Available online at www.eccomasproceedia.org
Eccomas Proceedia UNCECOMP (2025) 430-437

ECCOMAS
Proceedia

UNCECOMP 2025
6th ECCOMAS Thematic Conference on
Uncertainty Quantification in Computational Science and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou (eds.)
Rhodes Island, Greece, 15-18 June 2025

BAYESIAN CALIBRATION OF A NON-LINEAR CONTROLLED DYNAMIC SYSTEM WITH MISSING DATA : APPLICATION TO HIGH SPEED-TRAIN

R. Jorge Do Marco[1,2,3], G. Perrin[1], C. Funfschilling[2], and C. Soize[3]

[1]COSYS, Université Gustave Eiffel
14-20 Boulevard Newton, 77447 Marne-la-Vallée, France
e-mail: guillaume.perrin@univ-eiffel.fr

[2]SNCF, DTIPG
1-3 Av François Mitterrand, 93574 Saint-Denis, France
e-mail: {romain.jorge-domarco,christine.funfschilling}@sncf.fr

[3]Université Gustave Eiffel, MSME UMR 8208
5 Bd Descartes, 77454 Marne-La-Vallée, France
e-mail: christian.soize@univ-eiffel.fr

Abstract. *As systems become more complex, physical model calibration has become an essential industrial step. In industrial applications, model calibration is often hampered by the presence of unknown environmental variables that affect the behavior of the system. Traditional approaches assume that all input variables are well known, which can lead to significant errors if this assumption is violated. This issue is particularly critical in dynamic systems where inputs and outputs are functional, making direct measurements inherently approximate. To address this challenge, we adopt a Bayesian framework in which a probability distribution is assigned to the unknown environmental variables. Instead of calibrating the model for a single, arbitrarily chosen set of inputs, we compute the posterior distribution of the calibration parameters and average it over the distribution of the unknown variables. This process not only accounts for input uncertainty, but also helps to mitigate model errors, resulting in a more robust and reliable calibration. The effectiveness of this approach is demonstrated in a railway case study, where numerical experiments highlight its advantages over standard deterministic calibration methods.*

Keywords: Bayesian Calibration, Missing data, Dynamic system, Nonlinearity, Railway dynamic

ISSN:2623-3339 © 2025 The Authors. Published by Eccomas Proceedia.
Peer-review under responsibility of the organizing committee of UNCECOMP 2025.
doi: 10.7712/120225.12368.21114

1 INTRODUCTION

Modern engineering increasingly relies on complex physical models, the accuracy of which depends on precise parameter tuning through a process known as calibration. Calibration integrates multiple sources of information, including measurements and expert knowledge, to refin model predictions. However, industrial applications often suffer from prediction errors due to measurement inaccuracies, model incompleteness (epistemic errors) or inherent randomness (aleatory errors) [1]. In this context, uncertainty quantificatio is crucial for assessing confi dence in calibrated models and for tracing the origin of errors. Among existing calibration approaches, the Bayesian framework introduced by Kennedy and O'Hagan [2] provides a powerful way to incorporate uncertainty while combining multiple sources of information. However, its application to dynamical systems presents several challenges. These systems are governed by differential equations, have a high degree of freedom, and involve functional inputs and outputs that are often difficul to measure simultaneously. In practice, missing or imperfectly known inputs are a major obstacle to effective calibration [3]. While missing data problems have been extensively studied in statistics [4], their impact on model calibration, particularly for dynamic systems, remains an open question. This paper extends the framework of Kennedy and O'Hagan by explicitly addressing the challenge of calibrating models with not perfectly known input variables. Our approach is to assign a probability distribution to these partially known inputs and compute expectations over all plausible values. This process not only enhances robustness, but also helps to mitigate model errors by naturally integrating uncertainty propagation. However, this probabilistic treatment significantl increases the computational cost, necessitating a dual optimization strategy: (i) constructing a generator that ensures input-output coherence, and (ii) using surrogate models to maintain computational feasibility. In addition, we analyze how uncertainty propagates throughout the calibration process to improve the reliability of stochastic surrogate models. The proposed methodology is applied to the railway industry, where numerical experiments highlight its advantages over standard deterministic calibration methods.

The outline of this work is as follows. Section 2 presents the theoretical framework associated with the handling of missing data and model discrepancy. An application to the railway industry is then given in Section 3.

