

# MEASUREMENT AND VERIFICATION OF BUILDING SYSTEMS UNDER UNCERTAIN DATA: A GAUSSIAN PROCESS MODELING APPROACH

MICHAEL C. BURKHART <sup>1</sup>, YEONSOOK HEO <sup>2,†</sup>, AND VICTOR M. ZAVALA <sup>1</sup>

<sup>1</sup>*Mathematics and Computer Science Division, Argonne National Laboratory  
9700 South Cass Avenue, Argonne, IL 60439, USA*

<sup>2</sup>*Department of Architecture, University of Cambridge  
1-5 Scroope Terrace, Cambridge, CB2 1PX, UK*

ABSTRACT. Uncertainty in sensor data (e.g., weather, occupancy) complicates the construction of baseline models for measurement and verification (M&V). We present a Monte Carlo expectation maximization (MCEM) framework for constructing baseline Gaussian process (GP) models under uncertain input data. We demonstrate that the GP-MCEM framework yields more robust predictions and confidence levels compared with standard GP training approaches that neglect uncertainty. We argue that the approach can also reduce data needs because it implicitly expands the data range used for training and can thus be used as a mechanism to reduce data collection and sensor installation costs in M&V processes. We analyze the numerical behavior of the framework and conclude that robust predictions can be obtained with relatively few samples.

**Keywords:** Gaussian process modeling, data uncertainty, expectation maximization, measurement and verification.

## 1. INTRODUCTION

Gaussian process (GP) modeling is a powerful statistical modeling framework that proposes a structure for the covariance matrix of input variables to compute predictions of output variables [20]. A Bayesian framework is used to train hyperparameters of the input covariance matrix and to derive a predictive distribution for output variables at test input points. As a result, GP models can capture complex nonlinear relationships between multiple input and output variables and can provide mean predictions and associated uncertainty levels. These features make GP particularly attractive for measurement and verification (M&V), as they can help assess the amount and resolution of data required to reach desired uncertainty levels. Using real and simulated studies, Heo, Zavala, and coworkers [10, 8] have demonstrated that GP models can systematically capture the effect of multiple weather variables (e.g., ambient temperature, relative humidity) as well as occupancy variables on energy demands and can effectively quantify uncertainty levels.

The standard GP modeling framework [20] has been applied in a number of building applications. For M&V, GP models were developed to predict baseline energy use during the post-retrofit period [10], [30]. In fault diagnosis and detection, GP models served to predict system performance

---

<sup>†</sup> Corresponding author. Email: yh305@cam.ac.uk, Tel: (+44)1223760114.

baselines on the basis of measured operational data [24]. In addition, GP models have been used as surrogates of complex energy simulation models to reduce computational complexity of tasks such as model calibration [9], model-predictive control [25], and optimal design [14].

The *standard GP framework assumes that input data are known with full certainty*, but such is not the case in many situations. In particular, weather conditions and building occupancy (key variables suggested in M&V protocols [12, 1]) are often obtained from noisy sensors or they might be too expensive to measure on site. For instance, government-published weather data are measured at distant meteorological stations and hence may not capture microclimate conditions around the specific building site. This situation can lead to significant discrepancy in predicted energy use [21, 22]. In addition, building occupancy is difficult to measure in real time and with high accuracy [28]. Occupant densities in office buildings can vary between 4.3 m<sup>2</sup> and 22.8 m<sup>2</sup> per person, and this range significantly impacts internal heat gains [15, 11]. Ultimately, variations in input data can distort energy predictions and can lead to inconsistent confidence levels, making M&V conclusions unreliable.

This work extends the standard GP framework reported in the literature by explicitly accounting for uncertain input data in the learning of the hyperparameters (i.e., model training). We use the framework of Quiñonero-Candela [18] in which the hyperparameters are learned by maximizing the likelihood function with marginalized input data (also known as the marginal evidence). The maximization of the marginal evidence is performed by using a Monte Carlo expectation maximization (MCEM) algorithm. We call the resulting framework GP-MCEM, which we summarize as follows. The exact EM algorithm [5] is a coordinate ascent technique that iteratively searches for the maximum of the log marginal evidence by alternating expectation steps (E-steps) that maximize a lower bound of the log marginal evidence with respect to a target distribution and maximization steps (M-steps) that maximize the lower bound with respect to the hyperparameters. The M-step requires the maximization of the expected value of the likelihood function with respect to the input data distribution conditional to the output data and hyperparameters. Because the expectation function is a multidimensional integral and cannot be computed exactly, Monte Carlo estimates are used instead by sampling the conditional input data distribution. This gives rise to the so-called Monte Carlo EM (MCEM) method [23].

Using a simulation setting to emulate the energy performance of an advanced multivariable control system in an office building, we demonstrate that more consistent uncertainty estimates of energy demands can be obtained by using the GP-MCEM framework compared with standard GP approaches that neglect input uncertainty. In addition, we demonstrate that much more robust estimates of hyperparameters can be obtained. In particular, we argue that even an ideal GP approach (learning hyperparameters by using perfect input data) can suffer from prediction robustness when confronted with test points slightly outside the training set and, consequently, large amounts of data can thus be needed to mitigate this lack of robustness. The proposed GP-MCEM approach, on the other hand, implicitly expands the input data range used for training and can thus obtain more robust hyperparameters that perform well outside the training set and can thus reduce training data needs. We perform computational experiments with the GP-MCEM framework to validate performance with varying numbers of Monte Carlo samples.

The paper is structured as follows. In Section 2 we derive the GP-MCEM framework, in Section 3 we discuss the algorithmic implementation and convergence properties. In Section 4 we present a detailed numerical study to demonstrate the advantages of the proposed framework. The paper closes in Section 5 with conclusions and directions of future work.

## 2. GP-MCEM FRAMEWORK

In this section, we derive the framework to train GP models under uncertain input data.

**2.1. Setting.** Consider a training set with  $n$  output data points (e.g., daily energy demands)  $\{y^i\}$ , where  $y^i \in \mathbb{R}$ ,  $i = 1, \dots, n$ . Consider also a set of uncertain input data points (e.g., occupancy level, ambient temperature, relative humidity)  $\{\mathbf{x}^i\}$  where  $\mathbf{x}^i \in \mathbb{R}^d$ ,  $i = 1, \dots, n$ . Each element of vector  $\mathbf{x}^i$  is given by  $\mathbf{x}_j^i$ ,  $j = 1, \dots, d$  and  $d$  is the number of input variables. We also define the output vector  $\mathbf{y} := \{y^i\} \in \mathbb{R}^n$  and the input matrix  $X = \{\mathbf{x}_j^i\} \in \mathbb{R}^{n \times d}$ , where the  $i$ th input vector  $\mathbf{x}^i$  forms the  $i$ th row of the matrix  $X$ .

