Uncertainty Quantification for Model Parameters and Hidden States in Bayesian Dynamic Linear Models

Luong Ha Nguyen*, Ianis Gaudot, and James-A. Goulet
Department of Civil, Geologic and Mining Engineering
Ecole Polytechnique de Montreal, CANADA

October 2, 2018

Abstract

The quantification of uncertainty associated with the model parameters and the hidden state variables is a key missing aspect for the existing Bayesian Dynamic Linear Models. This paper proposes two procedures for carrying out the uncertainty quantification task: (1) the Maximum a Posteriori with the Laplace approximation procedure (LAP-P) and (2) the Hamiltonian Monte Carlo procedure (HMC-P). A comparative study of LAP-P with HMC-P is conducted on simulated data as well as real data collected on a dam in Canada. The results show that the LAP-P is capable to provide a reasonable estimation without requiring a high computation cost, yet it is prone to be trapped in local maxima. The HMC-P yields a more reliable estimation than LAP-P, but it is computationally demanding. The estimation results obtained from both LAP-P and HMC-P tend to the same values as the size of the training data increases. Therefore, a deployment of both LAP-P and HMC-P is suggested for ensuring an efficient and reliable estimation. LAP-P should first be employed for the model development and HMC-P should then be used to verify the estimation obtained using LAP-P.

Keywords: Uncertainty, Bayesian, Dynamic Linear Models, Kalmal filter, Structural Health Monitoring, Dam.

1 Introduction

Bayesian Dynamic Linear Models (BDLMs) are a class of *state-space models* (SSMs) which are well suited for sequential inference [40]. BDLMs rely on a transition model which is used to predict recursively future hidden state variables based on current hidden state variable. At each step, the predictions are updated with new observations. In BDLMs, the hidden state variables and observations are assumed to be Gaussian random variables, and the transition and observation model are linear. The analytical solutions for the prediction and update step are available through the Kalman filter equations. Following the Bayesian framework, the hidden state variables are described by a posterior *probability*

*Corresponding author: luongha.nguyen@gmail.com

density function (PDF) that combines information from a prediction (prior PDF) and from observations through the likelihood function. In BDLMs, the posterior PDF is fully described by its mean vector and covariance matrix, where the variance quantifies the uncertainty about each hidden state variable. The accurate quantification of uncertainty associated with state estimation is important because the primary objective of BDLMs is to serve decision-making [42].

In this paper, we are interested in quantifying the effect of model parameter uncertainties on hidden state uncertainties, in the context of the use of BLDMs for the long-term monitoring of civil infrastructure. The BDLM formulation has been introduced in [15], where the method is used to decompose the observed time-series related to the response of the civil structure (e.g. displacements or frequencies) into a set of hidden state variables. Recent applications have illustrated the potential of BDLMs to track time-varying baseline responses of civil structures from datasets containing multiple observations [16], and to detect anomalies [33]. BDLMs typically depend on several unknown parameters that need to be inferred from data. Previous studies employed the *Maximum Log-likelihood Estimation* (MLE) to infer the optimal model parameters from observations [10].

The MLE method suffers from the well-known drawbacks [25] of point estimation methods: (a) it ignores model parameter uncertainties, (b) it does not guarantee to converge towards the global maximum, and (c) its performance is sensitive to the choice of initial parameter values. These reasons currently limit the general applications of BDLMs for the structural health monitoring.

In SSMs, the classic point estimation method is the Expectation-Maximization (EM) [8], which is based on the maximization of the likelihood function obtained by analytically setting its derivative equals to zero [38]. The point estimation method for model parameters is justified by the asymptotic distributional convergence [34], which tells that the MLE converges to the true values as the training dataset length tends to infinity [10]. In practical situation, however, the number of observations available to compute model parameters is often limited, and it is difficult to know whether the asymptotic condition is met, or not. More importantly, there are some problems where the asymptotic rule does not hold, that is, when model parameters are not identifiable because there is no global maximum in the likelihood function [1, 2, 6]. For instance, such a situation arises when model parameters are strongly correlated with each other. In other cases, when likelihood function exhibits a surface that is not unimodal, the MLE can converge towards a local maximum instead of the global maximum, depending on the starting value of the model parameters. The *Maximum A Posteriori* (MAP) using a gradient-based maximization algorithm is an alternative to the MLE method, which allows us to account for our prior knowledge of model parameters [37]. Like the MLE, the MAP estimate corresponds to a Dirac delta function approximation to the posterior PDF for model parameters. The MLE can be considered as a special case of MAP when we employ a uniform prior PDF. The uncertainties around the MLE and MAP results can be computed using the Laplace approximation [10]. The Laplace approximation assumes that the target function around the estimate is close to a Gaussian, which may not be the case for a wide range of problems, as mentioned earlier.

Alternatively, Monte Carlo sampling approaches can be used to directly sample from the posterior PDF of model parameters. In contrast to MLE and MAP methods, which explores only a small portion of the target function, Monte-Carlo sampling techniques

have the potential to perform a global exploration. Markov chain Monte Carlo (MCMC) is a broad class of Monte Carlo sampling techniques which perform a random walk in the model parameter space, where each sample depends only on the previous one. The resulting Markov chain statistically converges towards a unique stationary distribution which is the required target function [5]. MCMC techniques are intrinsically Bayesian because they provide an approximation of the posterior PDF of model parameters rather than a point estimation. The use of MCMC methods have remained limited until the 90's due to their high computational costs. In recent years, the advances in computer performance, particularly in parallel computing [36], have led the MCMC approaches to be more popular in many domains of applications, including for model parameters inference in SSMs [14, 21, 34]. MCMC approaches are particularly attractive for practical applications because they allow quantifying the model parameter posterior PDF even if a small training dataset is available, or when the asymptotic rule does not hold [34]. In civil engineering, MCMC approaches have been widely used for Bayesian model updating [7, 23] and for model class selection [2].

