on its initialization. Moreover MCMC suffers from the curse of dimensionality, a strongly limited degree of parallelism and difficulties in exploring broad and multimodal posteriors. As far as the latter is concerned, because of its mode-jumping capability independence sampling can be considerably more efficient than random walk sampling. As opposed to the "simple" calibration of forward model parameters, multilevel model calibration pose different multilevel-specific MCMC challenges for sampling the posteriors Eqs. (3) and (7). An efficient sampling scheme for high-dimensional parameter spaces is beneficial for the former while the latter demands a sufficiently accurate and numerically efficient evaluation of the transformed likelihood. In total either of the approaches to Bayesian multilevel model calibration may require an enormous number of forward model runs. Generally this number easily exceeds the number of forward model runs necessary for "simple" Bayesian parameter calibration.

II.C. Posterior Fidelity

Due to Bayes' law, closed-form approximations introduced on the level of the likelihood directly induce approximations on the level of the posterior. However, if the posterior is explored by means of MCMC and calls to the likelihood function $\mathcal{L}$ are replaced by calls to a statistical estimator $\hat{\mathcal{L}}$, then a modification is introduced on the level of the Markov chain transition kernel.[25, 26] Consequentially there arises the question of as to what extent the induced equilibrium distribution is in congruence with the true posterior, i.e. the *posterior fidelity*. Moreover there is the question of how free algorithmic parameters, e.g. the number of response samples K and the kernel bandwidth $\boldsymbol{H}$, can be set in order to provide a convenient trade-off between posterior fidelity and computational feasibility, i.e. an "optimal" parameter tuning.

American Institute of Aeronautics and Astronautics

Endeavoring to maintain both fidelity and feasibility, we suppose that it is indeed possible to define certain criteria that parameter tuning can be based on. Even though this is beyond the scope of this paper, we have some preliminary comments. Given the current state $\boldsymbol{q}^{(t)}$ of the Markov chain, in the MH correction Eq. (11) a "random" decision is made whether to approve or to refuse a candidate state $\boldsymbol{q}^{(\star)}$. This decision follows the computation of the posterior ratio $\pi_1(\boldsymbol{q}^{(\star)})/\pi_1(\boldsymbol{q}^{(t)})$. Thus provided that the ratio of estimated likelihoods approximates the true ratio "reasonably well"

$$\frac{\hat{\mathcal{L}}(\boldsymbol{q}^{(\star)})}{\hat{\mathcal{L}}(\boldsymbol{q}^{(t)})} \approx \frac{\mathcal{L}(\boldsymbol{q}^{(\star)})}{\mathcal{L}(\boldsymbol{q}^{(t)})}, \tag{12}$$

an "appropriate" decision is being made. High posterior fidelity is ensured on condition that "appropriate" decisions are being frequently made over the course of the Markov process. The level of confidence that this is indeed the case depends on a complex interplay between the quality of the estimation $\hat{\mathcal{L}}$ of $\mathcal{L}$, the true posterior π_1 and the proposal distribution P.

III. The NASA Langley Multidisciplinary UQ Challenge

Now we will interpret the uncertainty characterization subproblem A of the NASA Langley UQ challenge[14, 15] in Bayesian terms. For that purpose we will compose an appropriate Bayesian multilevel model and formulate the main objective, i.e. the reduction of epistemic uncertainties, as Bayesian calibration of this multilevel model. Inputs $(p_1, p_2, p_3, p_4, p_5)$ of the forward model $\mathcal{M} \equiv h_1$ are subject to uncertainty. Model inputs $\boldsymbol{\zeta} \equiv p_3$ that are subject to aleatory variability constitute the category I parameters. There is epistemic uncertainty about the true value of the category II model parameter $\boldsymbol{m} \equiv p_2$. Category III subsumes those parameters $\boldsymbol{x} \equiv (p_1, p_4, p_5)$ that are subject to a mixed-type uncertainty.

III.A. Category I: Aleatory Uncertainty

For experiments $i = 1, \ldots, n$ category I model inputs $\boldsymbol{\zeta} \equiv p_3 \in [0, 1]$ take on experiment-specific realizations $p_{3,i}$. The population distribution is a uniform distribution $\mathcal{U}(a_3, b_3)$ determined by perfectly known hyperparameters $\boldsymbol{\theta_Z} \equiv \boldsymbol{\theta}_3 = (a_3, b_3)$ with $(a_3, b_3) = (0, 1)$. We will write this as follows

$$(P_{3,i}|\boldsymbol{\theta}_3) \sim f_3(p_{3,i}|\boldsymbol{\theta}_3) = \mathcal{U}(0, 1). \tag{13}$$

It corresponds to a prescribed aleatory variability or structural uncertainty that is irreducible in the sense that by analyzing available data $\tilde{y}_i$ for $i = 1, \ldots, n$ "past" realizations $p_{3,i}$ could be inferred in principle, whereas the knowledge about "future" realizations $p_{3,i'}$ with $i' > n$ cannot be improved.

III.B. Category II: Epistemic Uncertainty

Category II model inputs are physically fixed yet unknown model parameters $\boldsymbol{m} \equiv p_2 \in [0, 1]$. A given epistemic interval $\Delta = [0, 1]$ is known to contain the true value of p_2 prior to any data analysis. We translate this available information into a flat and uniform Bayesian prior probability density

$$P_2 \sim \pi_2(p_2) = \mathcal{U}(0, 1). \tag{14}$$

It represents an a priori degree of plausibility about the true value p_2. It is reducible in the sense that Bayesian updating provides an a posteriori degree of evidence. Note that the quantification of parametric Bayesian priors is always a controversial business, and that priors assign a relative evidence structure over the set of admissible values that goes beyond bare interval-like statements.

III.C. Category III: Mixed Uncertainty

Category III comprises those model inputs $\boldsymbol{x} \equiv (p_1, p_4, p_5)$ that are subject to aleatory variability across experiments $i = 1, \ldots, n$. The natural variability is parametrized by hyperparameters $\boldsymbol{\theta_X} \equiv (\boldsymbol{\theta}_1, \boldsymbol{\theta}_{45})$ that are themselves epistemically uncertain. This is a mixed-type uncertainty model that is sometimes referred to as imprecise probability or distributional p-box. In this context the proper Bayesian interpretation of imprecise probability is exchangeability.[16, 17]

American Institute of Aeronautics and Astronautics

III.C.1. Unimodal Beta

Model inputs $p_1 \in [0,1]$ are distributed according to a unimodal beta distribution. Beta distributions $\text{Beta}(\alpha_1, \beta_1)$ are commonly parametrized by shape hyperparameters $\alpha_1, \beta_1 > 0$. Instead we will here parametrize the beta distribution $\text{Beta}(\mu_1, \sigma_1^2)$ by its mean μ_1 and variance σ_1^2. Thus with unknown hyperparameters $\boldsymbol{\theta}_1 \equiv (\mu_1, \sigma_1^2)$ experiment-specific realizations $p_{1,i}$ are drawn from the population distribution

$$(P_{1,i}|\boldsymbol{\theta}_1) \sim f_1(p_{1,i}|\boldsymbol{\theta}_1) = \text{Beta}(\mu_1, \sigma_1^2). \tag{15}$$

On the one hand, provided the shape parameters (α_1, β_1), the expected value $\mu_1 = \text{E}[p_1]$ and the variance $\sigma_1^2 = \text{Var}[p_1]$ of the density function $\text{Beta}(\mu_1, \sigma_1^2)$ are given as

$$\mu_1 = \frac{\alpha_1}{\alpha_1 + \beta_1}, \quad \sigma_1^2 = \frac{\alpha_1 \beta_1}{(\alpha_1 + \beta_1)^2 (\alpha_1 + \beta_1 + 1)}. \tag{16}$$

On the other hand, given the statistical moments (μ_1, σ_1^2), the shape parameters (α_1, β_1) of the density function $\text{Beta}(\alpha_1, \beta_1)$ can be obtained by

$$\alpha_1 = \left(\frac{\sigma_1^2 + \mu_1^2 - \mu_1}{\sigma_1^2} \right) (-\mu_1), \quad \beta_1 = \left(\frac{\sigma_1^2 + \mu_1^2 - \mu_1}{\sigma_1^2} \right) (\mu_1 - 1). \tag{17}$$

The required unimodality, i.e. the fact that the distribution features a single mode within its support, translates into $\alpha_1, \beta_1 > 1$. Moreover it is required that $3/5 \leq \mu_1 \leq 4/5$ and $1/50 \leq \sigma_1^2 \leq 1/25$. In order to adopt this epistemic uncertainty model about the hyperparameters $\boldsymbol{\theta}_1$ we state the uniform hyperprior distribution

$$\begin{aligned} \boldsymbol{\Theta}_1 &\sim \pi_1(\boldsymbol{\theta}_1) = \mathcal{U}(\mathcal{D}_{\boldsymbol{\theta}_1}), \text{ with} \\ \mathcal{D}_{\boldsymbol{\theta}_1} &= \left\{ (\mu_1, \sigma_1^2) \in \mathbb{R}^2 \,\middle|\, 3/5 \leq \mu_1 \leq 4/5,\, 1/50 \leq \sigma_1^2 \leq 1/25,\, \alpha_1 > 1,\, \beta_1, > 1 \right\}. \end{aligned} \tag{18}$$