Each input data observation  $\mathbf{x}^i$  is assumed to be random and independent normally distributed as  $\mathbf{x}^i | \phi^i \sim \mathcal{N}(\mathbf{m}^i, \mathbf{S})$ ,  $i = 1, \dots, n$ . Here,  $\phi^i = \{\mathbf{m}^i, \mathbf{S}\}$  are the statistics of observation  $\mathbf{x}^i$  where  $\mathbf{m}^i$  is the mean of each observation  $\mathbf{x}^i$ ,  $i = 1, \dots, n$ . We assume that  $\mathbf{S} := \text{diag}(s_1^2, \dots, s_d^2) \in \mathbb{R}^{d \times d}$  where  $s_j^2 \in \mathbb{R}$  is the variance of variable  $\mathbf{x}_j^i$ ,  $j = 1, \dots, d$ . Consequently, the variance for each  $j$ th variable is constant across the observations  $i = 1, \dots, n$ . In practical terms, this amounts to assuming a different distribution for each input variable  $j = 1, \dots, d$ , but the variance of each variable does not depend on the observation  $i = 1, \dots, n$ . In addition, it assumes that input variables are uncorrelated. From a conceptual standpoint, these assumptions are not necessary for the proposed framework but they greatly simplify implementation of sampling procedures, which we will discuss later. The statistics (i.e., mean and variance) for  $X$  are grouped as  $\Phi = \{\phi^i\}$ ; and, because of the assumed structure, we have that the conditional prior of  $X$  (conditional distribution of  $X$  given the statistics  $\Phi$ ) is given by

$$(2.1) \quad p(X | \Phi) = \prod_{i=1}^n p(\mathbf{x}^i | \phi^i),$$

where

$$(2.2) \quad p(\mathbf{x}^i | \phi^i) = (2\pi)^{-d/2} |\mathbf{S}|^{-1/2} \exp\left(-\frac{1}{2}(\mathbf{x}^i - \mathbf{m}^i)^T \mathbf{S}^{-1}(\mathbf{x}^i - \mathbf{m}^i)\right), \quad i = 1, \dots, n.$$

Here, the matrix determinant is denoted as  $|\cdot|$ . The conditional distribution of  $y$  given the input data  $X$  and hyperparameters  $\theta \in \mathbb{R}^p$  is given by

$$(2.3) \quad p(y | X, \theta) = (2\pi)^{-n/2} |\mathbf{K}(\theta, X, X)|^{-1/2} \exp(-\frac{1}{2}\mathbf{y}^T \mathbf{K}(\theta, X, X)^{-1}\mathbf{y}).$$

Here,  $\mathbf{K}(\theta, X, X) = \{k(\theta, \mathbf{x}^i, \mathbf{x}^j)\} \in \mathbb{R}^{n \times n}$  is the input matrix covariance,  $\theta \in \mathbb{R}^p$  are the hyperparameters, and  $k : \mathbb{R}^p \times \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$  is the covariance function. While the methods developed here are applicable to any proper covariance function, we focus on the standard Gaussian covariance used in most of the literature [10, 30, 24, 9, 14],

$$(2.4) \quad k(\theta, \mathbf{x}^i, \mathbf{x}^j) = \theta_0 \exp\left(-\frac{\|\mathbf{x}^i - \mathbf{x}^j\|^2}{\theta_1}\right) + \theta_2 \delta_{ij}.$$

For alternative covariance functions see [20]. Here, we define the hyperparameter vector  $\theta := [\theta_0, \theta_1, \theta_2]$ , the Kronecker delta  $\delta_{ij}$ , and the Euclidean norm  $\|\cdot\|$ .

**2.2. Marginal Evidence Derivation.** Quiñero-Candela [18] proposes a couple of approaches to train a GP model under uncertain inputs. The first approach maximizes the joint posterior  $p(X, \theta | \mathbf{y}, \Phi) \propto p(\mathbf{y} | X, \theta)p(X | \Phi)$  with respect to  $\theta$  and  $X$ . This can in principle be done because  $p(\mathbf{y} | X, \theta)p(X | \Phi)$  has an explicit, given by the product of Equations (2.1) and (2.3). Consequently, the log of the joint posterior is separable. However, this approach proved to be unreliable because of the introduction of multiple local minima. The second approach, used in this work, consists of computing a maximum a posteriori estimate (MAP). This is achieved by maximizing the posterior  $p(\theta | \mathbf{y}, \Phi) \propto p(\mathbf{y} | \theta, \Phi)p(\theta)$ . Assuming a flat prior  $p(\theta)$ , this is equivalent to maximizing  $p(\mathbf{y} | \theta, \Phi)$ . In addition, we know that

$$\begin{aligned} p(\mathbf{y} | \theta, \Phi) &= \int p(\mathbf{y}, X | \theta, \Phi) dX \\ (2.5) \qquad &= \int p(\mathbf{y} | \theta, X)p(X | \Phi) dX, \end{aligned}$$

where  $\int p(\mathbf{y} | \theta, X)p(X | \Phi) dX$  is the so-called *marginal evidence*. Consequently, we can obtain the MAP estimate of the hyperparameters  $\theta$  by maximizing the log marginal evidence,

$$(2.6) \qquad \mathcal{L}(\theta) := \log \left( \int p(\mathbf{y} | \theta, X)p(X | \Phi) dX \right) = \log \left( \int p(\mathbf{y}, X | \theta, \Phi) dX \right).$$

From a practical standpoint, we note that the *marginal evidence is the likelihood function  $p(\mathbf{y} | \theta, X)$  averaged over the input data*. The likelihood function is typically maximized in standard parameter estimation procedures in which input data observations are assumed to be given (certain). This observation is important because it highlights that maximizing the marginal evidence has the implicit effect of capturing the entire range of input data and not only data points as in standard approaches. As we will demonstrate later, this significantly aids prediction robustness and can be helpful in reducing data needs.

**2.3. EM Algorithm.** The log marginal evidence shown in Equation (2.6) has a complicated form (the expected value is inside the logarithm) and thus direct maximization is not straightforward. Such maximization, however, can be done indirectly by using the expectation-maximization algorithm, which we now proceed to explain. First note that we have the following lower bound for the log marginal evidence:

$$\begin{aligned} \log \left( \int p(\mathbf{y}, X | \theta, \Phi) dX \right) &= \log \left( \int q(X) \frac{p(\mathbf{y}, X | \theta, \Phi)}{q(X)} dX \right) \\ &\geq \int q(X) \log \left( \frac{p(\mathbf{y}, X | \theta, \Phi)}{q(X)} \right) dX \\ (2.7) \qquad &:= \mathcal{F}(q(\cdot), \theta). \end{aligned}$$

Here,  $q(X)$  represents an arbitrary distribution for  $X$ , and Jensen's inequality establishes the lower bound. The EM algorithm maximizes the lower bound  $\mathcal{F}(q(\cdot), \theta)$  with respect to the density function  $q(\cdot)$  and the hyperparameters  $\theta$ . This can be done using a coordinate ascent technique.

In the E-step we maximize  $F(q(\cdot), \theta)$  with respect to  $q(\cdot)$ , leaving  $\theta$  fixed to a current guess, and in the M-step we maximize  $F(q(\cdot), \theta)$  with respect to  $\theta$ , leaving  $q(\cdot)$  fixed to the updated value.