The random walk Metropolis-Hasting (MH) [27, 28] is one of the most common MCMC algorithms. MH relies on a proposal distribution to transit from a current sample to the following one. The proposed sample is then accepted or rejected according to an acceptance probability [19]. MH suffers from some drawbacks, such as the tuning of the proposal distribution. In particular, in the case of non-Gaussian posterior, high correlations between the model parameters can substantially slow down the convergence speed [35]. More sophisticated MCMC methods have been developed to improve sampling performance. Adaptive MCMC techniques were developed to automatically adjust the proposal during the MCMC run [18]. Another approach consists in taking advantage of the gradient information for the target function to propose new samples [9]. This approach, known as *Hamiltonian Monte Carlo* (HMC) [3, 31], has been recently used in a wide range of applications including physics [9], biology [22], engineering [7, 17]. Few studies have investigated the use of HMC for model parameters inference in SSMs [26, 41].

This paper first proposes both a MAP approach coupled with a Laplace approximation, and an HMC-based method to approximate the posterior PDF of model parameters for the BDLMs. Secondly, we propose a Gaussian mixture approach to propagate the model parameters uncertainties in the hidden state variable estimates. The performance of the MAP coupled with a Laplace approximation is tested against the results obtained using an HMC-based method which serves as Benchmark. The results obtained using the two algorithms are compared as a function of the training dataset length.

The paper is organized as follows. The first section presents the state estimation theory for BDLMs. We then describe the MAP with the Laplace approximation and the HMC method applied to BDLMs. In a third section, we present the details of the Gaussian mixture approach to propagate the model parameter uncertainties on state estimation uncertainties . The fourth section presents a general procedure including the theories presented in Sections 2 & 3. In the fifth section, we illustrate the results obtained using the proposed approaches on simulated data as well as real displacement data measured on a dam in Canada. Finally, we discuss the main features of the approaches, reviewing their advantages and drawbacks.

2 Bayesian Dynamic Linear Models

This section presents the Bayesian dynamic linear models (BDLM), which is a special case of State-space models (SSMs). A BDLM consists of two linear models defined by an *observation* model and a *transition* model. The observation model is employed to describe the relation between data $\mathbf{y}_t$ and hidden state variables $\mathbf{x}_t$ at time $t \in [1:\mathtt{T}]$. The transition model describes the dynamics of the hidden states variables over time. The mathematical formulations for both models are defined as

Observation model

$$\mathbf{y}_t = \mathbf{C}_t\mathbf{x}_t + \mathbf{v}_t, \quad \begin{cases} \mathbf{y}_t \sim \mathcal{N}(\mathbb{E}[\mathbf{y}_t], \mathrm{cov}[\mathbf{y}_t]) \\ \mathbf{x}_t \sim \mathcal{N}(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_{\boldsymbol{t}}) \\ \mathbf{v}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{R}_{\mathbf{t}}) \end{cases} \tag{1}$$

Transition model

$$\mathbf{x}_t = \mathbf{A}_t\mathbf{x}_{t-1} + \mathbf{w}_t, \quad \left\{ \ \mathbf{w}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{Q}_{\mathbf{t}}), \right.$$

where $\mathbf{C}_t$ is the observation matrix, $\mathbf{v}_t$ is the Gaussian observation error with zero-mean and covariance matrix $\mathbf{R}_t$, $\mathbf{A}_t$ is the transition matrix, and $\mathbf{w}_t$ is the Gaussian model error with zero-mean and covariance matrix $\mathbf{Q}_t$.

The theory behind the BDLM is that the observation $\mathbf{y}_t$ is modeled by a vector of hidden state variables $\mathbf{x}_t$. In common cases, the vector of hidden state variables describes the baseline, periodic, and stochastic behavior of a system's response. Typically, a baseline is used to describe the structural behavior over time without external effects such as temperature and loading, a periodic component is employed to model the external effects, and an autoregressive component is used to capture the time-dependent model errors. The rate of change of the baseline component over time is described by a trend component. Like any other SSMs, the BDLM also employs either Kalman filter [30] or the UD filter [39] to estimate the hidden state variables $\mathbf{x}_t$ given the model parameter vector $\boldsymbol{\theta}$ and the available information $\mathbf{y}_{1:t}$,

$$p(\mathbf{x}_t|\boldsymbol{\theta}, \mathbf{y}_{1:t}) = \mathcal{N}\left(\mathbf{x}_t; \boldsymbol{\mu}_{t|t}, \boldsymbol{\Sigma}_{t|t}\right), \tag{2}$$

where the posterior expected value $\boldsymbol{\mu}_{t|t} \equiv \mathbb{E}\left[\mathbf{x}_t|\mathbf{y}_{1:t}\right]$ and posterior covariance $\boldsymbol{\Sigma}_{t|t} \equiv \mathrm{cov}\left[\mathbf{x}_t|\mathbf{y}_{1:t}\right]$ are obtained from the measurement step in the Kalman filter. The reduced form of the Kalman filter or UD filter can be written as

$$\left(\boldsymbol{\mu}_{t|t}, \boldsymbol{\Sigma}_{t|t}\right) = \mathrm{filter}\left(\boldsymbol{\mu}_{t-1|t-1}, \boldsymbol{\Sigma}_{t-1|t-1}, \mathbf{y}_t; \mathbf{A}_t, \mathbf{C}_t, \mathbf{Q}, \mathbf{R}_t\right). \tag{3}$$

Note that the UD filter yields numerically more stable performance than the Kalman filter yet, it is slightly more computational demanding. The full mathematical formulations associated with the hidden state variables as well as with the Kalman filter are described in well details by West & Harrison [40] and Goulet [15].