If $\lambda(\mathcal{D}_{\boldsymbol{\theta}_1})$ is the Lebesgue measure of the set $\mathcal{D}_{\boldsymbol{\theta}_1} \subset \mathbb{R}^2$, then the uniform density Eq. (18) is $1/\lambda(\mathcal{D}_{\boldsymbol{\theta}_1})$ on $\mathcal{D}_{\boldsymbol{\theta}_1}$ and zero elsewhere. In practice the normalization constant $\lambda(\mathcal{D}_{\boldsymbol{\theta}_1})$ is unknown, but since priors are flat and only ratios are compared in the MH correction Eq. (11), only the set membership of MCMC proposals has to be determined. Consequently we can treat the prior Eq. (18) as $\pi_1(\boldsymbol{\theta}_1) = \pi(\mu_1)\,\pi(\sigma_1^2)$ with independent marginals $\pi(\mu_1) = \mathcal{U}(3/5, 4/5)$ and $\pi(\sigma_1^2) = \mathcal{U}(1/50, 1/25)$ and reject proposals that do not respect $\alpha_1, \beta_1 > 1$ with the aid of Eq. (17). This practical prior choice is ambiguous in the sense that priors could be assumed for shape parameters (α_1, β_1), too. However, this could yield improper prior distributions. Moreover we consider (μ_1, σ_1^2) statistically more "natural" than the shape parameters. In addition they underlie strong prior constraints which is advantageous to exploring the posterior by means of MCMC.

III.C.2. Correlated Gaussian

The model inputs $p_4, p_5 \in \mathbb{R}$ will be modeled as possibly correlated Gaussian random variables. Across the experiments $i = 1, \ldots, n$ these model inputs take on different unknown realizations $(p_{4,i}, p_{5,i})$. This inherently aleatory variability is represented by the population distribution

$$\left((P_{4,i}, P_{5,i})|\boldsymbol{\theta}_{45} \right) \sim f_{45}\left((p_{4,i}, p_{5,i})|\boldsymbol{\theta}_{45} \right) = \mathcal{N}\left(\boldsymbol{\mu}_{45}, \boldsymbol{\Sigma}_{45} \right). \tag{19}$$

For $j = 4, 5$ the means $\mu_j = \text{E}[p_j]$, variances $\sigma_j^2 = \text{Var}[p_j]$ and the coefficient of correlation $\rho_{45} = \text{E}[(p_4 - \mu_4)(p_5 - \mu_5)]$ constitute the hyperparameters $\boldsymbol{\theta}_{45} \equiv (\mu_4, \sigma_4^2, \mu_5, \sigma_5^2, \rho_{45})$. Those hyperparameters are unknown constants that determine the mean $\boldsymbol{\mu}_{45}$ and the covariance matrix $\boldsymbol{\Sigma}_{45}$ of the bivariate normal density by

$$\boldsymbol{\mu}_{45} = \begin{pmatrix} \mu_4 \\ \mu_5 \end{pmatrix}, \quad \boldsymbol{\Sigma}_{45} = \begin{pmatrix} \sigma_4^2 & \rho_{45}\, \sigma_4\, \sigma_5 \\ \rho_{45}\, \sigma_4\, \sigma_5 & \sigma_5^2 \end{pmatrix}. \tag{20}$$

Besides the natural bounds $|\rho_{45}| \leq 1$ it is requested that $-5 \leq \mu_j \leq 5$ and $1/400 \leq \sigma_j^2 \leq 4$. We translate these intervals into flat and independent marginals $\pi(\mu_j)$, $\pi(\sigma_j^2)$ and $\pi(\rho_{45})$ of the common hyperprior $\pi_{45}(\boldsymbol{\theta}_{45})$ by

$$\left. \begin{aligned} \pi(\mu_j) &= \mathcal{U}(-5, 5), \\ \pi(\sigma_j^2) &= \mathcal{U}(1/400, 4), \\ \pi(\rho_{45}) &= \mathcal{U}(-1, 1), \end{aligned} \right\} \quad \boldsymbol{\Theta}_{45} \sim \pi_{45}(\boldsymbol{\theta}_{45}) = \left(\prod_{j=4}^{5} \pi(\mu_j)\,\pi(\sigma_j^2) \right) \pi(\rho_{45}). \tag{21}$$

The ambiguity in quantifying parametric Bayesian priors is especially obvious for spread hyperparameters, insofar as priors for spread hyperparameters could refer to standard deviations or variances alike.

III.D. Bayesian Problem Statement

The primary objective of the UQ challenge subproblem A is the reduction of epistemic uncertainties about the true values of the forward model parameter p_2 and the hyperparameters $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_{45})$.[14] In order to accomplish that goal, the forward model, data and prior knowledge is available. Preventing to reverse engineer its mathematical character and numerical implementation, the forward model h_1 is distributed as a protected MATLAB p-code file, i.e. a blackbox model. Available data $\langle \tilde{y}_i \rangle_{1 \le i \le 50}$ comprises $n = 50$ scalar observations $\tilde{y}_i = h_1(p_{1,i}, p_2, p_{3,i}, p_{4,i}, p_{5,i})$ which have been realized as forward model responses complying with the true uncertainty model of forward model inputs, i.e. the model parameter p_2 takes on its true value and $(\langle p_{1,i} \rangle, \langle p_{3,i} \rangle, \langle p_{4,i} \rangle, \langle p_{5,i} \rangle)$ have been randomly sampled from their true population distributions. Notwithstanding that the observations provided are "perfect", in general they might very well be subject to an additional model-measurement discrepancy,[15] i.e. "imperfect. Data have been arranged into two distinct configurations of observations $\langle \tilde{y}_i \rangle_{1 \le i \le 25}$ and $\langle \tilde{y}_i \rangle_{26 \le i \le 50}$ whose separate and joint analysis is envisaged to indicate how the number n of processed samples impacts the significance of the final results. The available prior knowledge has been translated into parametric and structural Bayesian prior distributions. We have pointed out that this formulation endows the problem with a subjectivist interpretation of probability and suffers from the ambiguity in the chosen parametric prior and its influence on the resulting posterior.

The problem statement as well as the framework and the algorithms introduced so far grant ample scope of formulating and solving the problem as Bayesian inference of the QoI $(p_2, \boldsymbol{\theta}_1, \boldsymbol{\theta}_{45})$ within a multilevel context. In the first place the Bayesian multilevel model Eq. (5), defined by parametric priors Eqs. (14), (18) and (21) and structural priors Eqs. (13), (15) and (19), establishes the natural framework for solving the original problem posed. For the sake of completeness the devised multilevel model is summarized as

$$(\tilde{Y}_i | p_2, \boldsymbol{\theta}_1, \boldsymbol{\theta}_{45}, \boldsymbol{\theta}_3) \sim f(\tilde{y}_i | p_2, \boldsymbol{\theta}_1, \boldsymbol{\theta}_{45}, \boldsymbol{\theta}_3), \tag{22a}$$

$$P_2 \sim \pi_2(p_2) = \mathcal{U}(0, 1), \tag{22b}$$

$$(P_{1,i} | \boldsymbol{\theta}_1) \sim f_1(p_{1,i} | \boldsymbol{\theta}_1) = \mathrm{Beta}(\mu_1, \sigma_1^2), \tag{22c}$$

$$((P_{4,i}, P_{5,i}) | \boldsymbol{\theta}_{45}) \sim f_{45}((p_{4,i}, p_{5,i}) | \boldsymbol{\theta}_{45}) = \mathcal{N}(\boldsymbol{\mu}_{45}, \boldsymbol{\Sigma}_{45}), \tag{22d}$$

$$\boldsymbol{\Theta}_1 \sim \pi_1(\boldsymbol{\theta}_1) = \mathcal{U}(\mathcal{D}_{\boldsymbol{\theta}_1}), \tag{22e}$$

$$\boldsymbol{\Theta}_{45} \sim \pi_{45}(\boldsymbol{\theta}_{45}) = \pi(\mu_4) \, \pi(\sigma_4^2) \, \pi(\mu_5) \, \pi(\sigma_5^2) \, \pi(\rho_{45}), \tag{22f}$$

$$(P_{3,i} | \boldsymbol{\theta}_3) \sim f_3(p_{3,i} | \boldsymbol{\theta}_3) = \mathcal{U}(0, 1). \tag{22g}$$

The posterior Eq. (7) of the QoI follows Bayesian data analysis of given forward model responses $\tilde{y}_i$, i.e. realizations of random variables $(\tilde{Y}_i | p_2, \boldsymbol{\theta}_1, \boldsymbol{\theta}_{45}, \boldsymbol{\theta}_3)$. In the second place one could solve the inverse problem posed in the presence of additional measurement noise. To that end the Bayesian multilevel model Eq. (1) establishes the proper framework. Synthetic and noisy observations $y_i = \tilde{y}_i + \varepsilon_i$ could be obtained by perturbing the given model responses $\tilde{y}_i$ with residuals ε_i that are randomly sampled from prescribed distributions $f_{E_i}(\varepsilon_i)$. Parameters of the residual model, e.g. the residual variances σ_i^2, could either be treated as knowns or as further unknowns. By analyzing "imperfect" data y_i, i.e. realizations of random variables $(Y_i | p_{1,i}, p_2, p_{3,i}, p_{4,i}, p_{5,i})$, and treating latent variables $(\langle p_{1,i} \rangle, \langle p_{3,i} \rangle, \langle p_{4,i} \rangle, \langle p_{5,i} \rangle)$ as nuisance, inference of the QoI would be based on the posterior Eq. (4). A DAG of the Bayesian multilevel model corresponding to our challenge problem interpretation with "perfect" and "imperfect" data, respectively, is depicted in Fig. 2.