For candidate distribution  $q(\cdot)$ , the EM algorithm uses the posterior distribution  $p(X | \mathbf{y}, \theta, \Phi)$ . This is motivated by the relationship,

$$(2.8) \quad \mathcal{L} = \mathcal{F} + \mathcal{D}(q(X), p(X | \mathbf{y}, \theta, \Phi)).$$

Here,  $\mathcal{D}(q(X), p(X | \mathbf{y}, \theta, \Phi))$  is the Kullback-Leibler divergence that [measures the distance between distributions  \$q\(X\)\$  and  \$p\(X | \mathbf{y}, \theta, \Phi\)\$](#) . Note that because the divergence is a positive quantity, we have  $\mathcal{L} \geq \mathcal{F}$ . This relationship thus suggests that, with  $\theta$  fixed,  $\mathcal{L}$  is maximized by setting the distribution  $q(\cdot)$  equal to the posterior  $p(X | \mathbf{y}, \theta_k, \Phi)$ , where  $\theta_k$  is a current guess. We thus define the current distribution guess as  $q_k(\cdot) = p(X | \mathbf{y}, \theta_k, \Phi)$ . This gives the E-step, and  $\mathcal{F}(q_k(\cdot), \theta)$  can be written as

$$(2.9) \quad \mathcal{F}(q_k(\cdot), \theta) = \int p(X | \mathbf{y}, \theta_k, \Phi) \log \left( \frac{p(\mathbf{y}, X | \theta, \Phi)}{p(X | \mathbf{y}, \theta_k, \Phi)} \right) dX.$$

In the M-step, we now seek to maximize the lower bound  $\mathcal{F}(q_k(\cdot), \theta)$  with respect to  $\theta$ . We note that  $p(\mathbf{y}, X | \theta, \Phi) \propto p(\mathbf{y} | X, \theta)p(X | \Phi)$ . Consequently, [Equation \(2.9\)](#) becomes

$$\begin{aligned} \mathcal{F}(q_k(\cdot), \theta) &= \int p(X | \mathbf{y}, \theta_k, \Phi) \log \left( \frac{p(\mathbf{y} | X, \theta)p(X | \Phi)}{p(X | \mathbf{y}, \theta_k, \Phi)} \right) dX \\ &= \int p(X | \mathbf{y}, \theta_k, \Phi) \log \left( \frac{p(\mathbf{y} | X, \theta)p(X | \Phi)}{p(\mathbf{y} | X, \theta_k) p(X | \Phi)} \right) dX \\ &= \int p(X | \mathbf{y}, \theta_k, \Phi) \log \left( \frac{p(\mathbf{y} | X, \theta)}{p(\mathbf{y} | X, \theta_k)} \right) dX \\ &= \int p(X | \mathbf{y}, \theta_k, \Phi) \log p(\mathbf{y} | X, \theta) dX - \int p(X | \mathbf{y}, \theta_k, \Phi) \log p(\mathbf{y} | X, \theta_k) dX \\ (2.10) \quad &:= \mathcal{Q}(\theta, \theta_k) + C. \end{aligned}$$

Here, the second equality follows from Bayes theorem which states that

$$(2.11) \quad p(X | \mathbf{y}, \theta_k, \Phi) \propto p(\mathbf{y} | X, \theta_k) p(X | \Phi).$$

We also have that

$$(2.12a) \quad \mathcal{Q}(\theta, \theta_k) := \int p(X | \mathbf{y}, \theta_k, \Phi) \log p(\mathbf{y} | X, \theta) dX$$

$$(2.12b) \quad C := - \int p(X | \mathbf{y}, \theta_k, \Phi) \log p(\mathbf{y} | X, \theta_k) dX.$$

Since  $C$  is a constant independent of  $\theta$ , it is irrelevant in the maximization. The next guess for  $\theta$  is thus obtained from the maximization or M-step,

$$(2.13) \quad \theta_{k+1} \leftarrow \underset{\xi}{\operatorname{argmax}} \mathcal{Q}(\xi, \theta_k).$$

After this, the conditional distribution  $p(X | \mathbf{y}, \theta_k, \Phi)$  is updated in the E-step and the M-step (2.13) is repeated. This recursion corresponds to the exact EM algorithm.

### 3. MCEM ALGORITHM

The integral in Equation (2.12a) is analytically intractable but can be approximated by Monte Carlo sampling, which yields the MCEM algorithm. We obtain  $N$  data samples  $X_k(\omega), \omega = 1, \dots, N$  from  $p(X | \mathbf{y}, \theta_k, \Phi)$  or, equivalently, from  $p(\mathbf{y} | X, \theta_k) p(X | \Phi)$  to obtain an empirical approximation of the integral. We then compute an approximate M-step of the form

$$(3.14) \quad \theta_{k+1} \leftarrow \underset{\xi}{\operatorname{argmax}} \mathcal{Q}^N(\xi, \theta_k),$$

where

$$(3.15) \quad \mathcal{Q}^N(\xi, \theta_k) = \frac{1}{N} \sum_{\omega=1}^N \log p(\mathbf{y} | X_k(\omega), \xi).$$

We now summarize the MCEM algorithm. At iteration  $k$ , the E-step samples the estimated posterior distribution  $p(X | \mathbf{y}, \theta_k, \Phi) \propto p(\mathbf{y} | X, \theta_k) p(X | \Phi)$ , where the structure of the distributions is given by Equations (2.1) and (2.3). We can thus sample from  $p(X | \mathbf{y}, \theta_k, \Phi)$  using a Markov chain Monte Carlo (MCMC) algorithm because all that is needed is a routine that evaluates the target distribution  $p(\mathbf{y} | X, \theta_k) p(X | \Phi)$ . We define a method of the form  $\{X_k(\omega)\} \leftarrow \text{MCMC}(p(\mathbf{y} | X, \theta_k) p(X | \Phi), N)$ . Using these samples, we then obtain an updated value for the hyperparameters by solving Equation (3.14). The algorithm is presented below.

**input** : Data  $\mathbf{y}$ , statistics  $\Phi = \{\mathbf{m}^i, \mathbf{S}\}$ , number of MCMC samples  $N$ , tolerance  $\tau > 0$ , and initial guess  $\theta$ .

**output**: Estimates for hyperparameters  $\theta$  that maximize  $\mathcal{L}$ .

Set  $\epsilon \leftarrow 2 \cdot \tau$

**while**  $\epsilon \geq \tau$  **do**

- $\{X(\omega)\} \leftarrow \text{MCMC}(p(\mathbf{y} | X, \theta) p(X | \Phi), N)$
- $\theta' \leftarrow \theta$
- $\theta \leftarrow \underset{\xi}{\operatorname{argmax}} \mathcal{Q}^N(\xi, \theta)$
- $\epsilon \leftarrow \|\theta - \theta'\|_1$

**end**

We highlight that the input posterior  $p(X | \mathbf{y}, \theta_k, \Phi)$  does not have a practical form amenable for sampling. Therefore, the relationship (2.11) is important. This relationship indicates that the input posterior is proportional to the product of the output posterior  $p(\mathbf{y} | X, \theta_k)$  and the input distribution  $p(X | \Phi)$ . This enables the use of MCMC methods such as slice sampling, which requires knowledge only of the product function  $p(\mathbf{y} | X, \theta_k) p(X | \Phi)$  or the log form  $\log(p(\mathbf{y} | X, \theta_k)) + \log(p(X | \Phi))$  which in our case has a known form.