2 METHODOLOGY

2.1 Problem definition

In this work, we are interested in estimating the physical properties of a dynamical system of interest based on $n_{\text{meas}} > 1$ independent measurements taken in different environments. For each measurement $i \in [\![1; n_{\text{meas}}]\!]$, let $t_{f,i}$ be the duration of the measurement i, $t \in [0; t_{f,i}]$ be the time, $\mathbf{x}_i := \{x_i(t) \in \mathbb{X} \subset \mathbb{R}^p, t \in [0; t_{f,i}]\}$ be the state vector of the system, $u_i := \{u_i(t), t \in [0; t_{f,i}]\}$ be the system control function, and $\mathcal{T}_i$ be the environmental properties. The function u_i is assumed to belong to $\mathbb{U}(t_{f,i}) := \mathcal{F}\left([0; t_{f,i}]\,; [-1; 1]\right)$, which is the set of functions define on $[0; t_{f,i}]$ whose values are in $[-1; 1]$. For a given control function $u_i \in \mathbb{U}(t_{f,i})$, the state vector $\mathbf{x}$ should verify a deterministic parametric equation given by

$$\forall t \in [0; t_{f,i}]\,, \quad \dot{\mathbf{x}}_i(t) = \mathbf{f}\left(\mathbf{x}_i(t), u_i(t), \mathcal{T}_i; \mathbf{z}_f\right). \tag{1}$$

In this equation, the model $\mathbf{f}$ is assumed to be known, but the vector of parameters $\mathbf{z}_f \in \mathbb{Z}_f$ on which it depends is *a priori* unknown. The model $\mathbf{f}$ is considered simplifie in the sense that it provides a reasonable representation of the dynamics of the system, without being perfect (presence of model discrepancies due to modeling simplifications) It is also considered deterministic in the sense that, for given values of $\mathcal{T}_i$ and u_i, and for a given initial state $\mathbf{x}_{0,i} = \mathbf{x}_i(0)$ and a given value of $\mathbf{z}_f$, it yields a unique value for the state function, denoted by

$$\mathbf{x}(\cdot; u_i, \mathbf{z}_f, \mathcal{T}_i, \mathbf{x}_{0,i}) := \begin{cases} [0; t_{f,i}] & \rightarrow & \mathbb{X} \\ t & \mapsto & \mathbf{x}(t; u_i, \mathbf{z}_f, \mathcal{T}_i, \mathbf{x}_{0,i}) \end{cases}. \tag{2}$$

The system dynamics must also satisfy a set of constraints, which can be represented by an envelope $\mathbb{X}_i^{\text{valid}}$ containing lower and upper bounds for each component of $\mathbf{x}_i$ at each time t. The performance evaluation of the system under consideration involves another real-valued function h, which also depends on $\mathbf{x}_i$, u_i and $\mathcal{T}_i$. The function h is assumed to be known, but may additionally depend on parameters that are not perfectly known, denoted by $\mathbf{z}_h \in \mathbb{Z}_h$. We then denote by e_i the integral in time of this quantity of interest, so that for each measurement $1 \leq i \leq n_{\text{meas}}$ and each $t \in [0; t_{f,i}]$,

$$e_i(t) = \int_0^t h(\mathbf{x}(\tau; u_i, \mathbf{z}_f, \mathcal{T}_i, \mathbf{x}_{0,i}), u_i; \mathbf{z}_h) d\tau. \tag{3}$$

Similar to the state vector, there may be constraints on the values of h in the form of time-dependent lower and upper bounds. In the same way as for the dynamics model, the function h can also be affected by a model discrepancy due to modeling simplifications The quantity e_i, as a composition of the models h and $\mathbf{f}$, is thus *a priori* affected by two different model errors, whose amplitudes are characterized by two scalars σ_f and σ_p (the precise definition of these scalars, which depend on the application considered, are given in Section 3.1). We also introduce $\boldsymbol{\sigma} = (\sigma_f, \sigma_p) \in \mathbb{S}$ as the vector of hyperparameters characterizing the model errors.