3 Parameter Estimation

This section presents the details of two methods for approximating the posterior PDF of model parameters: *Maximum A Posteriori with Laplace approximation* and *Hamiltonian*

Monte Carlo.

3.1 Maximum A Posteriori with Laplace Approximation

Maximum a posteriori consists in identifying the point estimates by maximizing the posterior PDF defined as

$$\begin{aligned} p(\boldsymbol{\theta}|\mathbf{y}_{1:\mathrm{T}}) &= \frac{p(\mathbf{y}_{1:\mathrm{T}}|\boldsymbol{\theta}) \cdot p(\boldsymbol{\theta})}{p(\mathbf{y}_{1:\mathrm{T}})} \\ &\propto p(\mathbf{y}_{1:\mathrm{T}}|\boldsymbol{\theta}) \cdot p(\boldsymbol{\theta}), \end{aligned} \tag{4}$$

where $p(\boldsymbol{\theta})$ is the *prior* PDF, $p(\mathbf{y}_{1:\mathrm{T}}|\boldsymbol{\theta})$ is the *likelihood* function, $p(\mathbf{y}_{1:\mathrm{T}})$ is a normalizing constant, and $\boldsymbol{\theta} = [\theta_1, \theta_2, \cdots, \theta_{\mathrm{P}}]$ is a vector of model parameters to be estimated. The prior probability represents the knowledge available for model parameter values before the data have been collected. The likelihood function is the *joint prior probability of observations*, that is, plausibility of the available observations $\mathbf{y}_{1:\mathrm{T}}$ given the parameter vector $\boldsymbol{\theta}$. Assuming that the observations are independent from each other, the joint likelihood function is defined as the product of the marginal likelihoods

$$p(\mathbf{y}_{1:\mathrm{T}}|\boldsymbol{\theta}) = \prod_{t=1}^{\mathrm{T}} p(\mathbf{y}_t|\mathbf{y}_{1:t-1}, \boldsymbol{\theta}). \tag{5}$$

In the BDLM framework, the likelihood function at time t is a Gaussian distribution following

$$p(\mathbf{y}_t|\mathbf{y}_{1:t-1}, \boldsymbol{\theta}) = \mathcal{N}(\mathbf{y}_t; \mathbf{C}_t\boldsymbol{\mu}_{t|t-1}, \mathbf{R}_t + \mathbf{C}\boldsymbol{\Sigma}_{t|t-1}\mathbf{C}^{\intercal}), \tag{6}$$

where the model matrices $\{\mathbf{C}_t, \mathbf{R}_t\}$ are defined in Section 2, the prior expected value $\boldsymbol{\mu}_{t|t-1}$ and prior covariance matrix $\boldsymbol{\Sigma}_{t|t-1}$ for the hidden state variables are obtained from the prediction step in the Kalman filter. Replacing the likelihood function in Equation 4 by Equation 5 results in

$$p(\boldsymbol{\theta}|\mathbf{y}_{1:\mathrm{T}}) \propto \prod_{t=1}^{\mathrm{T}} p(\mathbf{y}_t|\mathbf{y}_{1:t-1}, \boldsymbol{\theta}) \cdot p(\boldsymbol{\theta}). \tag{7}$$

In order to avoid zero underflow, the posterior probability density is transformed into the log space. Equation 7 then becomes

$$\ln p(\boldsymbol{\theta}|\mathbf{y}_{1:\mathrm{T}}) \propto \underbrace{\sum_{t=1}^{\mathrm{T}} \ln p(\mathbf{y}_t|\mathbf{y}_{1:t-1}\boldsymbol{\theta}) + \ln p(\boldsymbol{\theta})}_{\mathcal{L}_p(\boldsymbol{\theta})}, \tag{8}$$

where $\mathcal{L}_p(\boldsymbol{\theta})$ denotes *log-posterior* function. The MAP method identifies the optimal parameter estimates $\boldsymbol{\theta}^*$ that maximize the log-posterior function

$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} \left[\mathcal{L}_p(\boldsymbol{\theta})\right].$$

In this paper, the *Newton-Raphson* (NR) algorithm [10] is employed to carry on the optimization task. The stopping criterion for the NR algorithm is defined as

$$\begin{cases} \mathcal{L}_p(\boldsymbol{\theta}^{i-1}) & < \quad \mathcal{L}_p(\boldsymbol{\theta}^{i}) \\ \left|\mathcal{L}_p(\boldsymbol{\theta}^{i}) - \mathcal{L}_p(\boldsymbol{\theta}^{i-1})\right| & \leq \quad \tau \cdot \left|\mathcal{L}_p(\boldsymbol{\theta}^{i-1})\right| \end{cases}, \tag{9}$$

where i corresponds to optimization loop and τ is a termination tolerance on the log-posterior function.

Because the regularization of the optimization is provided by the prior PDF, the performance of the MAP method is sensitive to the selection of the prior PDF. An example that provides the details regarding the choice of the prior PDF is presented in Section 6.1.2. In addition to the sensitivity towards the prior distribution, the MAP by itself is a point estimation method so that it does not take into account the *uncertainty* in the parameter estimates. This challenge can be addressed using the Laplace approximation [10]. Hence, the posterior distribution of parameters is approximated with a Gaussian distribution

$$p(\boldsymbol{\theta}|\mathbf{y}_{1:\mathrm{T}}) \approx \mathcal{N}\left(\boldsymbol{\theta}; \boldsymbol{\theta}^*, -\mathcal{H}(\boldsymbol{\theta}^*)^{-1}\right), \tag{10}$$

where $\mathcal{H}(\boldsymbol{\theta}^*)$ is the second derivative of the log-posterior function $\mathcal{L}_p(\boldsymbol{\theta})$ evaluated at the MAP optimal parameter values $\boldsymbol{\theta}^*$.