**3.1. Convergence.** The exact EM algorithm has a guaranteed improvement in the log marginal evidence  $\mathcal{L}$  at each iteration. This results from the fact that, in the E-step at iteration  $k$ , the lower bound is maximized with respect to  $\theta$  and, consequently, the log marginal evidence increases, having  $q(\cdot)$  fixed at  $p(X | \mathbf{y}, \theta_k, \Phi)$ . Once the hyperparameters are updated in the M-step to  $\theta_{k+1}$ , the Kullback-Leibler divergence becomes nonzero, because the distribution in the previous iteration

was  $q_k(\cdot) = p(X | \mathbf{y}, \theta_k, \Phi)$  and not  $p(X | \mathbf{y}, \theta_{k+1}, \Phi)$ , which maximizes the log marginal evidence with respect to  $q(\cdot)$ . Consequently, the increase in the log marginal evidence will be larger than the increase in the lower bound.

The transition to MCEM sacrifices the guaranteed increase in the log marginal evidence because the M-step cannot be performed exactly. Numerous papers, however, establish theoretical convergence results for MCEM algorithms. Fort and Moulines establish convergence but require that the number of MCMC samples at each iteration increases without bound [6]. Booth and Hobert assert that, even as the sample size increases, the Monte Carlo error remains in the estimate [2]. Delyon, Lavielle, and Moulines proposed a stochastic approximation EM algorithm that replaces the MC update with a weighted average of the current and previous  $\mathcal{Q}(\cdot)$  functions [4]; in this way they obtain convergence without increasing the number of samples. In the following section, we explore the empirical behavior of the MCEM method as the number of samples increases.

We highlight that, even in the exact EM algorithm, one cannot to monitor stationarity of  $\mathcal{L}$  because of its complicated form. Therefore, we stop the algorithm once the change in the hyperparameters is below a tolerance.

**3.2. Computational Details.** The explicit form of function  $\mathcal{Q}^N(\theta, \theta_0)$  is given by

$$\begin{aligned} \mathcal{Q}^N(\theta, \theta_k) &= \frac{1}{N} \sum_{\omega=1}^N \log p(\mathbf{y} | X_k(\omega), \theta) \\ (3.16) \quad &= \frac{1}{N} \sum_{\omega=1}^N \left( -\frac{n}{2} \log(2\pi) - \frac{1}{2} \log |\mathbf{K}(\theta, X_k(\omega), X_k(\omega))| - \frac{1}{2} \mathbf{y}^\top \mathbf{K}(\theta, X_k(\omega), X_k(\omega))^{-1} \mathbf{y} \right). \end{aligned}$$

Each term in the summation is the log likelihood function maximized in standard GP [20]. The first and second derivatives of this function are well known [29]. In this work, we use the trust-region Newton method implemented in Matlab [16]. The gradient is provided in explicit form and second derivatives are estimated by using finite differences. We use the slice sample procedure implemented in Matlab to compute the MCMC samples.

Once hyperparameters have been trained, it is desired to make predictions at a set of test data points. For prediction at a test point  $X_t$ , we propose to use a predictive distribution with mean and covariance,

$$(3.17a) \quad \bar{\mu}(\theta, X, X_t) := \mathbf{K}(\theta, X, X_t) \mathbf{K}(\theta, X, X)^{-1} \mathbf{y},$$

$$(3.17b) \quad \bar{\mathbf{K}}(\theta, X, X_t) := \mathbf{K}(\theta, X_t, X_t) - \mathbf{K}(\theta, X, X_t) \mathbf{K}(\theta, X, X)^{-1} \mathbf{K}(\theta, X, X_t).$$

Here,  $X$  and  $X_t$  contain only the mean of the uncertain inputs (i.e., we assume zero variance in the input data). We emphasize that this *predictive* distribution does not consider the effect of input data uncertainty. In other words, our approach considers input uncertainty *only in the training phase*. As we demonstrate in the next section, this already gives significant improvements in prediction performance because it robustifies the estimates of the hyperparameters. Predictive distributions that consider input data uncertainty are much more elaborate and are presented in [19, 18]. We leave this as a topic of future research.

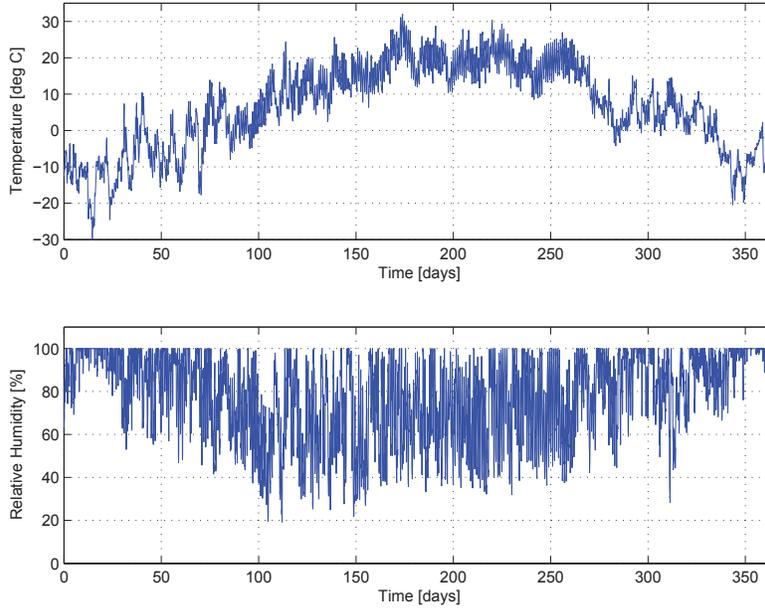


FIGURE 1. Hourly ambient temperature and relative humidity profiles for Chicago area, 2006.

#### 4. CASE STUDY

In this section, we present a case study to demonstrate that the GP-MCEM approach yields much more robust predictions compared to standard GP approaches. In addition, we discuss numerical performance.

**4.1. Prediction Performance.** We consider a simulated study in which an advanced model predictive control (MPC) system is used to optimize the operation of the HVAC system of an office building. Here, we seek to baseline the energy performance of the MPC system. The office building comprises a well-mixed single space conditioned by an air-handling unit (AHU). Dynamic models of temperature, species concentration (moisture and carbon dioxide), and pressure are used to simulate the comfort and air quality conditions of the building space. The MPC system uses all the available degrees of freedom of the HVAC system (heating and cooling for AHU, moisture removal, ambient, recycle and exhaust flows) to minimize energy while satisfying comfort and air quality constraints. This optimization is done adaptively as external variables (occupancy and weather conditions) change in time. A detailed formulation of the MPC controller and a description of the building model are presented in [27, 26, 10]. The MPC controller introduces a high degree of coupling between the building control variables, external variables, and the total energy demand; and it induces nonlinear behavior. This makes the construction of baseline models complicated because energy performance becomes a strong function of multiple input variables such as occupancy and ambient conditions.