IV. Bayesian Data Analysis

We will now apply the inferential machinery of Bayesian multilevel calibration for solving the Bayesian incarnation of the uncertainty characterization subproblem A of the NASA Langley multidisciplinary UQ challenge. The problem will be solved in its original formulation involving "perfect" data. Motivated by findings from first preliminary problem analyses, posterior densities of the QoI will be sampled by a suitable independence MCMC algorithm. This sampler will be implemented in MATLAB and serially run on a modern CPU. Nevertheless we will discuss possible parallelization strategies. The total data $\langle \tilde{y}_i \rangle_{1 \le i \le 50}$ and its subconfigurations $\langle \tilde{y}_i \rangle_{1 \le i \le 25}$ and $\langle \tilde{y}_i \rangle_{26 \le i \le 50}$ will be analyzed with the devised algorithm. Based on heuristic parameter tuning and plausibility checks we will try to assess the fidelity of the posterior. Promising a boost of posterior fidelity we will lastly devise a hybrid MCMC scheme which is based on data augmentation and both independence and random walk sampling.
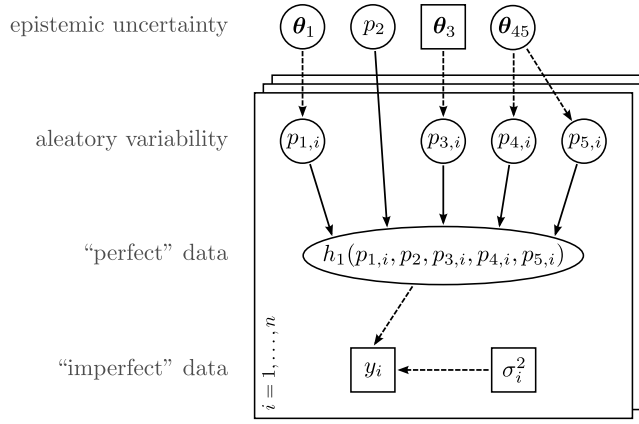
Figure 2. DAG of the NASA UQ challenge subproblem A. The hyperparameters $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_{45}$ and the forward model parameter p_2 located at the "highest" hierarchical level are the QoI. Realizations ($\langle p_{1,i} \rangle, \langle p_{3,i} \rangle, \langle p_{4,i} \rangle, \langle p_{5,i} \rangle$) on the "intermediate" problem level are considered nuisance. "Perfect" $\tilde{y}_i = h_1(p_{1,i}, p_2, p_{3,i}, p_{4,i}, p_{5,i})$ or "imperfect" data $y_i = \tilde{y}_i + \varepsilon_i$ constitute the "lowest" model layer.

IV.A. Preliminary Analyses

A basic understanding of an inverse problem under consideration allows to judge the performance of various types of MCMC schemes. Since this allows to design efficient algorithms and prevents from obtaining misleading results that are due to inapplicable samplers, gaining first insights into the multilevel calibration problem is indispensable. Initial MCMC runs had therefore been based on crude random walk Metropolis sampling. The principal nature of the posteriors Eqs. (3) and (7) could be provisionally assessed. Main findings from sampling the posterior Eq. (7) indicate that posterior marginals of the QoI ($p_2, \boldsymbol{\theta}_1, \boldsymbol{\theta}_{45}$) can be multimodal and broad distributions that significantly overlap with the marginal priors. Sampling the joint posterior Eq. (3) of the entirety of unknowns ($\langle p_{1,i} \rangle, p_2, \langle p_{3,i} \rangle, \langle p_{4,i} \rangle, \langle p_{5,i} \rangle, \boldsymbol{\theta}_1, \boldsymbol{\theta}_{45}$) has provided additional insight. Experiment-specific unknowns $\langle p_{1,i} \rangle$ occur to be identifiable, i.e. the corresponding posterior features a single mode. Those are valuable information that will eventually motivate the final MCMC samplers.

IV.B. "Perfect" Data Analysis

For the calibration of the Bayesian multilevel model Eq. (5) we devise a blockwise independence MCMC sampler. Since the algorithm is based on MCMC, MC and KS techniques, hereinafter it will be referred to as MC³KS. QoI are grouped in blocks (p_2), (μ_1, σ_1^2), ($\mu_4, \sigma_4^2, \rho_{45}$) and ($\mu_5, \sigma_5^2$) that are consecutively updated by sampling blockwise candidates from the corresponding prior distributions. In many cases independence sampling from the priors is inefficient due to a negligible overlap between the priors and the posterior distributions and the resulting low acceptance rates. However, on account of the multimodality of the posteriors and their overlap with the the priors, that were indicated by first analyses, independence sampling promises rapid mixing for the problem at hand. Moreover in the context of Eq. (12) we suppose that wide jumps in the parameter space, that are induced by independence sampling on average, are beneficial in terms of posterior fidelity. Another advantage of the devised MCMC scheme over random walk sampling is that it does not require extensive fine-tuning of the proposal distribution. Updating in blocks intents to minimize the number of calls to the likelihood Eq. (10) that are necessary for each block in each MCMC iteration, while maintaining high acceptance rates. With the help of Eq. (17) the constraints $\alpha_1, \beta_1 > 1$ are enforced by rejecting nonconforming proposals. The MC³KS sampler is initialized by setting parameters in the middle of their admissible intervals. Generally speaking we expect that forward model parameters and mean hyperparameters are easier to identify than spread or correlation hyperparameters.

IV.B.1. Likelihood Estimation & Posterior Fidelity

For the estimation Eq. (10) of the transformed likelihood Eq. (6) we choose kernel functions $\mathcal{K}$ of Gaussian type. In order to achieve a convenient trade-off between the conflicting endeavors posterior fidelity and ease of its computation, the number of samples K and the bandwidth h have to be set. In practice resource limitations restrict the total number of affordable forward model runs, hence we approach parameter tuning from the situation of given K. Owing to the absence of a rigorous means to define a corresponding and

American Institute of Aeronautics and Astronautics

"optimal" bandwidth h, we study the posteriors obtained for fixed $K = 10^4$ and decreasing h in a cascade of runs. We observe an initial shrinkage of the posterior, i.e. evolving from the flat prior it takes on definite shape, and an eventual collapse, i.e. the posterior flattens out again and looses its structure. The initial shrinkage is associated with significant changes of the posterior shape, the eventual breakdown is QoI-dependent, and in between the posterior is relatively stable with respect to h. We remark that this behavior is consistent with Eqs. (11) and (12). Significant oversmoothing the target density Eq. (9), i.e. a strongly biased estimator Eq. (10), can falsely assign posterior mass to QoI-values that do not well-explain or even contradict the data. Considerable undersmoothing of the target density, i.e. a high variance of the estimator Eq. (10), can cause "arbitrary" acceptances in the MH correction. We speculate that in between those extremes, the more stable the posterior is with respect to small changes in h, the more confident we can be to have revealed the true posterior. Beyond that we presume that a high degree of distinctiveness of the posterior with respect to the prior indicates high posterior fidelity, even though the converse statement does not hold. In addition to those heuristics we perform plausibility checks based on monitoring the accuracy of evaluating likelihood ratios Eq. (12) over the Markov chain.