The aim of this work is to propose a method for estimating the value of $\mathbf{z} := (\mathbf{z}_f, \mathbf{z}_h) \in \mathbb{Z} := \mathbb{Z}_f \times \mathbb{Z}_h$ as accurately as possible from point measurements of $\mathbf{x}_i$ and e_i for each $i \in [\![1; n_{\text{meas}}]\!]$ in the context of missing data, in the sense that the control values u_i applied during the n_{meas} measurements are imperfectly known.

2.2 Bayesian framework

In this paper we adopt a Bayesian framework for estimating the vector $\mathbf{z}$. It is therefore modeled as a random vector $\mathbf{Z}$, and we denote its Probability Density Function (PDF) by $f_{\mathbf{Z}}$. Obtaining this PDF depends on the prior available information about this vector [1, 5], an example is given in Section 3. In addition to this prior knowledge, we have access to discrete and noisy observations of the system behavior, which are assumed to be gathered in the vector $\mathbf{y} \in \mathbb{Y}$. Estimating $\mathbf{Z}$ thus amounts to approximating the PDF of $\mathbf{Z}|\mathbf{Y} = \mathbf{y}$, which can be written as follows using Bayes' theorem:

$$f_{\mathbf{Z}|\mathbf{Y}=\mathbf{y}}(\mathbf{z}) = c \times f_{\mathbf{Z}}(\mathbf{z}) \times f_{\mathbf{Y}|\mathbf{Z}=\mathbf{z}}(\mathbf{y}), \tag{4}$$

where c is a constant that does not depend on $\mathbf{z}$ and $f_{\mathbf{Y}|\mathbf{Z}=\mathbf{z}}$ is the likelihood function. To make this likelihood function explicit, let us denote by $\mathbf{y}^{\text{mod}}(\mathbf{z}, \mathbf{u})$ the vector of predictions of the quantities $\mathbf{y}$ that can be made by evaluating the codes $\mathbf{f}$ and h. This vector depends on the vector $\mathbf{z}$ to be identified but also requires knowledge of the various control functions u_i applied during each measurement, which are grouped together in the command list $\mathbf{u}$. The link between the observations and the code predictions is then written:

$$\mathbf{y} = \mathbf{y}^{\text{mod}}(\mathbf{z}, \mathbf{u}) + \boldsymbol{\varepsilon}^{\text{meas}} + \boldsymbol{\varepsilon}^{\text{mod}}(\boldsymbol{\sigma}), \tag{5}$$

where $\varepsilon^{\text{meas}}$ is the measurement noise, modeled as a random vector with known PDF, and $\varepsilon^{\text{mod}}(\boldsymbol{\sigma})$ is the model error, also modeled as a random vector whose PDF is chosen in a parametric family indexed by $\boldsymbol{\sigma}$. Since this vector $\boldsymbol{\sigma}$ is *a priori* unknown, we also propose to model it as a random vector noted $\mathbf{S}$ with prior PDF $f_{\mathbf{S}}$. It comes:

$$f_{\mathbf{Z}|\mathbf{Y}=\mathbf{y}}(\mathbf{z}) = c \times f_{\mathbf{Z}}(\mathbf{z}) \times \mathbb{E}_{\mathbf{S}|\mathbf{Z}=\mathbf{z}} \left[f_{\mathbf{Y}|\mathbf{Z}=\mathbf{z},\mathbf{S}}(\mathbf{y}) \right] . \tag{6}$$

The peculiarity of this work also stems from the fact that the commands gathered in $\mathbf{u}$ are only partially known. We therefore propose to model them again as random quantities, in order to incorporate their uncertain nature into the calibration process. We then denote by $\mathbf{U}$ the list of random commands and we assume that we are capable of generating independent realizations of U. This allows us to adapt Eq. (6) to obtain

$$f_{\mathbf{Z}|\mathbf{Y}=\mathbf{y}}(\mathbf{z}) = c \times f_{\mathbf{Z}}(\mathbf{z}) \times \mathbb{E}_{\mathbf{U}|\mathbf{Z}=\mathbf{z}} \left[\mathbb{E}_{\mathbf{S}|\mathbf{Z}=\mathbf{z},\mathbf{U}} \left[f_{\mathbf{Y}|\mathbf{Z}=\mathbf{z},\mathbf{U},\mathbf{S}}(\mathbf{y}) \right] \right] , \tag{7}$$