The mathematical formulation of the MPC controller has the form

$$\begin{aligned}
(4.18a) \quad & \min_{\mathbf{u}(\cdot)} \int_t^{t+T} y(\tau) d\tau \\
(4.18b) \quad & \text{s.t.} \quad \frac{d\mathbf{x}}{dt}(\tau) = f_x(\mathbf{x}(\tau), \mathbf{z}(\tau), \mathbf{u}(\tau), \omega(\tau)) \\
(4.18c) \quad & 0 = f_z(\mathbf{x}(\tau), \mathbf{z}(\tau), \mathbf{u}(\tau), \omega(\tau)) \\
(4.18d) \quad & 0 \leq g(\mathbf{x}(\tau), \mathbf{z}(\tau), \mathbf{u}(\tau), \omega(\tau)) \\
(4.18e) \quad & \chi(\tau) = \Xi(\mathbf{x}(\tau), \mathbf{z}(\tau), \mathbf{u}(\tau)) \\
(4.18f) \quad & \mathbf{x}(t) = \text{given}, \quad \tau \in [t, t+T].
\end{aligned}$$

Here,  $t$  is the current time,  $T$  is the prediction horizon,  $\mathbf{x}(\cdot)$  are the dynamic states of the building (zone temperature, CO<sub>2</sub> and H<sub>2</sub>O concentrations, building air mass),  $\mathbf{z}(\cdot)$  are the algebraic states (relative humidity, supply air temperature),  $\mathbf{u}(\cdot)$  are the degrees of freedom or controls (AHU electrical energy, ambient air flow, recycle flow), and  $\omega(\tau)$  are external variables such as occupancy heat and CO<sub>2</sub>/H<sub>2</sub>O loads and ambient conditions (CO<sub>2</sub>, temperature, and relative humidity). The variable  $\chi(\tau)$  represents the total HVAC electrical energy in the HVAC system.

The performance of the MPC controller (4.18) is simulated by running the system using a receding horizon of 24 hours with time steps of one hour for an entire year. We used real weather data in the Chicago area to perform the year-long simulations (see Figure 1). The total daily HVAC energy is used as the output, and we consider two certain input variables (average daily relative humidity and average ambient temperature) and an uncertain input variable (average daily occupancy). For the certain inputs we consider the values presented in Figure 1, and we assume zero variance. For the uncertain input we consider two cases. The first case assumes that the average daily occupancy varies *mildly* as  $\mathcal{N}(500, 50^2)$ . The second case assumes that occupancy varies *strongly* as  $\mathcal{N}(500, 100^2)$ . For each case, we generate a *training data set* and a *validation (test) data set* using two different samples from the corresponding distributions. The *true occupancy values* are the samples obtained from these distributions. Note that both the training and validation data sets use the same weather information; occupancy is the only variable that is assumed uncertain.

We compare the performance of three different GP strategies. The first strategy is the *ideal-GP* strategy in which we assume that occupancy is known with full certainty, and this is used for training the model. The second strategy is the *standard-GP* strategy in which we neglect the presence of occupancy variations; this is equivalent to assuming that occupancy remains constant between days and, consequently, is not used for training the model. The third strategy is the *GP-MCEM* strategy in which we train the model by capturing occupancy variations.

We first demonstrate that GP modeling can be used to accurately predict the energy outputs of the MPC controller given three different input variables. We use the ideal-GP strategy in which we use true occupancy values of the training sample for training. We then test the predictions of the model using the validation data set.

In the left panel of Figure 2 we present the fit and the 95% confidence intervals for the training set obtained with the ideal-GP approach. The fit is accurate, demonstrating that GP is able to capture the multivariable interactions in an effective way. This is a remarkable feature given the complex structure of the building model and control system used. Also note that the confidence intervals

are narrow implying that the weather and occupancy data are sufficient to explain energy behavior. In the right panel we present the GP predictions (using the hyperparameters obtained from the training set) obtained for the validation set. Surprisingly, we observe significant instability at certain points, reflected as large confidence intervals. In Figure 3 we present the fit and prediction for the ideal-GP case with high occupancy variability. Note that the higher variability leads to energy demands outside the main trend but the fit of the GP model remains highly accurate. This further illustrates the ability of GP modeling to capture multivariable interactions. Note, however, that the instability of the predictions in the validation points increases significantly and large amounts of data might be needed to eliminate such instability.

The high sensitivity of the model to perturbations in input occupancy data can be explained by the fact that the training data set did not fully cover the variable range. Consequently, it does not capture the test points in the validation set. This is troubling from a practical point of view because it implies that significant uncertainty will be observed if the M&V baseline model is used to predict energy performance (say, next year) at days with similar weather conditions but slightly different occupancy conditions. The modeler will then be tempted to add those days to the training set to mitigate uncertainty. Continuous data collection will proceed until the predictions stabilize; but they will always be vulnerable to occupancy levels outside the collected range, and large amounts of data will be required. We can thus conclude that the use of “perfect” information does yield a highly accurate training model but this comes at the expense of significant vulnerability to even slight data perturbations.

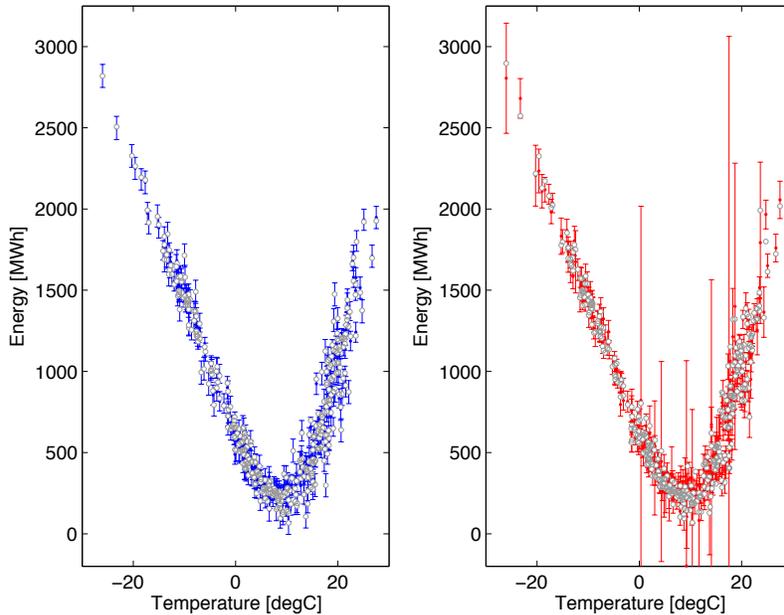


FIGURE 2. Ideal-GP case for mild occupancy variation. Fit of training set (left) and prediction at validation set (right).

In Figure 4 we compare the training performance of the GP-MCEM approach with that of the standard-GP approach. For the MCEM approach, we used a total of 1000 samples in the E-step.