Following this discussion $K = 10^4$ and $h = 0.002$ constitutes our final parameter setup. The principle of estimating the density Eq. (9) and the transformed likelihood Eq. (6) is visualized in Fig. 3. Samples of $K = 10^4$ and $K = 10^7$ forward model responses are simulated for two different (hyper)parameter values $(p_2, \boldsymbol{\theta}_1, \boldsymbol{\theta}_{45})_{\text{high}}$ and $(p_2, \boldsymbol{\theta}_1, \boldsymbol{\theta}_{45})_{\text{low}}$. As judged from our final results, these are (hyper)parameter values of high and low degree of posterior evidence, respectively. For the smaller sample with $K = 10^4$ estimates of the sought densities $f(\tilde{y}_i | p_2, \boldsymbol{\theta}_1, \boldsymbol{\theta}_{45}, \boldsymbol{\theta}_3)$ are shown. For reference purposes a histogram of the larger sample with $K = 10^7$ is shown. It can be seen that response densities $f(\tilde{y}_i | p_2, \boldsymbol{\theta}_1, \boldsymbol{\theta}_{45}, \boldsymbol{\theta}_3)$ for $(p_2, \boldsymbol{\theta}_1, \boldsymbol{\theta}_{45})_{\text{high}}$ and $(p_2, \boldsymbol{\theta}_1, \boldsymbol{\theta}_{45})_{\text{low}}$ significantly overlap. This is a problem characteristic that complicates the identification of the QoI $(p_2, \boldsymbol{\theta}_1, \boldsymbol{\theta}_{45})$. It can also be seen that the employed bandwidth $h = 0.002$ amounts to a slight undersmoothing of the target density, i.e. a bias-variance trade-off favoring lower bias yet acceptable variance. This is advantageous because it allows to capture local small-scale features of the target density, e.g. sharp peaks and edges, in the posterior. Since the target density significantly differs from a normal distribution, automatic bandwidth selection cannot be based on the normal reference rule. The resulting oversmoothing of the target density, i.e. a significantly biased KDE, would veil its important characteristics. Finally the (hyper)parameter values $(p_2, \boldsymbol{\theta}_1, \boldsymbol{\theta}_{45})_{\text{high}}$ can be seen to lead to a response density that explains the data sample $\langle \tilde{y}_i \rangle$ reasonably well.
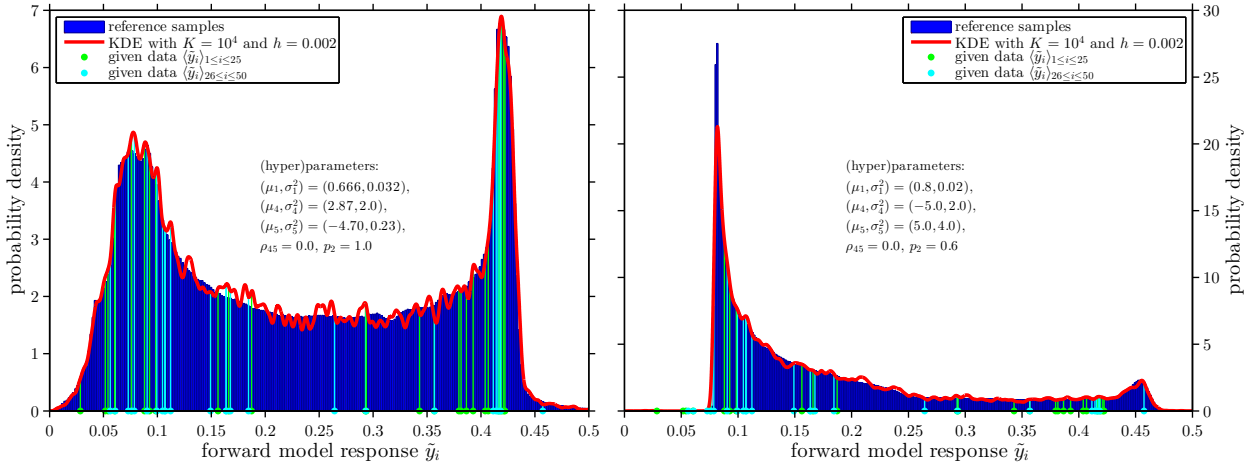

Figure 3. **Estimation of $f(\tilde{y}_i | p_2, \boldsymbol{\theta}_1, \boldsymbol{\theta}_{45}, \boldsymbol{\theta}_3)$.** Evaluating the transformed likelihood Eq. (6) for MC^{3}KS is based on the forward model response density Eq. (9). For two different values of the (hyper)parameters $(p_2, \boldsymbol{\theta}_1, \boldsymbol{\theta}_{45})$ a KDE of $f(\tilde{y}_i | p_2, \boldsymbol{\theta}_1, \boldsymbol{\theta}_{45}, \boldsymbol{\theta}_3)$ with $K = 10^4$ and $h = 0.002$ is shown. As a reference a histogram is shown for a larger number $K = 10^7$ of forward model responses.

IV.B.2. Final Results

First of all we analyze the total data $\langle \tilde{y}_i \rangle_{1 \leq i \leq 50}$. For $N = 10^5$ iterations of the MC^{3}KS algorithm the total program runtime amounts to $t \approx 30\,\text{h}$ on a single core. Blockwise acceptance rates were found to be ca. $20\,\%$ for (p_2), $40\,\%$ for (μ_1, σ_1^2), $60\,\%$ for $(\mu_4, \sigma_4^2, \rho_{45})$ and $10\,\%$ for (μ_5, σ_5^2). With Eq. (16) a number of 10327 blockwise proposals (μ_1, σ_1^2) had been rejected because of violating the prior requirement $\alpha_1, \beta_1 > 1$.

Posterior marginals of the QoI are shown in Figs. 4 to 7. The densities shown have been obtained by kernel smoothing of the MCMC posterior samples based on appropriate boundary correction methods. We assign an acceptable degree of fidelity to the posteriors obtained, i.e. we are confident to have revealed the true posteriors, regardless of whether some are flat and only weakly informative.

Note that Bayesian probability densities feature a richer structure than mere epistemic intervals. They are read as a relative degree of evidence and may contain complex dependency structures. Parameters that were assumed to be statistically independent a priori, can be statistically dependent a posteriori. The marginal densities shown hide this possibly existing posterior correlations. Small negative correlations between μ_1 and σ_1^2 or μ_4 with linear Pearson's coefficients of correlation $r_{\mu_1, \sigma_1^2} = -0.08$ and $r_{\mu_1, \mu_4} = -0.22$ were discovered. In order to provide final results of interval-like character one could define suitable Bayesian credible intervals or sets that accumulate a certain proportion, e.g. 95 %, of the total posterior mass. However, the definition of such intervals is ambiguous and would still bear the probabilistic interpretation, therefore we refrain from defining Bayesian credible intervals.

We also analyze the data subconfigurations $\langle \tilde{y}_i \rangle_{1 \leq i \leq 25}$ and $\langle \tilde{y}_i \rangle_{26 \leq i \leq 50}$ separately. The posterior densities produced by separate analyses may differ considerably. With respect to the posteriors yielded by analyzing $\langle \tilde{y}_i \rangle_{1 \leq i \leq 50}$, the two data subconfigurations are representative to a different degree. Those findings indicate that $n = 25$ is a comparably low number of observations while $n = 50$ is moderately satisfying for the Bayesian calibration of mean hyperparameters and the forward model parameter. Properly identifying the variance and correlation hyperparameters would require a higher number of observations. This is hardly surprising regarding the complex uncertainty setup, the number of unknowns, the unknown character of the forward model, and the inverse nature of the calibration problem.

IV.B.3. Conclusion

With the proposed MC^{3}KS sampling scheme we have solved the Bayesian formulation of the challenge problem. The shortcoming of the approach was the unclear assessment of the fidelity of the posterior, i.e. its dependency on algorithmic tuning parameters. We had to base parameter tuning on heuristic criteria and plausibility checks. The fidelity of the posteriors could only be provisionally assessed. The efficiency of the proposed independence sampling MCMC scheme can be easily increased. Obtained posterior distributions may be approximated by suitable distributions that are easy to sample. Utilizing these posterior approximations as proposal distributions will lead to higher acceptance rates and better mixing properties than sampling from the priors. Most Bayesian computations can only be parallelized by running several Markov chains simultaneously. Another parallelization strategy of the devised algorithm is to parallelize the estimation of the transformed likelihood on the level of single forward model runs. This suggests the possibility of studying the posterior for significantly larger K and smaller h. Moreover different classes of kernel functions $\mathcal{K}$, e.g. with bounded nonzero support, or more advanced KDE techniques, e.g. locally adaptive schemes, can be employed.
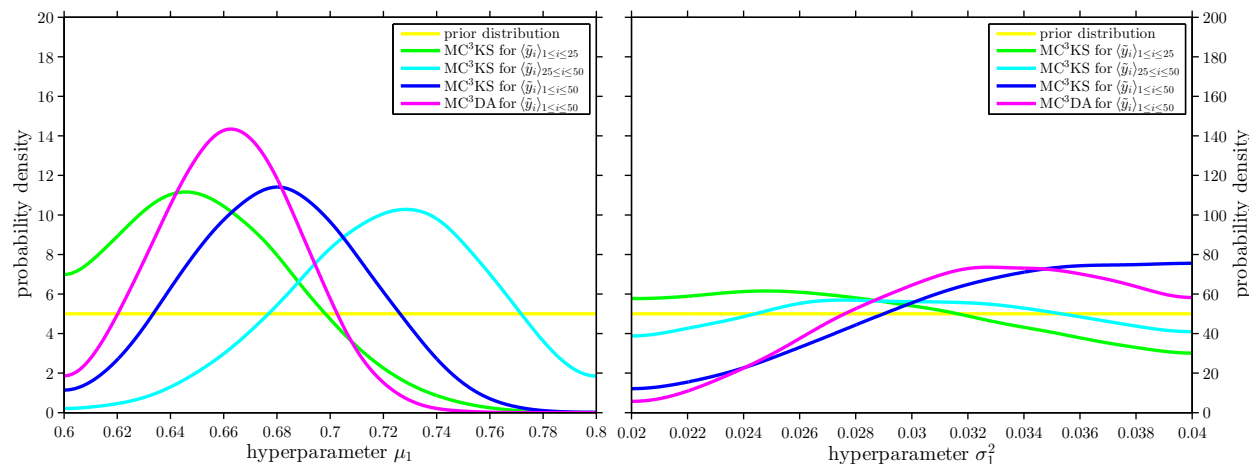