3.2 Hamiltonian Monte Carlo

Hamiltonian Monte Carlo (HMC) [9,31] is known as a Markov Chain Monte Carlo (MCMC) method for approximating the posterior PDF $p(\boldsymbol{\theta}|\mathbf{y}_{1:\mathrm{T}})$ of the parameters of interest $\boldsymbol{\theta}$, given the training data $\mathbf{y}_{1:\mathrm{T}}$. The particularity of the HMC algorithm is that an auxiliary momentum variable r_i is added to each parameter θ_i. The joint probability density for a parameter vector $\boldsymbol{\theta}$ and its momentum variable vector $\boldsymbol{r}$ is defined as

$$p(\boldsymbol{\theta}, \boldsymbol{r}|\mathbf{y}_{1:\mathrm{T}}) = p(\boldsymbol{r}|\boldsymbol{\theta}, \mathbf{y}_{1:\mathrm{T}}) \cdot p(\boldsymbol{\theta}|\mathbf{y}_{1:\mathrm{T}}), \tag{11}$$

where $p(\boldsymbol{r}|\boldsymbol{\theta}, \mathbf{y}_{1:\mathrm{T}})$ is a conditional probability density of $\boldsymbol{r}$ given $\boldsymbol{\theta}$. The joint density $p(\boldsymbol{\theta}, \boldsymbol{r}|\mathbf{y}_{1:\mathrm{T}})$ is also called the *canonical distribution* that is independent from the choice of parameterization [3]. Hence, the joint probability density can be written in another form using an invariant *Hamiltonian* function $H(\boldsymbol{\theta}, \boldsymbol{r})$ as

$$p(\boldsymbol{\theta}, \boldsymbol{r}|\mathbf{y}_{1:\mathrm{T}}) = \exp\left[-H(\boldsymbol{\theta}, \boldsymbol{r})\right]. \tag{12}$$

The Hamiltonian function originally comes from the classical mechanics where it refers to the energy at specific points and is *conservative* over time. In most cases, $H(\boldsymbol{\theta}, \boldsymbol{r})$ is decomposed into two terms

$$H(\boldsymbol{\theta}, \boldsymbol{r}) = T(\boldsymbol{\theta}, \boldsymbol{r}) + V(\boldsymbol{\theta}),$$

where $T(\boldsymbol{\theta}, \boldsymbol{r})$ is the *kinetic* energy and $V(\boldsymbol{\theta})$ is the *potential* energy. In the case of HMC, $H(\boldsymbol{\theta}, \boldsymbol{r})$ can be obtained using the Equations 11 and 12, so that

$$H(\boldsymbol{\theta}, \boldsymbol{r}) = \underbrace{-\ln\, p(\boldsymbol{r})}_{T(\boldsymbol{r})} \quad \underbrace{-\ln\, p(\boldsymbol{\theta}|\mathbf{y}_{1:\mathrm{T}})}_{V(\boldsymbol{\theta})} \tag{13}$$

with the assumption that the momentum variables $\boldsymbol{r}$ do not depend on the parameters $\boldsymbol{\theta}$. In common cases, the kinetic energy is defined as

$$T(\boldsymbol{r}) = \frac{1}{2}\boldsymbol{r}^{\intercal}\mathbf{M}^{-1}\boldsymbol{r}, \tag{14}$$

where $\mathbf{M}$ is symmetric and positive-definite *mass matrix*. $\mathbf{M}$ corresponds to the inverted covariance matrix of $p(\boldsymbol{\theta}|\mathbf{y}_{1:\mathrm{T}})$. Once the kinetic and potential energies are identified, the Hamiltonian's equations over time can be written as

$$\begin{aligned} \frac{d\boldsymbol{r}}{dt} &= -\nabla_{\boldsymbol{\theta}}\, V(\boldsymbol{\theta}) \\ \frac{d\boldsymbol{\theta}}{dt} &= \nabla_{\boldsymbol{r}}\, T(\boldsymbol{r}), \end{aligned} \tag{15}$$

where ∇ is the gradient operator. Equation 15 is employed to propose new samples in HMC. For the practical implementation, Equation 15 can be approximated using the *leapfrog* method, that is, a *symplectic integrator*, that allows simulating the trajectories for an efficient exploration of the posterior density $p(\boldsymbol{\theta}|\mathbf{y}_{1:\mathrm{T}})$. Given a time discretization parameter n, the main steps in the leapfrog method are written as

$$\begin{aligned} \boldsymbol{r}^{n+\frac{1}{2}} &= \boldsymbol{r}^{n} - \tfrac{\epsilon}{2}\,\nabla_{\boldsymbol{\theta}}\, V(\boldsymbol{\theta}^{n}) \\ \boldsymbol{\theta}^{n+1} &= \boldsymbol{\theta}^{n} + \epsilon\,\nabla_{\boldsymbol{r}}\, T(\boldsymbol{r}^{n+\frac{1}{2}}) \\ \boldsymbol{r}^{n+1} &= \boldsymbol{r}^{n+\frac{1}{2}} - \tfrac{\epsilon}{2}\,\nabla_{\boldsymbol{\theta}}\, V(\boldsymbol{\theta}^{n+1}), \end{aligned}$$

where ϵ is defined as the step size. A half step for the momentum vector $\boldsymbol{r}^{n+\frac{1}{2}}$ is first evaluated. Then, a full step for the parameter vector $\boldsymbol{\theta}^{n+1}$ is updated using $\boldsymbol{r}^{n+\frac{1}{2}}$. Finally, the other half step for the momentum vector $\boldsymbol{r}^{n+1}$ is computed using $\boldsymbol{\theta}^{n+1}$. The iterative process is repeated a number of steps L. A limitation of the leapfrog method is that it can introduce errors during discretizations, leading to a bias. Therefore, an acceptance probability β is defined to ensure the validity of the Markov chain:

$$\begin{aligned} \beta &= \min\left\{1,\, \frac{\exp\left[-H(\boldsymbol{\theta}^{n+1}, \boldsymbol{r}^{n+1})\right]}{\exp\left[-H(\boldsymbol{\theta}^{n}, \boldsymbol{r}^{n})\right]}\right\} \\ &= \min\left\{1,\, \exp\left[-T(\boldsymbol{r}^{n+1}) - V(\boldsymbol{\theta}^{n+1}) + T(\boldsymbol{r}^{n}) + V(\boldsymbol{\theta}^{n})\right]\right\}. \end{aligned} \tag{16}$$