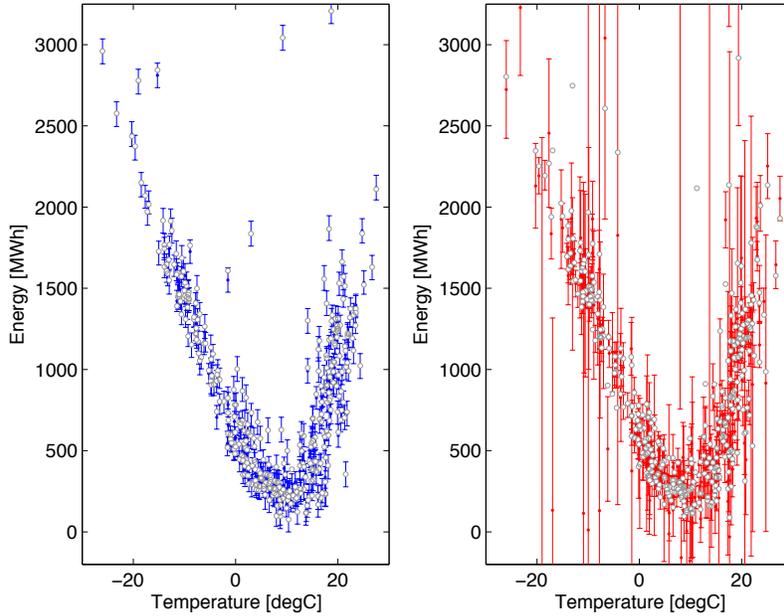


FIGURE 3. Ideal-GP case for high occupancy variation. Fit of training set (left) and prediction at validation set (right).

As can be seen, the confidence levels of the standard-GP approach are significantly larger than those of the MCEM approach. The reason is that the standard approach treats the variability in the outputs as noise when in fact a hidden input variable is responsible for the variation. In Figure 5 we compare the performance of both approaches in the validation set. Figures 6 and 7 compare performance of standard-GP and GP-MCEM for the case of high occupancy variation. We can see that the confidence intervals of the standard-GP approach reach impractical levels. *The confidence intervals and predictions of the MCEM approach, however, remain robust even in the presence of high variability.*

The high volatility of the confidence levels observed in the ideal-GP approach is eliminated with the GP-MCEM approach. For the mild variation case, we see that neglecting occupancy information (standard-GP approach) yields even less volatility than does the ideal-GP approach. The performance of standard-GP, however, is not competitive as we increase variability. These results raise an interesting property from a data collection standpoint. The GP-MCEM can *effectively reduce data needs by capturing variations in input data because this effectively expands the variable range used for training.* Consequently, even if occupancy data were available, we can use the proposed approach to robustify training. In addition, the GP-MCEM approach provides a systematic framework to assess the effect of input data variations for different variables on prediction stability.

We highlight that the proposed GP-MCEM approach can also be used to construct models when limited and/or sparse measurement data are available. For instance, consider the case in which a sensor fails frequently or is simply too expensive to install. In this case, we can construct a GP model by providing a typical range for such measurement (e.g., a uniform distribution). Consequently, the proposed approach can be seen as a mechanism to reduce data collection and sensor costs of M&V processes.

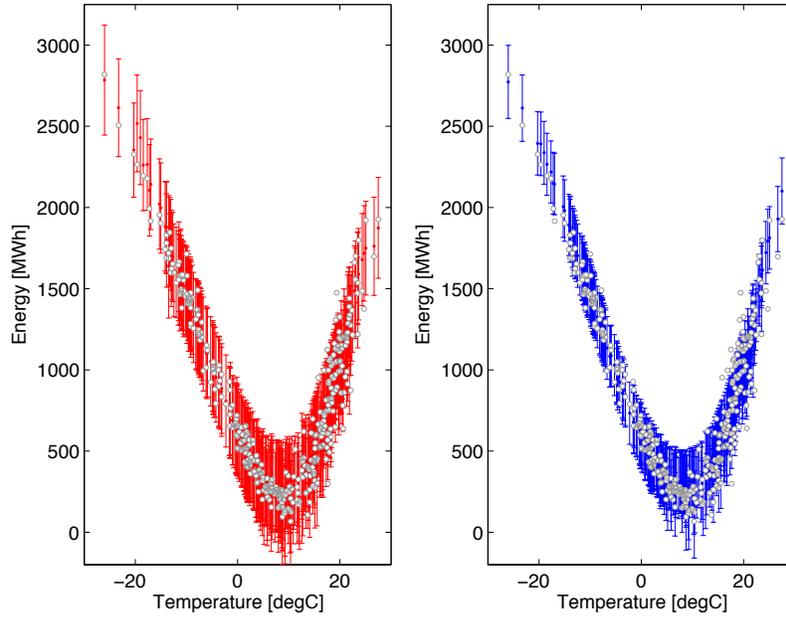


FIGURE 4. Fit of standard-GP (left) and GP-MCEM (right) approaches for mild occupancy variation (training set).

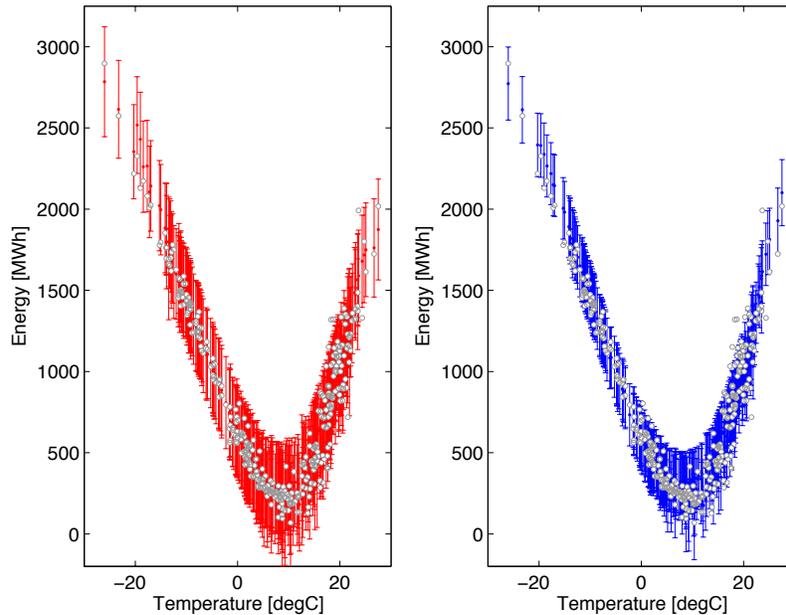


FIGURE 5. Prediction of standard-GP (left) and GP-MCEM (right) approaches for mild occupancy variation (validation set).