where for each couple of realizations $(\mathbf{u}, \boldsymbol{\sigma})$ of $(\mathbf{U}, \mathbf{S})$, it is now possible to calculate the PDF $f_{\mathbf{Y}|\mathbf{Z}=\mathbf{z},\mathbf{U}=\mathbf{u},\mathbf{S}=\mathbf{s}}(\mathbf{y})$ by evaluating the numerical codes $\mathbf{f}$ and h at each measurement. However, the presence of this double expectation makes it very difficul to evaluate $f_{\mathbf{Z}|\mathbf{Y}=\mathbf{y}}$ at any value of $\mathbf{z}$. Significan adaptations and simplification (by combining Markov Chain Monte Carlo (MCMC) (see [1]) approaches with machine learning approaches such as Gaussian process regression [6]) have therefore to be proposed to obtain a set of values of $\mathbf{z}$ that can be assumed to correspond to a sample of identical and identically distributed (i.i.d.) realizations of $\mathbf{Z}$ according to $f_{\mathbf{Z}|\mathbf{Y}=\mathbf{y}}$.

3 IDENTIFICATION OF THE PARAMETERS OF A HIGH-SPEED TRAIN

3.1 Context

This work focuses on a railway application, and we are interested in calibrating two simplifie models (see [7, 8] for their precise definiti n): a dynamic model $\mathbf{f}$ and an energetic model h. Each of these models is parameterized, and we note $\mathbf{z}$ the vector of parameters that needs to be specifie for these models to be evaluated. Vector $\mathbf{z}$ is made of 9 components,

$$\mathbf{z} = (m, a, b, c, a_\eta, b_\eta, c_\eta, d_\eta, p_{\text{aux}}) , \tag{8}$$

where m is the mass of the train, a, b, and c are Davis coefficient (see [9]), a_η, b_η, c_η and d_η are yield coefficient and p_{aux} is the train auxiliary power. Applying the maximum entropy principle (see [1]), it is possible to associate prior PDFs to the components of $\mathbf{z}$, which are shown in Figure 1. The two mechanical models are assumed to be affected by model errors, which are modeled by two Gaussian white noises of amplitudes σ_f and σ_p respectively: σ_f is the standard deviation of the noise for the train dynamical efforts, and σ_p is the standard deviation of the noise affecting the modeling of the electric power. These two errors are gathered in $\boldsymbol{\sigma}$, and we also associate a prior PDF to this vector (we are concentrating here on Jeffreys' priors [10] to minimize the impact of this choice on the calibration results).

In this work, we focus on a numerical experiment using synthetic data. We consider 20 measurements over several railway tracks and several running environments for the train. For these different measurements, simulated values of train dynamics and energy consumption are then calculated for a reference value of $\mathbf{z}$ noted $\mathbf{z}^*$. From these simulations, we extract the train speed and the power consumed by the train as a function of the curvilinear abscissa of the track at a discretization step of $0.2s$. The reference values for these quantities are then noised to give

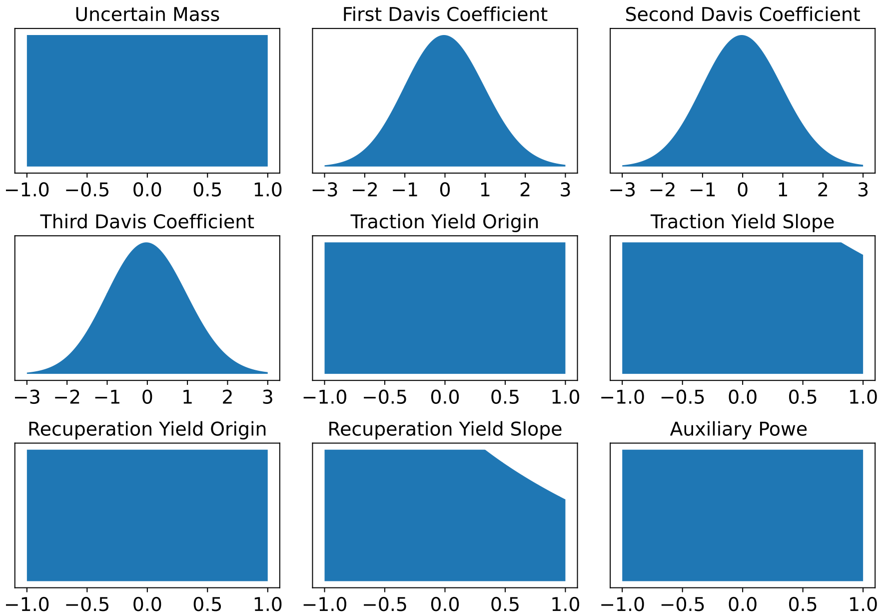

Figure 1: Normalized prior of $\mathbf{Z}$.