The key challenge in HMC is to tune the parameters such as the step size ϵ and the number of steps L [3, 20]. A small step size provides a more accurate approximation and effective exploration, yet it is computationally more demanding. A large step size leads to inaccurate simulations and yields low acceptance probabilities. Similarly, a small number of steps yields a high autocorrelation between the successive samples. A larger number of steps causes back loop trajectories [20, 31], leading to a poor exploration. Optimal values for ϵ and L are tuned based on the acceptance probability in Equation 16 using either the Dual Averaging method [32] or the No-U-Turn Sampler method [20].

In order to measure the efficiency exploration in HMC, we employ the convergence diagnostic statistic $\hat{\mathbf{R}}$ [12], i.e. *estimated potential scale reduction*. The idea behind is to interpret the stationarity of multiple, parallel Markov chains based on the quantity $\hat{\mathbf{R}}$. If $\hat{\mathbf{R}}$ is approximating to 1, the estimates obtained from the Markov chains are reliable.

4 Gaussian Mixture for Hidden State Variables

In the BDLMs, a vector of hidden state variables is assumed to be a multivariate Gaussian distribution, as presented in Equation 1. Because the new parameter estimation methods presented in Section 3, each model parameter is now represented by a probability distribution. Therefore, the vector of hidden state variables at each time step t is then described by a population of multivariate Gaussian distributions. In order to make an intuitive interpretation, this population of multivariate Gaussian distributions is approximated with a single multivariate Gaussian distribution using the *Gaussian Mixture* (GM) approach [30]. This section presents the GM approach proposed for including the model parameter uncertainty in the estimation of the hidden state variables. Assuming that we have N samples from the parameter posterior PDF approximated using either the MAP with the Laplace approximation, or the HMC method, the mixture density of hidden state variables is built from a linear combination of N Gaussian densities, each obtained using Equation 2. The mathematical formulation for the Gaussian mixture density is written as

$$\begin{aligned} p(\hat{\mathbf{x}}_t|\mathbf{y}_{1:T}) &= \frac{\sum_{n=1}^{N} p(\mathbf{x}_t^n|\boldsymbol{\theta}^n, \mathbf{y}_{1:t}) \cdot w^n}{\sum_{n=1}^{N} w^n} \\ &= \frac{\sum_{n=1}^{N} \mathcal{N}(\mathbf{x}_t^n; \boldsymbol{\mu}_{t|t}^n, \Sigma_{t|t}^n) \cdot w^n}{\sum_{n=1}^{N} w^n}, \end{aligned}$$

where $\boldsymbol{\mu}_{t|t}^n$ and $\boldsymbol{\Sigma}_{t|t}^n$ that are obtained using Equation 3 and w^n are the mixing coefficients. $p(\hat{\mathbf{x}}_t|\mathbf{y}_{1:T})$ is approximated by a Gaussian distribution with mean $\hat{\boldsymbol{\mu}}_{t|t}$ and covariance matrix $\hat{\boldsymbol{\Sigma}}_{t|t}$ that can be calculated following

$$\begin{aligned} \hat{\boldsymbol{\mu}}_{t|t} &= \frac{\sum_{n=1}^{N} \boldsymbol{\mu}_{t|t}^n \cdot w^n}{\sum_{n=1}^{N} w^n} \\ \hat{\boldsymbol{\Sigma}}_{t|t} &= \frac{\sum_{n=1}^{N} \boldsymbol{\Sigma}_{t|t}^n \cdot w^n}{\sum_{n=1}^{N} w^n} + \frac{\sum_{n=1}^{N} (\boldsymbol{\mu}_{t|t}^n - \hat{\boldsymbol{\mu}}_{t|t})(\boldsymbol{\mu}_{t|t}^n - \hat{\boldsymbol{\mu}}_{t|t})^\intercal \cdot w^n}{\sum_{n=1}^{N} w^n}. \end{aligned} \tag{17}$$

This Gaussian distribution is the closest one to the true mixture distribution [24, 29]. Because samples $\boldsymbol{\theta}^n$ are realizations of the posterior PDF, w^n are all equals to one and Equation 17 becomes

$$\begin{aligned} \hat{\boldsymbol{\mu}}_{t|t} &= \frac{1}{N} \sum_{n=1}^{N} \boldsymbol{\mu}_{t|t}^n \\ \hat{\boldsymbol{\Sigma}}_{t|t} &= \frac{1}{N} \left[\sum_{n=1}^{N} \boldsymbol{\Sigma}_{t|t}^n + \sum_{n=1}^{N} (\boldsymbol{\mu}_{t|t}^n - \hat{\boldsymbol{\mu}}_{t|t})(\boldsymbol{\mu}_{t|t}^n - \hat{\boldsymbol{\mu}}_{t|t})^\intercal \right]. \end{aligned} \tag{18}$$

5 Methodology

This section presents two procedures (1) the maximum a posteriori with the Laplace approximation procedure (LAP-P) and (2) the Hamiltonian Monte Carlo procedure (HMC-P) for approximating the posterior PDF of model parameters and for estimating the hidden state variables.