**4.2. Numerical Behavior.** We compare the performance of the GP-MCEM algorithm as a function of the number of samples used in the E-step. We performed three replications for every number of samples tried. The results are presented in Table 1. As can be seen, the number of iterations taken by the GP-MCEM algorithm stabilizes as the number of samples increases. The variability

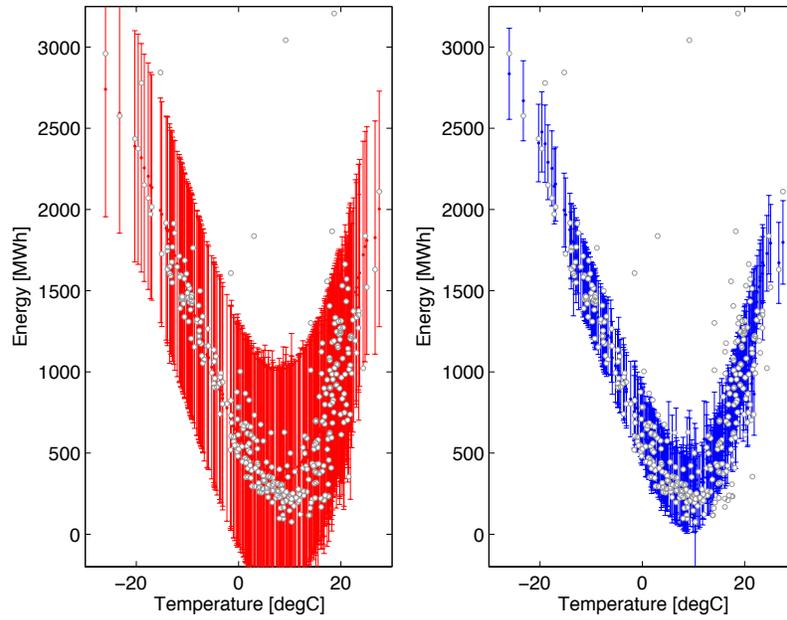


FIGURE 6. Fit of standard-GP (left) and GP-MCEM(right) approaches for high occupancy variation (training set).

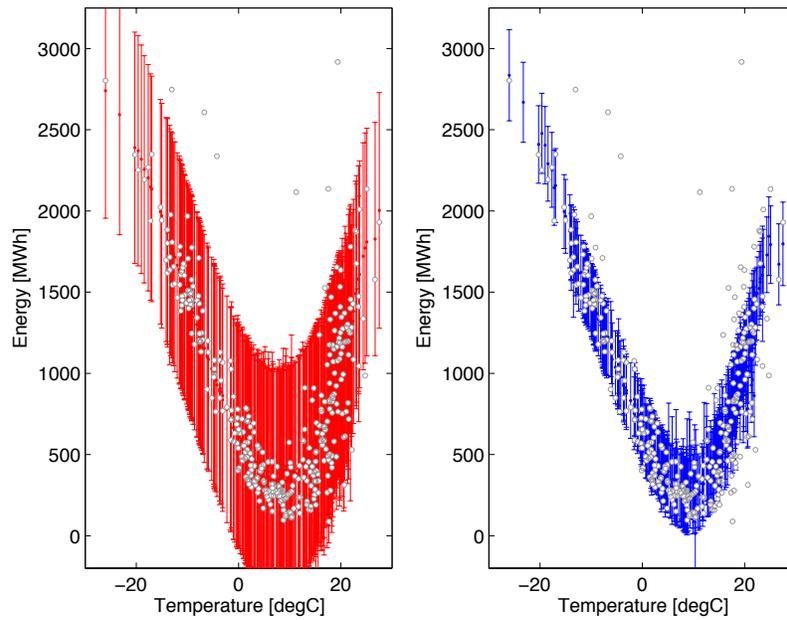


FIGURE 7. Prediction of standard-GP (left) and GP-MCEM (right) approaches for high occupancy variation (validation set).

of the hyperparameter estimates tends to decrease as well. Some persistent error exists, however, even in the case with 1,000 samples. Note that this is a general deficiency of stochastic algorithms including the one implemented here. For this class of algorithms, convergence can be guaranteed only asymptotically, or one must resort to empirical tests of convergence. Such tests are difficult

TABLE 1. Numerical behavior of GP approaches.

<b>Approach</b>	$N$	<b>Iter</b>	$\log \theta_0$	$\log \theta_1$	$\log \theta_2$
Ideal-GP	-	-	3.206	7.453	3.290
Standard-GP	-	-	2.899	7.245	4.906
	40	56	3.563	7.925	4.531
GP-MCEM	40	4	3.284	7.480	4.452
	40	9	3.447	7.653	4.461
	200	40	3.658	8.127	4.263
GP-MCEM	200	12	3.795	8.480	4.313
	200	12	3.725	8.309	4.340
	300	9	3.881	8.592	4.272
GP-MCEM	300	11	3.838	8.440	4.312
	300	10	3.613	8.016	4.220
	1000	10	3.872	8.576	4.212
GP-MCEM	1000	9	3.898	8.644	4.089
	1000	9	3.863	8.535	4.193

to implement in the EM context because one cannot easily construct estimators for the log marginal evidence shown in Equation (2.6). In practice, however, we have not observed significant performance differences in terms of training fit as we vary the number of samples. Consequently, one might consider terminating the algorithm once the fit of the training set stabilizes for a given number of replicates. We note, however, that this approach is adhoc and thus more research is needed on appropriate termination tests.

From Table 1 we can also observe that the output noise parameter  $\theta_2$  is significantly overestimated in the standard-GP approach compared with the other approaches. For the standard approach the logarithmic value is 4.906 while for the GP-MCEM approach the value is around 4.29. At real scale, the former corresponds to 135.1 and the latter 73.1. This is a difference of nearly 100%, manifested in much wider confidence levels.

## 5. CONCLUSIONS AND POTENTIAL EXTENSIONS

Accounting for input data uncertainty permits a much more realistic and robust approach to Gaussian process modeling. In particular, we have demonstrated that significant volatility in predictions can result even in the ideal case where perfect input data is assumed for training. Because the standard GP method ignores uncertainty in input data, the accuracy of energy savings estimates rely more strongly on the quality of data collected for M&V (e.g., number of measured variables, frequency of measurement, accuracy of sensors), which can substantially increase M&V expenses. To overcome this limitation, we proposed a GP framework coupled to a Monte Carlo expectation maximization algorithm (GP-MCEM) to train the GP model under input data uncertainty. We have demonstrated that the approach can significantly increase robustness and mitigate volatility in the predictions. The proposed approach can also be used to construct models when limited and sparse measurement data is available (e.g., by replacing sensor readings with crude estimates and

providing ranges). This can in turn reduce data collection and sensor deployment costs of M&V processes.

Several opportunities for future research promise to build upon the demonstrated successes, particularly in the prediction phase in which we explicitly account for input uncertainty [7, 19]. This paper focused on demonstrating the benefits of the GP-MCEM framework through an illustrative case study in which only occupancy uncertainty was considered. From a pragmatic standpoint, the current GP-MCEM framework is capable of accounting for uncertainty in multiple input variables as long as they are uncorrelated. For variables exhibiting correlations such as weather variables (outdoor temperature, solar radiation, relative humidity), the current framework will tend to over-estimate uncertainty. Consequently, extensions are needed to deal with correlated variables; in particular, more efficient MCMC procedures are needed that are capable of sampling from significantly more complex input distributions in a computationally efficient manner. From an algorithmic standpoint, an ascent-based MCEM method capable of recovering the ascent property of EM with high probability can significantly aid numerical performance [3, 13, 23, 2]. In addition, methods are needed that exploit parallelism and enable matrix-free settings, in order to accelerate computational performance [17].