Figure 4. Posteriors of μ_1 and σ_1^2. The posterior of μ_1 features a clear structure as compared to the prior. While separate analyses of $\langle \tilde{y}_i \rangle_{1 \leq i \leq 25}$ and $\langle \tilde{y}_i \rangle_{26 \leq i \leq 50}$ lead to different posterior modes, the joint analysis of $\langle \tilde{y}_i \rangle_{1 \leq i \leq 50}$ leads to a mode in between the aforementioned ones. The posterior of σ_1^2 is seen to be comparably structureless and therefore less informative.
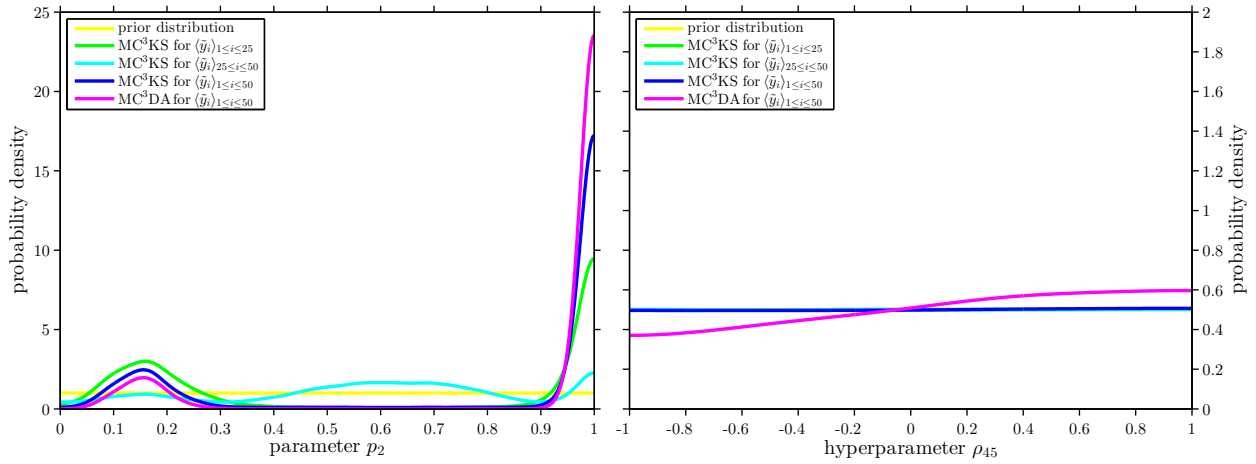
American Institute of Aeronautics and Astronautics

Figure 5. Posteriors of p_2 and ρ_{45}. Due to the fact that data subconfigurations $\langle \tilde{y}_i \rangle_{1 \leq i \leq 25}$ and $\langle \tilde{y}_i \rangle_{26 \leq i \leq 50}$ are to a different degree informative about hyperparameters, the posteriors obtained for the constant model parameter p_2 deviate as well. Analyzing $\langle \tilde{y}_i \rangle_{1 \leq i \leq 50}$ reveals two clear and separated posterior modes. As expected the posterior of the correlation hyperparameter ρ_{45} is flat.
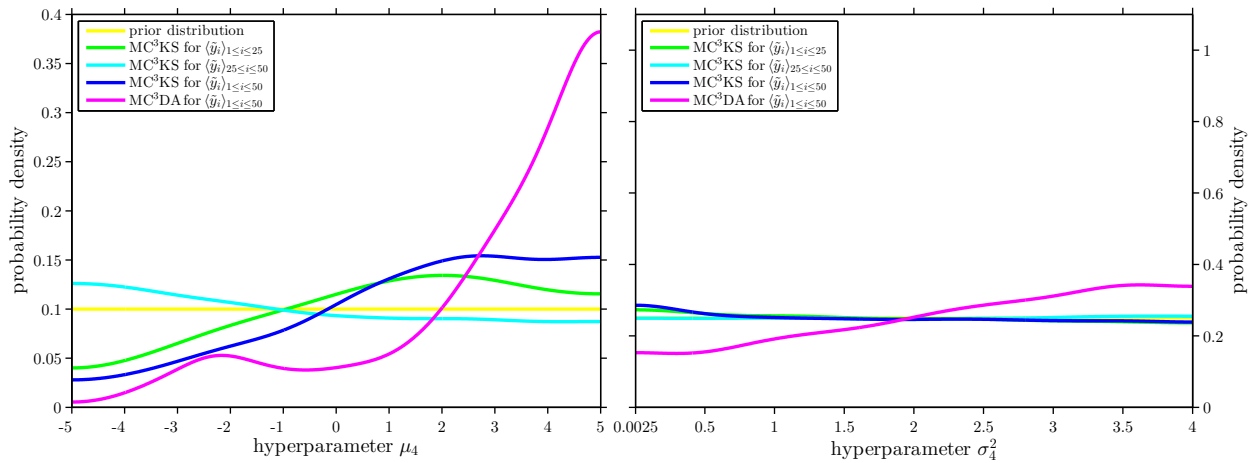
Figure 6. Posteriors of μ_4 and σ_4^2. Both the posteriors of μ_4 and σ_4^2 that were sampled by MC^{3}KS are rather flat and uninformative. On the contrary, the posterior of μ_4 explored by MC^{3}DA features more definite structure. This questions the fidelity of the posterior.
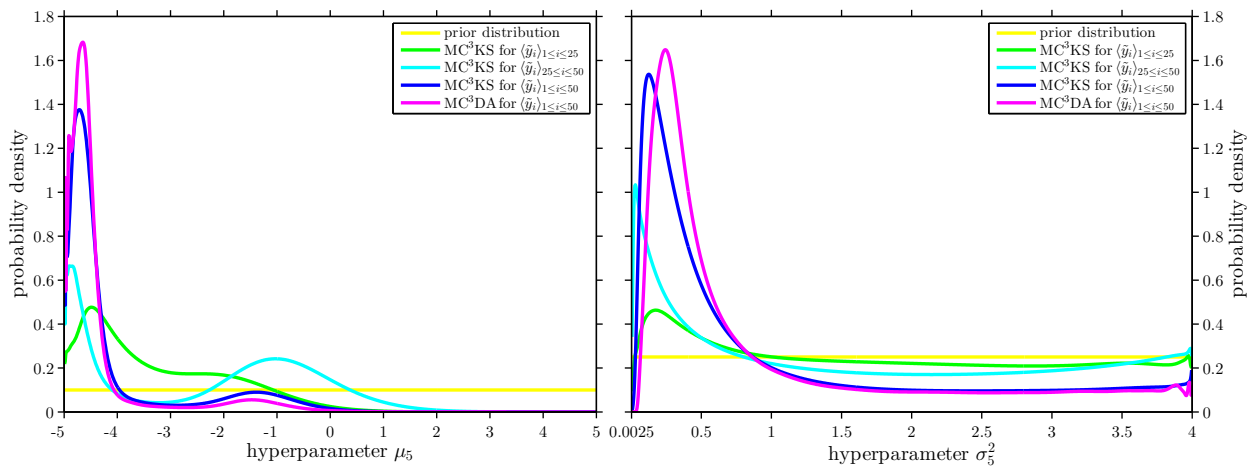

Figure 7. Posteriors of μ_5 and σ_5^2. The posteriors marginals of μ_5 and σ_5^2 feature a distinctive structure as compared to the priors. The posterior of μ_5 is multimodal whereas the one of σ_5^2 is unimodal. With respect to the posteriors sampled by MC^{3}KS, the posteriors that are due to MC^{3}DA are only slightly more evolved in structure. This indicates an acceptable degree of posterior fidelity.