the observations available for calibrating the model parameters. Half of these measurements (the firs ten) are used for the calibration phase, while the other half (the last ten measurements) will be used to assess the relevance of the results obtained. The observations associated with the firs ten measurements are gathered in the vector $\mathbf{y}$. In our case, the model error arises from a difference in the discretization of the differential equations modeling the movement of the train. Whereas the reference values are based on a Runge Kutta numerical scheme of order 4 with a step of $0.01s$, the model we are trying to calibrate is solved by a Runge Kutta numerical scheme of order 4 but with a step of $0.1s$.

3.2 Control definition

The dynamic models are highly sensitive to the choice of the driver commands u_i. In our case, a value of $u_i(t) = 1$ indicates maximum mechanical traction, while a value of $u_i(t) = -1$ indicates maximum braking. For this work, we assume that we only have a discretization of these commands every $0.2s$ (using the same time step as the speed and power measurements). This discretization step does not exactly match the command changing instants (it is nevertheless assumed that by discretizing every $0.2s$, no command changes are missed). We would therefore like to quantify the influenc of this loss of information on the estimate of $\mathbf{z}^*$. To this end, we study three cases:

1. We call *true control* the case where we assume that the value of the driver commands is known. This configuratio will be used as a reference for the results.

2. We call *unique control* the case where we assume that the true control is exactly given by a constant piecewise interpolation of the available discretizations of the u_i.

3. We call *Generator control* the case where we do not consider a single control in the inversion, but a collection of controls, all associated with control change times uniformly distributed over the $0.2s$ separating the potential control jumps (see Figure 2 for an illustration).

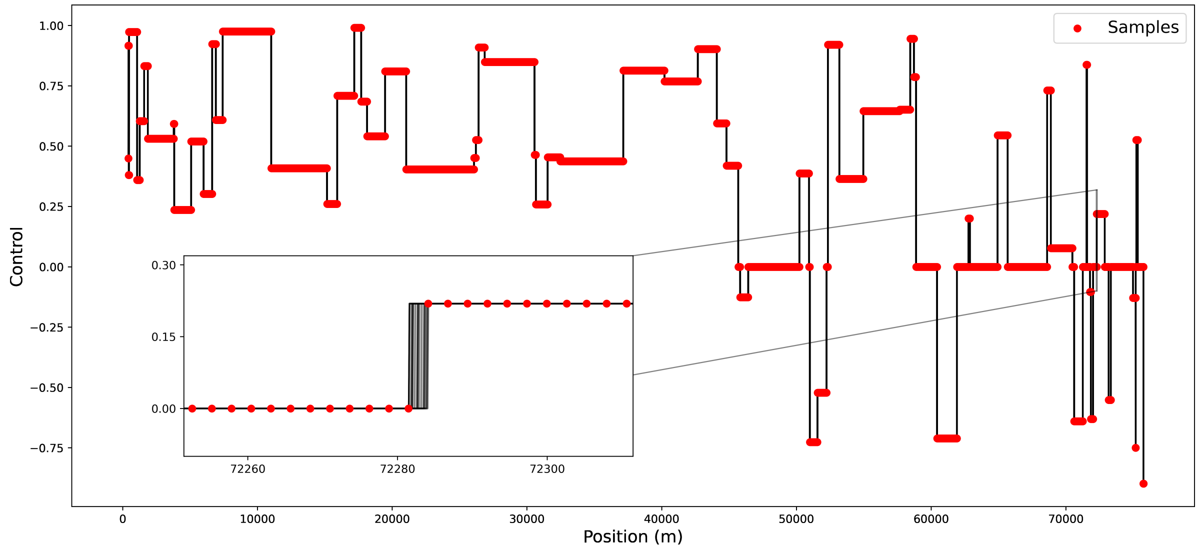

Figure 2: Representation of several command functions (solid black lines) associated with a same discretization (red dots) of the true but unknown driver command.

Note that the two firs cases are specifi draws of the third case.