#### ACKNOWLEDGMENTS

This work was supported by the U.S. Department of Energy, Office of Science, under Contract DE-AC02-06CH11357.

#### REFERENCES

- [1] ASHRAE, *Ashrae guideline 14: Measurement of energy and demand savings*, American Society of Heating, Refrigerating, and Air-Conditioning Engineers, Inc., Atlanta, GA, 2002.
- [2] J. G. Booth and J. P. Hobert, *Maximizing general linear mixed model likelihoods with an automated Monte Carlo EM algorithm*, J. Roy. Statist. Soc. Ser. B Stat. Methodol. **61** (1999), no. 1, 265–285.
- [3] B. S. Caffo, W. Jank, and G. L. Jones, *Ascent-based Monte Carlo expectation-maximization*, J. Roy. Statist. Soc. Ser. B Stat. Methodol. **67** (2005), no. 2, 235–251.
- [4] B. Delyon, M. Lavielle, and E. Moulines, *Convergence of a stochastic approximation version of the EM algorithm*, Ann. Statist. **27** (1999), no. 1, 235–251.
- [5] A. P. Dempster, N. M. Laird, and D. B. Rubin, *Maximum likelihood from incomplete data via the EM algorithm*, J. Roy. Statist. Soc. Ser. B Stat. Methodol. **39** (1999), no. 1, 1–38.
- [6] G. Fort and E. Moulines, *Convergence of the Monte Carlo Expectation Maximization for curved exponential families*, Ann. Statist. **31** (2003), no. 4, 1220–1259.
- [7] A. Girard, C. E. Rasmussen, and R. Murray-Smith, *Gaussian process priors with uncertain inputs: Multiple-step ahead prediction*, Tech. Report TR-2002-119, Department of Computing Science, Glasgow University, Glasgow, Scotland, 2002.
- [8] Y. Heo, D. Graziano, V.M. Zavala, P. Dickinson, M. Kamrath, and M. Kirshenbaum, *Cost-effective measurement and verification method for determining energy savings under uncertainty*, Proceedings of ASHRAE Annual Conference (2013).
- [9] Y. Heo, Choudhary R., and Augenbroe G. A., *Calibration of building energy models for retrofit analysis under uncertainty*, Energy and Buildings **47** (2012), 550–560.
- [10] Y. Heo and V. M. Zavala, *Gaussian process modeling for measurement and verification*, Energy & Buildings **53** (2012), 7–18.

- [11] P. Hoes, J. L. M. Hensen, M. G. L. C. Loomans, B. de Vries, and D. Bourgeois, *User behaviour in whole building simulation*, *Energy and Buildings* **41** (2009), no. 3, 295–302.
- [12] IPMVP, *International performance measurement and verification protocol: Concepts and options for determining energy and water savings, volume 1*, Efficiency Valuation Organization, 2010.
- [13] W. Jank, *Implementing and diagnosing the stochastic approximation EM algorithm*, *J. Comput. Graph. Statist.* **15** (2006), no. 4, 803–829.
- [14] Y. J. Kim, K. U. Ahn, C. S. Park, and I. H. Kim, *Gaussian emulator for stochastic optimal design of a double glazing system*, Proceedings of 13th International Building Performance Simulation Association Conference, 2013.
- [15] I. P. Knight and G. N. Dunn, *Evaluation of heat gains in UK office environments*, Worldwide CIBSE/ASHRAE Gathering of the Building Services Industry, Edinburgh, Scotland, 2003.
- [16] Matlab, *Optimization toolbox users guide R2013b*, The MathWorks Inc., 2013.
- [17] A. Mihai, J. Chen, and L. Wang, *A matrix-free approach for solving the parametric Gaussian process maximum likelihood problem*, *SIAM Journal on Scientific Computing* **34** (2012), no. 1, A240–A262.
- [18] J. Quiñonero-Candela, *Learning with uncertainty – gaussian processes and relevance vector machines*, Ph.D. thesis, Technical University of Denmark, 2004.
- [19] J. Quiñonero-Candela, A. Girard, J. Larsen, and C. E. Rasmussen, *Propagation of uncertainty in bayesian kernels models - application to multiple-step ahead forecasting*, International Conference on Acoustics, Speech and Signal Processing **2** (2003).
- [20] C. E. Rasmussen and C. K. Williams, *Gaussian processes for machine learning*, MIT Press, Cambridge, MA, 2006.
- [21] Y. Sun and G. Augenbroe, *Modeling the urban heat island effect and assessing the impact on building energy consumption*, Proceedings of the 2nd International Conference on Building Energy and Environment, Boulder, 2012.
- [22] Y. Sun, Y. Heo, H. Xie, M. Tan, J. Wu, and G. Augenbroe, *Uncertainty quantification of microclimate variables in building energy simulation*, *Journal of Building Performance Simulation* (2013).
- [23] G. C. G. Wei and M. A. Tanner, *A Monte Carlo implementation of the EM algorithm and the poor man’s data augmentation algorithms*, *J. Amer. Statist. Assoc.* **85** (1990), no. 411, 699–704.
- [24] B. Yan and A. M. Malkawi, *A Bayesian approach for predicting building cooling and heating consumption*, Proceedings of 13th International Building Performance Simulation Association Conference, 2013.
- [25] J. E. Yan, Y. J. Kim, K. U. Ahn, and C. S. Park, *Gaussian process emulator for optimal operation of a high rise office building*, Proceedings of 13th International Building Performance Simulation Association Conference, 2013.
- [26] V. M. Zavala, *Real-time optimization strategies for building systems*, *Industrial & Engineering Chemistry Research* **52** (2012), no. 9, 3137–3150.
- [27] V. M. Zavala, *Real-time resolution of conflicting objectives in building energy management: An utopia-tracking approach*, In Proceedings of SimBuild (2012).
- [28] V. M. Zavala, *Inference of building occupancy signals using moving horizon estimation and fourier regularization*, *Journal of Process Control* **In Press** (2013).
- [29] Y. Zhang and W. E. Leithead, *Exploiting Hessian matrix and trust-region algorithm in hyperparameters estimation of Gaussian process*, *Appl. Math. Comput.* **171** (2005), no. 2, 1264–1281.
- [30] Y. Zhang, Z. O’Neil, T. Wagner, and G. Augenbroe, *An inverse model with uncertainty quantification to estimate the energy performance of an office building*, Proceedings of 13th International Building Performance Simulation Association Conference, 2013.

The submitted manuscript has been created by UChicago Argonne, LLC, Operator of Argonne National Laboratory ("Argonne"). Argonne, a U.S. Department of Energy Office of Science laboratory, is operated under Contract No. DE-AC02-06CH11357. The U.S. Government retains for itself, and others acting on its behalf, a paid-up, nonexclusive, irrevocable worldwide license in said article to reproduce, prepare derivative works, distribute copies to the public, and perform publicly and display publicly, by or on behalf of the Government.