American Institute of Aeronautics and Astronautics

IV.C. Partial Data Augmentation

As a potential improvement over the employed MC^3KS sampler we will devise a new hybrid MCMC sampling scheme. Since the scheme will be based on data augmentation (DA), henceforth it will be referred to as MC^3DA. Traditionally DA can be a powerful tool for enhancing the computational efficiency of MCMC posterior sampling.[9, 10] Instead we will herein utilize DA as a means to reformulate the multilevel model calibration problem in a complementary way, such that it allows for more adequate likelihood estimations which in turn promises enhanced posterior fidelity. Rather than directly sampling the posterior of the QoI $(p_2, \boldsymbol{\theta}_1, \boldsymbol{\theta}_{45})$, one can sample the posterior of an augmented number of unknowns $(\langle p_{1,i}\rangle, p_2, \boldsymbol{\theta}_1, \boldsymbol{\theta}_{45})$ and obtain the posterior of the QoI by subsequently marginalizing over nuisance $\langle p_{1,i}\rangle$. Presuming that sampling from $\pi(\langle p_{1,i}\rangle, p_2, \boldsymbol{\theta}_1, \boldsymbol{\theta}_{45}|\langle \tilde{y}_i\rangle, \boldsymbol{\theta}_3)$ is "easier" to accomplish than straightforwardly sampling from $\pi(p_2, \boldsymbol{\theta}_1, \boldsymbol{\theta}_{45}|\langle \tilde{y}_i\rangle, \boldsymbol{\theta}_3)$, a de facto improvement is achieved. The introduction of $\langle p_{1,i}\rangle$ as auxiliary variables is a partial form of data augmentation. As indicated by preliminary problem analyses, the forward model h_1 seems to be in such a way dependent on its input p_1, that data $\langle \tilde{y}_i\rangle$ can be inverted for the unknown $\langle p_{1,i}\rangle$, under uncertainty of $(p_2, \langle p_{3,i}\rangle, \langle p_{4,i}\rangle, \langle p_{5,i}\rangle)$. Moreover the likelihood corresponding to partial data augmentation can be estimated more adequately and the approach will allow for automatic kernel bandwidth selection based on a classical yet well-approved criterion, namely the normal reference rule. Indeed the aforementioned facts will allow to sample $\pi(\langle p_{1,i}\rangle, p_2, \boldsymbol{\theta}_1, \boldsymbol{\theta}_{45}|\langle \tilde{y}_i\rangle, \boldsymbol{\theta}_3)$ with higher fidelity than sampling $\pi(p_2, \boldsymbol{\theta}_1, \boldsymbol{\theta}_{45}|\langle \tilde{y}_i\rangle, \boldsymbol{\theta}_3)$. We will introduce the formalism of partial data augmentation below.

IV.C.1. Augmented Multilevel Model

If the unknowns $(p_2, \boldsymbol{\theta}_1, \boldsymbol{\theta}_{45})$ of the multilevel model Eq. (5) are augmented by $\langle p_{1,i}\rangle$, then the collective of unknowns $(\langle p_{1,i}\rangle, p_2, \boldsymbol{\theta}_1, \boldsymbol{\theta}_{45})$ has to be inferred explicitly. The associated Bayesian prior density is given as

$$\pi\big(\langle p_{1,i}\rangle, p_2, \boldsymbol{\theta}_1, \boldsymbol{\theta}_{45}\big) = \left(\prod_{i=1}^{n} f_1(p_{1,i}|\boldsymbol{\theta}_1)\right)\pi_1(\boldsymbol{\theta}_1)\,\pi_2(p_2)\,\pi_{45}(\boldsymbol{\theta}_{45}). \tag{23}$$

It comprises both parametric and structural prior knowledge. Given the density $f(\tilde{y}_i|p_{1,i}, p_2, \boldsymbol{\theta}_{45}, \boldsymbol{\theta}_3)$ of the distribution of random variables $(\tilde{Y}_i|p_{1,i}, p_2, \boldsymbol{\theta}_{45}, \boldsymbol{\theta}_3)$, the corresponding *augmented likelihood* follows as

$$\mathcal{L}\big(\langle \tilde{y}_i\rangle|\langle p_{1,i}\rangle, p_2, \boldsymbol{\theta}_{45}, \boldsymbol{\theta}_3\big) = \prod_{i=1}^{n} f\big(\tilde{y}_i|p_{1,i}, p_2, \boldsymbol{\theta}_{45}, \boldsymbol{\theta}_3\big). \tag{24}$$

Adopting the formalism and notation of Eq. (9), but omitting the technical details, the definition of the augmented likelihood Eq. (24) rests on the correspondingly transformed probability density

$$f\big(\tilde{y}_i|p_{1,i}, p_2, \boldsymbol{\theta}_{45}, \boldsymbol{\theta}_3\big) = \int\limits_0^1 \int\limits_{-\infty}^{+\infty} \int\limits_{-\infty}^{+\infty} \delta\big(\tilde{y}_i - \mathcal{M}_{p_{1,i},p_2}(p_3, p_4, p_5)\big)\, f_3\big(p_3|\boldsymbol{\theta}_3\big)\, f_{45}\big((p_4, p_5)|\boldsymbol{\theta}_{45}\big)\, \mathrm{d}p_3\,\mathrm{d}p_4\,\mathrm{d}p_5. \tag{25}$$

Here δ denotes the Dirac delta function and $\mathcal{M}_{p_{1,i},p_2}\colon (p_3, p_4, p_5) \mapsto \mathcal{M}(p_{1,i}, p_2, p_3, p_4, p_5)$ formalizes the map that the forward model $\mathcal{M} \equiv h_1$ defines for fixed inputs $(p_{1,i}, p_2)$ and functional arguments (p_3, p_4, p_5). With the combined parametric and structural prior Eq. (23) and the augmented likelihood Eq. (24), the augmented posterior of the unknowns $(\langle p_{1,i}\rangle, p_2, \boldsymbol{\theta}_1, \boldsymbol{\theta}_{45})$ is according to Bayes' law proportional to

$$\pi\big(\langle p_{1,i}\rangle, p_2, \boldsymbol{\theta}_1, \boldsymbol{\theta}_{45}|\langle \tilde{y}_i\rangle, \boldsymbol{\theta}_3\big) \propto \mathcal{L}\big(\langle \tilde{y}_i\rangle|\langle p_{1,i}\rangle, p_2, \boldsymbol{\theta}_{45}, \boldsymbol{\theta}_3\big)\,\pi\big(\langle p_{1,i}\rangle, p_2, \boldsymbol{\theta}_1, \boldsymbol{\theta}_{45}\big). \tag{26}$$

Since we are not interested in inferring experiment-specific realizations $\langle p_{1,i}\rangle$ per se, they are treated as nuisance. Thus the posterior of the QoI $(p_2, \boldsymbol{\theta}_1, \boldsymbol{\theta})$ is found by marginalizing Eq. (26) as follows

$$\pi\big(p_2, \boldsymbol{\theta}_1, \boldsymbol{\theta}_{45}|\langle \tilde{y}_i\rangle, \boldsymbol{\theta}_3\big) \propto \int\limits_0^1 \cdots \int\limits_0^1 \pi\big(\langle p_{1,i}\rangle, p_2, \boldsymbol{\theta}_1, \boldsymbol{\theta}_{45}|\langle \tilde{y}_i\rangle, \boldsymbol{\theta}_3\big)\,\mathrm{d}\langle p_{1,i}\rangle, \tag{27}$$

where $\mathrm{d}\langle p_{1,i}\rangle = \mathrm{d}p_{1,1}\ldots\mathrm{d}p_{1,n}$ as before. In practice the marginal posterior Eq. (27) can be accessed by probing the joint posterior Eq. (26) and simply discarding samples of $\langle p_{1,i}\rangle$. Indeed this marginalization is a prominent application of MCMC sampling.

American Institute of Aeronautics and Astronautics

IV.C.2. Augmented Likelihood Estimation

In practical terms the augmented likelihood Eq. (24) can be estimated analogously to Eq. (10). To that end the response density Eq. (25) is estimated for each $p_{1,i}$ and subsequently evaluated for the given responses $\tilde{y}_i$, respectively. Hence a KDE-based estimate of the augmented likelihood is given as

$$\hat{\mathcal{L}}_{\mathrm{DA}}\big(\langle\tilde{y}_i\rangle|\langle p_{1,i}\rangle, p_2, \boldsymbol{\theta}_{45}, \boldsymbol{\theta}_3\big) = \prod_{i=1}^{n}\left(\frac{1}{K}\sum_{k=1}^{K}\mathcal{K}_{h_i}\left(\tilde{y}_i - \tilde{y}_i^{(k)}\right)\right),$$

$$\text{with}\quad \begin{cases} p_3^{(k)} \sim f_3(p_3^{(k)}|\boldsymbol{\theta}_3), \\ \big(p_4^{(k)}, p_5^{(k)}\big) \sim f_{45}\big((p_4^{(k)}, p_5^{(k)})|\boldsymbol{\theta}_{45}\big), \\ \tilde{y}_i^{(k)} = \mathcal{M}_{p_{1,i}, p_2}(p_3^{(k)}, p_4^{(k)}, p_5^{(k)}), \\ h_i = (4/3K)^{1/5}\,\hat{\sigma}_i. \end{cases} \tag{28}$$

For $k = 1, \ldots, K$ inputs $p_3^{(k)} \sim f_3(p_3^{(k)}|\boldsymbol{\theta}_3)$ and $(p_4^{(k)}, p_5^{(k)}) \sim f_{45}((p_4^{(k)}, p_5^{(k)})|\boldsymbol{\theta}_{45})$ are sampled from the corresponding population distributions, responses $\tilde{y}_i^{(k)} = \mathcal{M}_{p_{1,i}, p_2}(p_3^{(k)}, p_4^{(k)}, p_5^{(k)})$ are computed accordingly and $\hat{\sigma}_i$ denotes the standard deviation of the response samples $(\tilde{y}_i^{(1)}, \ldots, \tilde{y}_i^{(K)})$. The number of samples for each of these estimations is set to $K = 10^3$ and selection of the bandwidths follows the normal reference rule $h_i = (4/3K)^{1/5}\,\hat{\sigma}_i$. Note that for sampling the posterior Eq. (26) by MC^{3}DA the density $f(\tilde{y}_i|p_{1,i}, p_2, \boldsymbol{\theta}_{45}, \boldsymbol{\theta}_3)$ has to be individually estimated for each $p_{1,i}$ with $i = 1, \ldots, n$.