3.3 Results

The *a posteriori* PDF of the model parameters was calculated independently in the three configuration listed above using Markov Chain Monte Carlo (MCMC) algorithms. The evaluation of the results is based on several criteria:

1. The Mean Square Error (MSE) of $\mathbf{Z}^{\text{post}} = (\mathbf{Z}|\mathbf{Y} = \mathbf{y})$ is noted

$$\delta_{\text{mse}} = \mathbb{E}\left[\|\mathbf{Z}^{\text{post}} - \mathbf{z}^*\|_2^2\right] \tag{9}$$

   and assesses the extent to which the obtained *a posteriori* PDF is focused on true reference value $\mathbf{z}^*$;

2. The estimated amplitudes of the model error terms, which are define such that

$$\boldsymbol{\sigma}^* = \arg\left(\max_{\boldsymbol{\sigma}\in\mathbb{S}}\left(\max_{\mathbf{z}\in\mathbb{Z}}\left(f_{\mathbf{Z},\mathbf{S}|\mathbf{Y}=\mathbf{y}}(\mathbf{z},\boldsymbol{\sigma})\right)\right)\right), \tag{10}$$

   allow us to quantify the impact on the mechanical models of not perfectly knowing the true command;

3. The reconstruction errors are given by

$$\forall i \in [\![1;10]\!], \quad \delta^i_{\text{rec,tot}} = \frac{\|\dot{s}_{\text{true},i} - \mathbb{E}_{\mathbf{Z}}\left[\dot{s}\left(\mathbf{Z}, u_i^*\right)\right]\|_2^2}{\|\dot{s}_{\text{true},i}\|_2^2} + \frac{\|\dot{e}_{\text{true},i} - \mathbb{E}_{\mathbf{Z}}\left[\dot{e}\left(\mathbf{Z}, u_i^*\right)\right]\|_2^2}{\|\dot{e}_{\text{true},i}\|_2^2}, \tag{11}$$

   where for each measurement i that was not used during the calibration phase, $\dot{s}_{\text{true},i}$ and $\dot{e}_{\text{true},i}$ are the true train speed and true consumed electrical power (that is to say associated with the true control function u_i^* and the true model parameters $\mathbf{z}^*$, whereas $\dot{s}\left(\mathbf{z}, u_i^*\right)$ and $\dot{e}\left(\mathbf{z}, u_i^*\right)$ are the train speed and consumed electrical power we obtain when running the models by setting the model parameters to $\mathbf{z}$.

The results are presented in Table 1 and Figure 3. They show that averaging over the control (*Generator control*) allows a better calibration with smaller model errors, a smaller MSE and a smaller reconstruction error. Surprisingly, we also note that we have a lower value for σ_p^* for the *unique control* case compared to the *true control* case: this is mainly due to the fact that the error in the effort (mainly governed by σ_f) is also propagated in the electrical power. Thus, the increased value of σ_f^* may cause a small decrease in σ_p^*. Behind a smaller model error we expect a better identificatio of the parameters. This is the case as shown in Table 1 (note that the high values of the MSE obtained for the *generator control* and the *unique control* cases are led by a unique poorly identifie model parameter). According to Figure 3, we also observe that the *generator control* case leads to reconstruction results that are between the ones associated with the *unique control* and the *true control* cases, which is exactly what we wanted to do. Indeed, by averaging over the driver controls, we expect to reduce the model errors that need to be introduced to match the measurements to the simulations, and through this, we expect a more accurate estimation of the model parameters.

Metric	**True control**	**Generator**	**Unique control**
δ_{mse}	1.18	5.34	5.72
σ_f^*	3.07	3.10	3.45
σ_p^*	5.89	5.87	5.88

Table 1: Synthesis of the results obtained for the MSE (obtained by averaging over 50 000 individuals drawn from the posterior) and for the estimated model errors.