5.1 Maximum A Posteriori with Laplace Approximation Procedure (LAP-P)

The LAP-P consists in two main steps: *Posterior Density Approximation* (PDA) and *Uncertainty Marginalization* (UM). The PDA-step is to approximate the parameter posterior density $p(\boldsymbol{\theta}|\mathbf{y}_{1:\mathrm{T}})$ using the MAP and the Laplace approximation presented in Section 3.1. The UM-step is related to the estimation of the expected values for hidden state variables $\hat{\boldsymbol{\mu}}_{t|t}$ and its covariance matrix $\hat{\boldsymbol{\Sigma}}_{t|t}$ based on Gaussian Mixture approach presented in Section 4.

In the PDA-step, the optimal parameter vector $\boldsymbol{\theta}^*$ of a model is first learned from a training set $\mathbf{y}_{1:\mathrm{T}}$ using the Newton-Raphson algorithm. The parameter posterior density $p(\boldsymbol{\theta}|\mathbf{y}_{1:\mathrm{T}})$ is then approximated using the Laplace approximation. Note that this density is supposed to be a Gaussian density with mean $\boldsymbol{\theta}^*$ and covariance matrix $-\mathcal{H}(\boldsymbol{\theta}^*)^{-1}$, where the operator $\mathcal{H}$ corresponds to the Hessian of the log-posterior function $\mathcal{L}_p(\boldsymbol{\theta})$ defined in Equation 8. The UM-step marginalizes the parameter uncertainty estimation using Equation 18 to estimate $\hat{\boldsymbol{\mu}}_{t|t}$ and $\hat{\boldsymbol{\Sigma}}_{t|t}$ at each time t. The LAP-P is summarized in Figure 1.

5.2 Hamiltonian Monte Carlo Procedure (HMC-P)

HMC-P has the same two steps as LAP-P, except that the PDA-step employs the HMC-based method presented in Section 3.2 for approximating the parameter posterior density $p(\boldsymbol{\theta}|\mathbf{y}_{1:\mathrm{T}})$. The shematic architecture of a HMC-P is illustrated in Figure 2.

To ensure an efficient performance for the HMC-P, the parameters $\{\epsilon, L\}$ for the leapfrog method presented in Section 3.2 need to be tuned in the PDA-step. For this purpose, the appropriate start point $\boldsymbol{\theta}^{\mathrm{start}}$ for the parameters along with the HMC sampler are required. Once the leapfrog parameters are identified, the samples are then drawn from the constructed sampler for approximating $p(\boldsymbol{\theta}|\mathbf{y}_{1:\mathrm{T}})$.

6 Case-Study

This section compares the LAP-P with the HMC-P for approximating the posterior PDF of model parameters and the estimation of the hidden state variables with respect to different training-set lengths (TSLs) on two case-studies: a simulated dataset and a real dataset for a dam in Canada. For this purpose, five tests associated with the TSL of $30, 90, 180, 365$ and 1095 days are employed in both case-studies.

6.1 Illustrative Example

The illustrative example studies a simulated dataset where the true values for the hidden state variables and the model parameters are known. The objective is to compare the performance of the LAP-P with the HMC-P based on the true values.

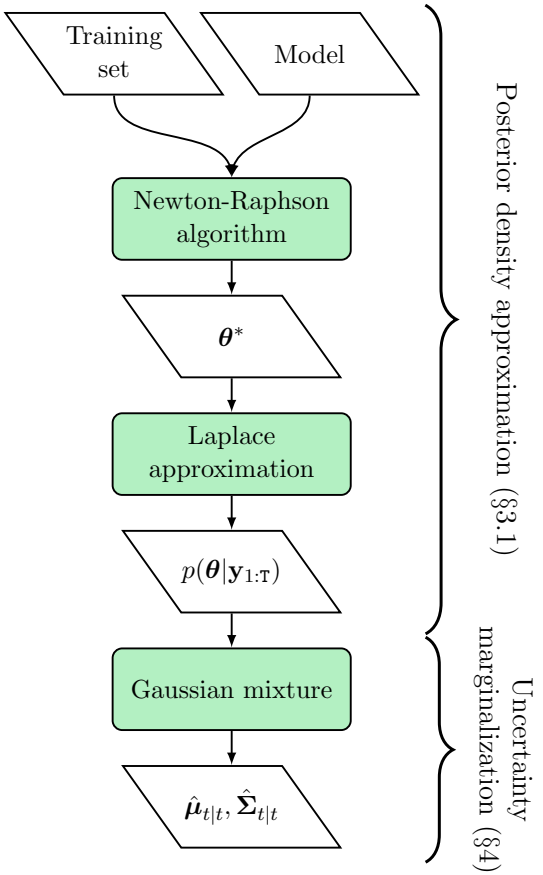


Figure 1: Illustration of the general procedure for approximating the posterior density of parameters $p(\boldsymbol{\theta}|\mathbf{y}_{1:\mathrm{T}})$ and the mean values of the hidden state variables and its covariance matrix at each time t i.e. $\{\hat{\boldsymbol{\mu}}_{t|t}, \hat{\boldsymbol{\Sigma}}_{t|t}\}$ using the combination of the MAP with the Laplace approximation and Gaussian Mixture approach.