Let us compare the transformed densities Eqs. (9) and (25). The density $f(\tilde{y}_i|p_2, \boldsymbol{\theta}_1, \boldsymbol{\theta}_{45}, \boldsymbol{\theta}_3)$ explicitly depends on $\boldsymbol{\theta}_1$ and thus implicitly involves the uncertainty of $p_{1,i}$. In contrast $f(\tilde{y}_i|p_{1,i}, p_2, \boldsymbol{\theta}_{45}, \boldsymbol{\theta}_3)$ directly depends on $p_{1,i}$ but does not bear reference to $\boldsymbol{\theta}_1$. Hence $f(\tilde{y}_i|p_2, \boldsymbol{\theta}_1, \boldsymbol{\theta}_{45}, \boldsymbol{\theta}_3)$ is the more complex and broader density, whereas $f(\tilde{y}_i|p_{1,i}, p_2, \boldsymbol{\theta}_{45}, \boldsymbol{\theta}_3)$ is simpler and easier to estimate. In turn this means that likelihood estimations for MC^{3}DA are less biased and have a smaller variance than for MC^{3}KS and thus the approach promises a higher degree of posterior fidelity. In Fig. 8 the density $f(\tilde{y}_i|p_{1,i}, p_2, \boldsymbol{\theta}_{45}, \boldsymbol{\theta}_3)$ is shown for values $(p_2, \boldsymbol{\theta}_{45})_{\text{high}}$ and $(p_2, \boldsymbol{\theta}_{45})_{\text{low}}$ that have been chosen as the same values already used in Fig. 3. Following our final results, the value $p_{1,i}$ has been exemplarily chosen as the posterior mean of $p_{1,i}$ with $i = 36$. When comparing Figs. 3 and 8 one can clearly see the essential difference between the densities $f(\tilde{y}_i|p_2, \boldsymbol{\theta}_1, \boldsymbol{\theta}_{45}, \boldsymbol{\theta}_3)$ and $f(\tilde{y}_i|p_{1,i}, p_2, \boldsymbol{\theta}_{45}, \boldsymbol{\theta}_3)$. The latter is distinctly simpler and clearly better resembling a Gaussian density. Moreover it can be seen that the chosen values $p_{1,36}$ and $(p_2, \boldsymbol{\theta}_{45})_{\text{high}}$ lead to a response density consistent with the observation $\tilde{y}_{36}$.
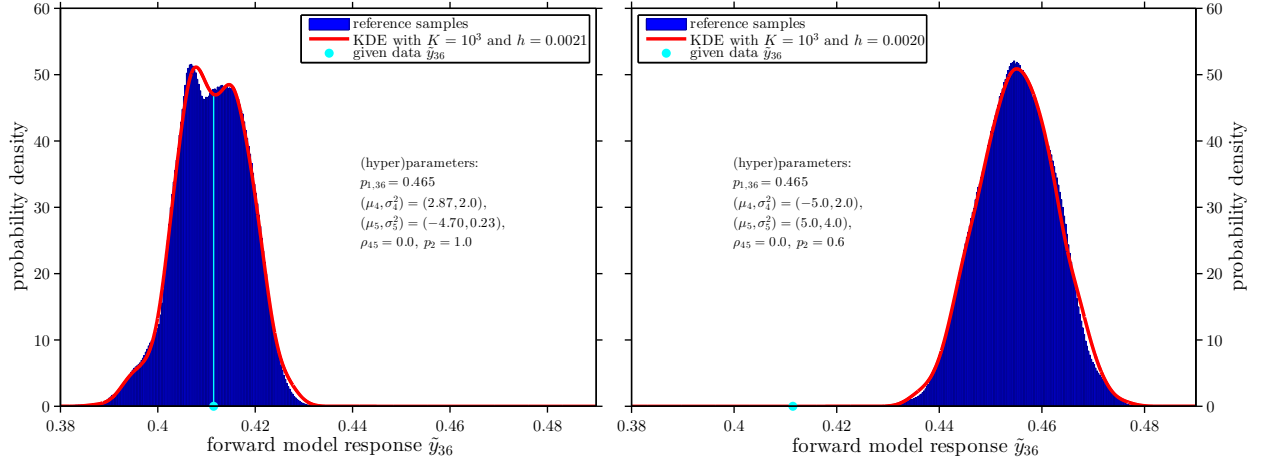

Figure 8. Estimation of $f(\tilde{y}_i|p_{1,i}, p_2, \boldsymbol{\theta}_{45}, \boldsymbol{\theta}_3)$**.** Evaluating the augmented likelihood Eq. (24) for MC^{3}DA is based on the response density Eq. (25). For $p_{1,36} = 0.465$ and two different values of the (hyper)parameters $(p_2, \boldsymbol{\theta}_1, \boldsymbol{\theta}_{45})$ a KDE of $f(\tilde{y}_{36}|p_{1,36}, p_2, \boldsymbol{\theta}_{45}, \boldsymbol{\theta}_3)$ is shown with $K = 10^3$ and automatic bandwidth selection according to the normal reference rule. For reference purposes a histogram is shown for a larger number $K = 10^7$ of simulated forward model responses.

IV.C.3. MCMC

The augmented posterior Eq. (26) will be explored by means of a dedicated MC^{3}DA sampler. Updating is done in blocks $(\langle p_{1,i}\rangle)$, (μ_1, σ_1^2), (p_2), (μ_4), (μ_5) and $(\sigma_4^2, \sigma_5^2, \rho_{45})$. Each $p_{1,i}$ in the block $(\langle p_{1,i}\rangle)$ is concurrently updated with a random walk Metropolis sampler based on independent Gaussian proposals with standard deviation $\sigma_{p_{1,i}} = 0.01$. As before the remaining blocks are initialized in the middle of the corresponding epistemic intervals and updated with independent prior proposals. Acceptance rates amounted to ca. 10 % for $(\langle p_{1,i}\rangle)$, (p_2), and (μ_5), 15 % for (μ_4) and $(\sigma_4^2, \sigma_5^2, \rho_{45})$, and 30 % for (μ_1, σ_1^2). A number of 10219 proposals in the last-mentioned block were rejected due to violating $\alpha_1, \beta_1 > 1$. We start with preliminary MCMC runs with $K = 10^3$ and $h_i = 0.02$ in order to identify the posterior modes of $\langle p_{1,i}\rangle$. Experiment-specific realizations $\langle p_{1,i}\rangle$ are initialized in the middle of their epistemic intervals and converge within ca. 1000 MCMC iterations. The initial convergence and final posterior of an experiment-specific realization $p_{1,i}$ with $i = 10$ are shown in Fig. 9. This shows that individual experiment-specific realizations $\langle p_{1,i}\rangle$ can indeed be inferred. As opposed to classical Bayesian inversion, the width of the posterior is not due to a prior and residual model, but due to posterior uncertainty of the remaining unknowns. The danger of the approach is that missing further posterior modes of $\langle p_{1,i}\rangle$ would alter the sampled posteriors of the remaining unknowns, above all the one of $\boldsymbol{\theta}_1$. Convergence checks have therefore been accomplished by initializing $\langle p_{1,i}\rangle$ within admissible regions of the parameter space that have not been visited in previous runs. Ultimately the chain converged to the same posterior modes which were found before. We therefore initialize the final sampler within these posterior modes of $\langle p_{1,i}\rangle$. With $K = 10^3$ and automatic selection of the bandwidths h_i we draw $N = 10^5$ posterior samples. Total execution time amounts to $t \approx 90\,\mathrm{h}$ on a single core. The resulting posterior marginals of the QoI are added to Figs. 4 to 7. As compared to the results obtained by MC^{3}KS the posteriors found by MC^{3}DA have been slightly shrunk and evolved in structure. Resting upon the assumption that the posterior modes of $\langle p_{1,i}\rangle$ have been correctly identified, we take this as an indication of a gain in posterior fidelity.
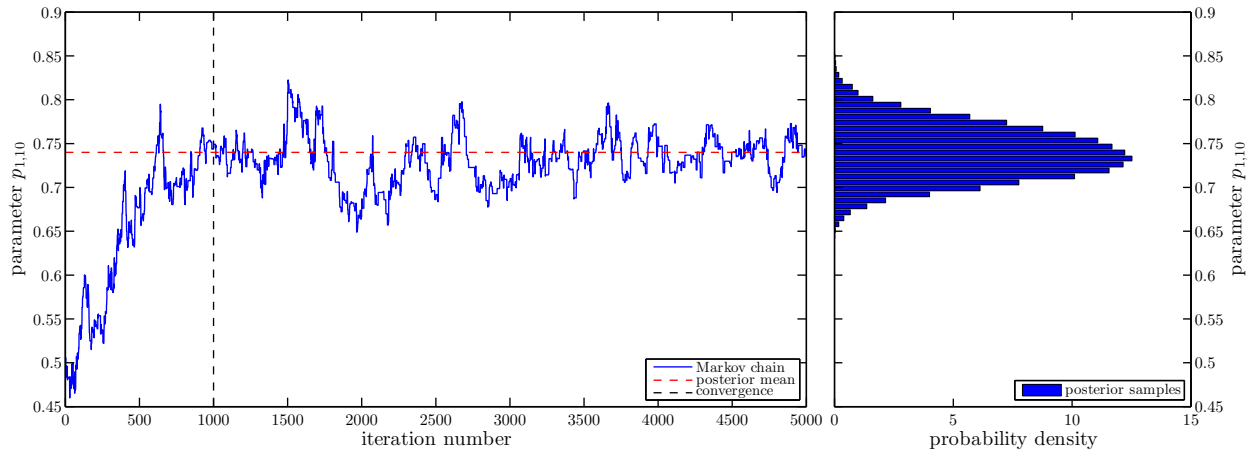