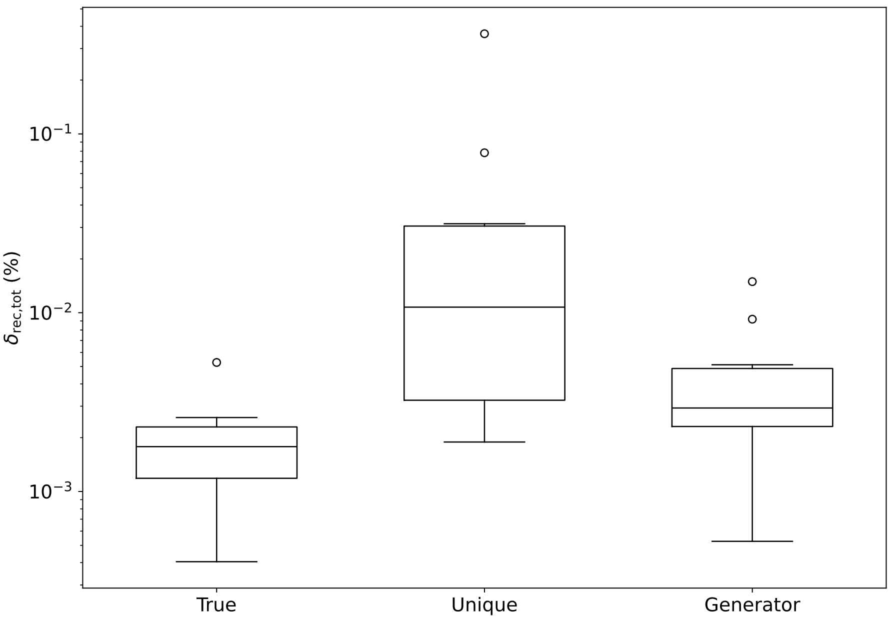

Figure 3: Reconstruction errors obtained for both speed profile and power profile for the 10 measurements that were not used in the calibration process.

4 CONCLUSIONS

In conclusion, this work proposes an application of the Kennedy and O'Hagan framework to handle configuration in presence of missing data. By considering, in the model calibration phase, a control generator rather than focusing on a unique control, we better explain the model variability and thus reduce the model errors. Furthermore, we also show that the reconstruction error is reduced when a control generator is used. As a perspective for this work, we could study the influenc of the generator definition Indeed, in our case the lack of knowledge is limited. It could be useful to study cases where the lack of knowledge increases and tends towards the case of completely unmeasured controls.

References

1. Soize, C. *Uncertainty quantification: An accelerated course with advanced applications in computational engineering* ISBN: 978-3-319-85372-7 (Springer International Publishing, Cham, Switzerland, 2018).
2. Kennedy, M. C. & O'Hagan, A. Bayesian calibration of computer models. en (2001).
3. Perrin, G. & Durantin, C. Taking into account input uncertainties in the Bayesian calibration of time-consuming simulators. *Journal de la société française de statistique* **160,** 24–46 (2019).
4. Little, R. J. A. & Rubin, D. B. *Statistical analysis with missing data* 3rd ed. ISBN: 978-0-470-52679-8 (John Wiley & Sons, Nashville, TN, 2019).
5. Robert, C. P. *The Bayesian choice: from decision-theoretic foundations to computational implementation* 2nd ed. en. ISBN: 978-0-387-71598-8 (Springer, New York, 2007).
6. Rasmussen, C. E. & Williams, C. K. I. *Gaussian processes for machine learning* OCLC: ocm61285753. ISBN: 978-0-262-18253-9 (MIT Press, Cambridge, Mass, 2006).
7. Jorge Do Marco, R., Perrin, G., Funfschilling, C. & Soize, C. *Constrained Optimization of Driver Control to Limit Energy Consumption* in *Proceedings of the Sixth International Conference on Railway Technology: Research, Development and Maintenance* (ed Pombo, J.) **CCC 7** (Civil-Comp Press, Edinburgh, UK, 2024), Paper 21.3.
8. Nespoulous, J., Soize, C., Funfschilling, C. & Perrin, G. Optimisation of train speed to limit energy consumption. en. *Vehicle System Dynamics* **60,** 3540–3557. ISSN: 0042-3114, 1744-5159. https://www.tandfonline.com/doi/full/10.1080/00423114.2021.1965628 (2023) (Oct. 2022).
9. Davis, W. J. *The tractive resistance of electric locomotives and cars* en (General Electric, 1926).
10. Jeffreys, H. An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences* **186.** _eprint: https://royalsocietypublishing.org/doi/pdf/10.1098/rspa.1946.0056, 453–461. https://royalsocietypublishing.org/doi/abs/10.1098/rspa.1946.0056 (1946).