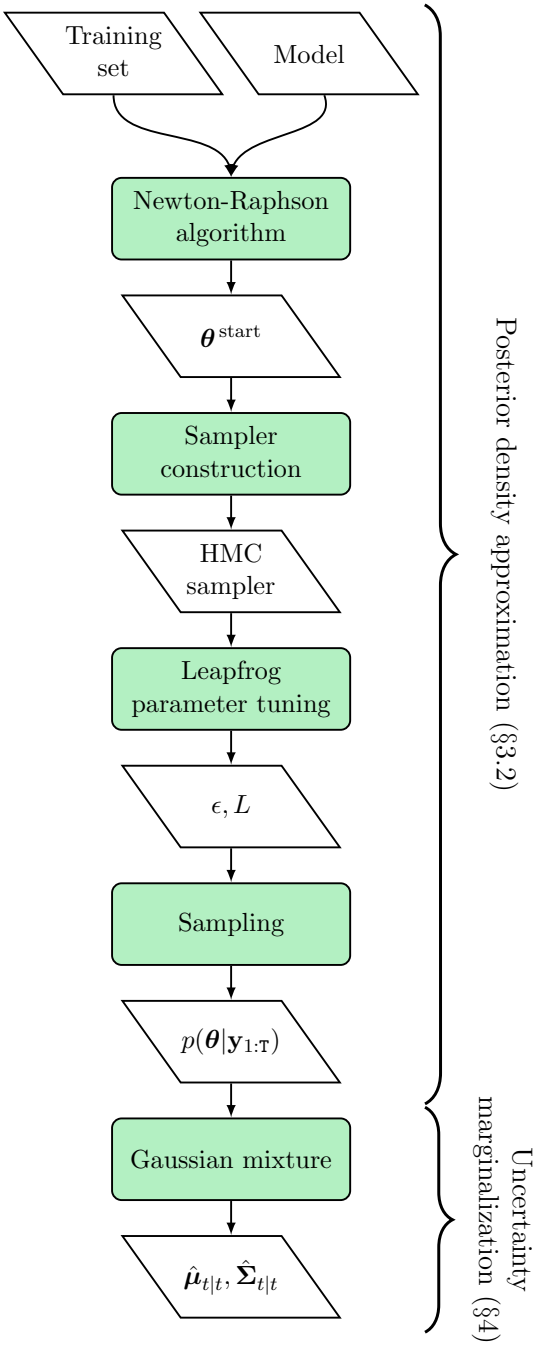


Figure 2: The two main steps for approximating the posterior density of parameters $p(\boldsymbol{\theta}|\mathbf{y}_{1:\text{T}})$ and the mean values of hidden state variables and its covariance matrix at each time t i.e. $\{\hat{\boldsymbol{\theta}}_{t|t}, \hat{\boldsymbol{\Sigma}}_{t|t}\}$ using the combination of the HMC method with Gaussian Mixture approach.

6.1.1 Simulated Data

The experiment is conducted on simulated data that are generated to be representative of the data recorded on civil infrastructure such as a dam. For this purpose, a dataset of displacement (D) measurements is generated including a *baseline* (B) to present the structural behavior over time, a *seasonal cycle* (S) to describe the thermal effect of environmental conditions on the displacement, an *autoregressive* (AR) process to artificially introduce model prediction errors over time, and observation errors (v_t). The dataset is formulated following

$$y_t^{\mathtt{D}} = x_t^{\mathtt{B}} + x_t^{\mathtt{S}} + x_t^{\mathtt{AR}} + v_t \qquad [\text{mm}],$$

where each component is generated using the following formulation

$$\begin{aligned} x_t^{\mathtt{B}} &= 3 + w_t^{\mathtt{B}}, & w_t^{\mathtt{B}} &\sim \mathcal{N}(0, (\underbrace{10^{-5}}_{\sigma_w^{\mathtt{B}}})^2) \\ x_t^{\mathtt{S}} &= 4\sin\left[\tfrac{2\pi}{365.24}\cdot(t+15)\right] \\ x_t^{\mathtt{AR}} &= \underbrace{0.866}_{\phi^{\mathtt{AR}}}\cdot x_{t-1}^{\mathtt{AR}} + w_t^{\mathtt{AR}}, & w_t^{\mathtt{AR}} &\sim \mathcal{N}(0, (\underbrace{0.05}_{\sigma_w^{\mathtt{AR}}})^2) \\ v_t &\sim \mathcal{N}(0, (\underbrace{0.1}_{\sigma_v})^2). \end{aligned}$$

A four-years-dataset (1461 observations) with a uniform time-step length of 24 hours is generated from the simulated setting, as shown in Figure 3.

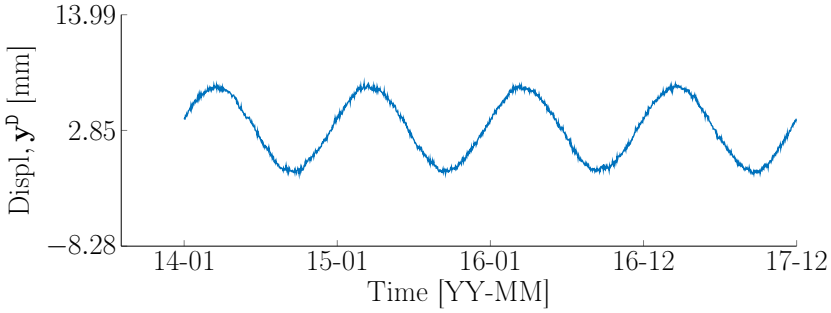


Figure 3: Illustration of 4 years of simulated data.