Figure 9. Convergence and identifiability of $p_{1,i}$. With $N = 10^5$ iterations of the MC^{3}DA algorithm the augmented posterior Eq. (26) is explored by analyzing $\langle \tilde{y}_i\rangle_{1\leq i 50}$. For $K = 10^3$ and $h_i = 0.02$ the converging Markov chain and the resulting posterior of an experiment-specific $p_{1,10}$ is shown. The width of the posterior is governed by the posterior uncertainty of the remaining unknowns.

V.   Conclusion & Outlook

Addressing the uncertainty characterization subproblem of the NASA UQ challenge has turned out to be a challenging yet rewarding task. We devised a Bayesian multilevel model involving "perfect" data in order to formulate the problem as statistical inversion and proposed computational techniques dedicated to sampling the corresponding posterior distribution. In turn this problem solution has given rise to new questions relevant to MCMC posterior exploration and statistical likelihood estimation in the context of Bayesian multilevel calibration. First related thoughts were given and an in-depth consideration has been initiated. In sum we hope that these efforts prove to be a solid contribution to the NASA challenge problem in particular and to the theory and practice of Bayesian data analysis and uncertainty quantification in general. Future research work encompasses the design of more sophisticated methods to simulate the likelihood function and the rigorous assessment of the posterior fidelity.

References

[1]Davidian, M. and Giltinan, D. M., "Nonlinear models for repeated measurement data: An overview and update," *Journal of Agricultural, Biological, and Environmental Statistics*, Vol. 8, 2003, pp. 387–419.

[2]Banks, H. T., Kenz, Z. R., and Thompson, W. C., "A review of selected techniques in inverse problem nonparametric probability distribution estimation," *Journal of Inverse and Ill-posed Problems*, Vol. 20, No. 4, 2012, pp. 429–460.

[3]Wakefield, J., "The Bayesian Analysis of Population Pharmacokinetic Models," *Journal of the American Statistical Association*, Vol. 91, No. 433, 1996, pp. 62–75.

[4]Lunn, D. J., "Bayesian Analysis of Population Pharmacokinetic/Pharmacodynamic Models," *Probabilistic Modeling in Bioinformatics and Medical Informatics*, edited by D. Husmeier, R. Dybowski, and S. Roberts, Advanced Information and Knowledge Processing, Springer, London, 2005, pp. 351–370.

[5]Browne, W. J. and Draper, D., "A comparison of Bayesian and likelihood-based methods for fitting multilevel models," *Bayesian Analysis*, Vol. 1, No. 3, 2006, pp. 473–514.

[6]Gelman, A., "Prior distributions for variance parameters in hierarchical models (Comment on Article by Browne and Draper)," *Bayesian Analysis*, Vol. 1, No. 3, 2006, pp. 515–534.

[7]Wakefield, J. C., Smith, A. F. M., Racine-Poon, A., and Gelfand, A. E., "Bayesian Analysis of Linear and Non-Linear Population Models by Using the Gibbs Sampler," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, Vol. 43, No. 1, 1994, pp. 201–221.

[8]Damien, P., Wakefield, J., and Walker, S., "Gibbs sampling for Bayesian non-conjugate and hierarchical models by using auxiliary variables," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, Vol. 61, No. 2, 1999, pp. 331–344.

[9]Dyk, D. A. and Meng, X.-L., "The Art of Data Augmentation," *Journal of Computational and Graphical Statistics*, Vol. 10, No. 1, 2001, pp. 1–50.

[10]Dyk, D. A., "Hierarchical Models, Data Augmentation, and Markov Chain Monte Carlo," *Statistical Challenges in Astronomy*, Springer, New York, 2003, pp. 41–55.

[11]Nagel, J. B. and Sudret, B., "Probabilistic Inversion for Estimating the Variability of Material Properties: A Bayesian Multilevel Approach," *Proc. 11th Int. Probabilistic Workshop, Brno (Czech Republic), November 6-8*, 2013.

[12]Nagel, J. B. and Sudret, B., "A Bayesian Multilevel Approach to Optimally Estimate Material Properties," *Proc. ASCE-ICVRAM-ISUMA 2014, Liverpool (UK), July 13-16*, 2014 (submitted for publication).

[13]Draper, D., Gaver, D. P., Goel, P. K., Greenhouse, J. B., Hedges, L. V., Morris, C. N., and Waternaux, C. M., *Combining Information: Statistical Issues and Opportunities for Research*, Panel on Statistical Issues and Opportunities for Research in the Combination of Information, National Research Council (U.S.), 1992.

[14]Crespo, L. G., Kenny, S. P., and Giesy, D. P., "The NASA Langley Multidisciplinary Uncertainty Quantification Challenge," 2013, NASA Langley Research Center.

[15]Crespo, L. G., Kenny, S. P., and Giesy, D. P., "NASA LaRC UQ Challenge 2014: Challenge Problem FAQ," http://uqtools.larc.nasa.gov/nda-uq-challenge-problem-2014/faq/, 2013.

[16]Draper, D., Hodges, J. S., Mallows, C. L., and Pregibon, D., "Exchangeability and Data Analysis," *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, Vol. 156, No. 1, 1993, pp. 9–37.

[17]Bernardo, J. M., "The Concept of Exchangeability and its Applications," *Far East Journal of Mathematical Sciences*, Vol. 4, 1996, pp. 111–121.

[18]Berger, J. O., Liseo, B., and Wolpert, R. L., "Integrated likelihood methods for eliminating nuisance parameters," *Statistical Science*, Vol. 14, No. 1, 1999, pp. 1–28.

[19]Liseo, B., "The Elimination of Nuisance Parameters," *Bayesian Thinking Modeling and Computation*, edited by D. Dey and C. Rao, Vol. 25 of *Handbook of Statistics*, Elsevier, Amsterdam, Netherlands, 2005, pp. 193–219.

[20]Wand, M. P. and Jones, M. C., *Kernel Smoothing*, Chapman & Hall/CRC Monographs on Statistics & Applied Probability, Vol. 60, Chapman and Hall/CRC, London, 1st ed., 1994.

[21]Robert, C. P. and Casella, G., *Monte Carlo Statistical Method*, Springer Series in Statistics, Springer, New York, 2nd ed., 2004.

[22]Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E., "Equation of State Calculations by Fast Computing Machines," *The Journal of Chemical Physics*, Vol. 21, No. 6, 1953, pp. 1087–1092.

[23]Hastings, W. K., "Monte Carlo sampling methods using Markov chains and their applications," *Biometrika*, Vol. 57, No. 1, 1970, pp. 97–109.

[24]Casella, G. and George, E. I., "Explaining the Gibbs Sampler," *The American Statistician*, Vol. 46, No. 3, 1992, pp. 167–174.

[25]Bal, G., Langmore, I., and Marzouk, Y., "Bayesian inverse problems with Monte Carlo forward models," *Inverse Problems and Imaging*, Vol. 7, No. 1, 2013, pp. 81–105.

[26]Andrieu, C. and Roberts, G. O., "The Pseudo-Marginal Approach for Efficient Monte Carlo Computations," *The Annals of Statistics*, Vol. 37, No. 2, 2009, pp. 697–725.