6.1.2 Model Construction

A common model is built for the experiment using both LAP-P and HMC-P. In this model, each observation is decomposed into a *baseline* component to model the structural behavior over time, a *periodic* component with a period of 365.24 days to model the environmental conditions, and an *autoregressive* component to describe model prediction errors. Hence, the vector of hidden state variables is defined as

$$\mathbf{x}_t = \left[\underbrace{x_t^{\mathtt{B}}}_{\text{baseline}}, \underbrace{x_t^{\mathtt{S1,T1}}, x_t^{\mathtt{S2,T1}}}_{\text{cycle, } p=365.24\,\text{days}}, \underbrace{x_t^{\mathtt{AR}}}_{\text{AR}}\right]^{\mathtt{T}}. \tag{19}$$

The model involves a vector of unknown parameters that are defined following

$$\boldsymbol{\theta} = \left[\sigma_w^{\mathtt{B}}, \phi^{\mathtt{AR}}, \sigma_w^{\mathtt{AR}}, \sigma_v\right]^{\mathtt{T}}, \tag{20}$$

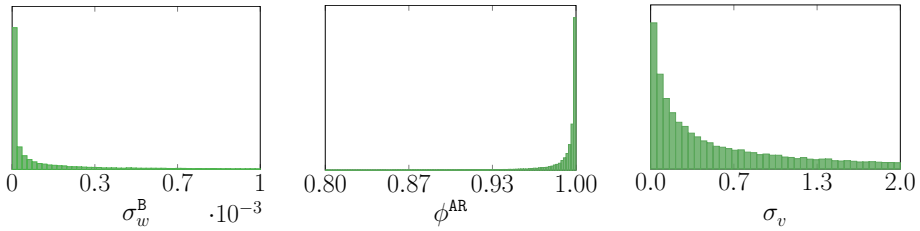
where $\sigma_w^{\mathtt{B}}$ is the *baseline standard deviation*, $\phi^{\mathtt{AR}}$ is the *autocorrelation coefficient*, $\sigma_w^{\mathtt{AR}}$ is the *autocorrelation standard deviation*, and σ_v is the *observation error standard deviation*. $\sigma_w^{\mathtt{B}}, \sigma_w^{\mathtt{AR}}$ and σ_v are positive real numbers $\mathbb{R}^+$, whereas $\phi^{\mathtt{AR}}$ is defined in range between 0 and 1. The full model matrices can be found in Appendix A. The parameters being estimated are commonly transformed to unbounded spaces for an efficient estimation [10]. For this purpose, a logarithms-base-10 and sigmoid functions are applied to the standard deviations; $\sigma \in (0, \infty)$ and the autocorrelation coefficient; $\phi^{\mathtt{AR}} \in (0, 1)$, respectively, so that

$$\begin{aligned} \sigma^{\mathtt{TR}} &= \log_{10}(\sigma) \\ \phi^{\mathtt{AR,TR}} &= \frac{1}{1+\exp(-4\,\phi^{\mathtt{AR}})}, \end{aligned}$$

where the superscript $\mathtt{TR}$ stands for transformed space. The initial parameter values in the original space for the model are

$$\boldsymbol{\theta}^0 = \left[\underbrace{10^{-4}}_{\sigma_w^{\mathtt{B}}}, \underbrace{0.7}_{\phi^{\mathtt{AR}}}, \underbrace{0.01}_{\sigma_w^{\mathtt{AR}}}, \underbrace{0.026}_{\sigma_v}\right]^{\mathtt{T}}.$$

In the BDLM framework, we expect the baseline standard deviation $\sigma_w^{\mathtt{B}}$ to be small because the error between the model prediction and the observation at each time step will be captured in the autoregressive component $\mathbf{x}^{\mathtt{AR}}$. Hence, it yields to high autocorrelations in the model prediction errors between time steps. The autocorrelation coefficient $\phi^{\mathtt{AR}}$ is assumed to be close to one. This prior knowledge defines the prior PDF for $\sigma_w^{\mathtt{B}}$ and $\phi^{\mathtt{AR}}$ in order to ensure a reliable estimation [11, 13]. The remaining parameters are expected to be near to a unit scale. The prior distributions in the transformed space associated with each parameter in Equation 20 are set as follows $f(\sigma_w^{\mathtt{B,TR}}) = \mathcal{N}(-4, 2)$, $f(\phi^{\mathtt{AR,TR}}) = \mathcal{N}(1.5, 0.5)$, $f(\sigma_w^{\mathtt{AR,TR}}) = \mathcal{N}(0, 1)$, and $f(\sigma_v^{\mathtt{TR}}) = \mathcal{N}(0, 1)$. Figure 4 illustrates the different prior distributions corresponding to three parameters $\left\{\sigma_w^{\mathtt{B}}, \phi^{\mathtt{AR}}, \sigma_v\right\}$ represented in the original space.



(a) $\sigma_w^{\mathtt{B}}$ (b) $\phi^{\mathtt{AR}}$ (c) σ_v

Figure 4: Illustration of prior distribution choices for model parameters in the original space.

Because the nature of LAP-P and HMC-P is different from each other, some setup options for parameter estimation must be tuned separately. In the LAP-P, the termination tolerance τ in Equation 9 is set to 10^{-7} and the initial parameter values in the original space are equal to $\boldsymbol{\theta}^0$. The HMC-P employs 4 parallel Markov-chains (C_c) for each training set length, where each Markov-chain has its own initial parameter values. These initial parameter values are directly defined in the transformed space as follows

$$\begin{aligned} \boldsymbol{\theta}^{0,\mathrm{C}_1} &= \boldsymbol{\theta}^{\,\mathrm{start}} \\ [\boldsymbol{\theta}^{0,\mathrm{C}_2}, \boldsymbol{\theta}^{0,\mathrm{C}_3}, \boldsymbol{\theta}^{0,\mathrm{C}_4}] &= \mathcal{N}\left(\boldsymbol{\theta}^{\,\mathrm{start}}, \mathrm{diag}([1\ 1\ 1])\right), \end{aligned}$$

where $\boldsymbol{\theta}^{\,\mathrm{start}}$ is a vector of pre-estimated parameters as presented in Section 5.2. The mass matrix $\mathbf{M}$ in Equation 15 is chosen as the negative diagonal Hessian matrix of $\ln p(\boldsymbol{\theta}|\mathbf{y}_{1:\mathrm{T}})$. Here, the stopping criterion is when the quantity $\hat{\mathbf{R}}$ is less than 1.01 (See Section 3.2).

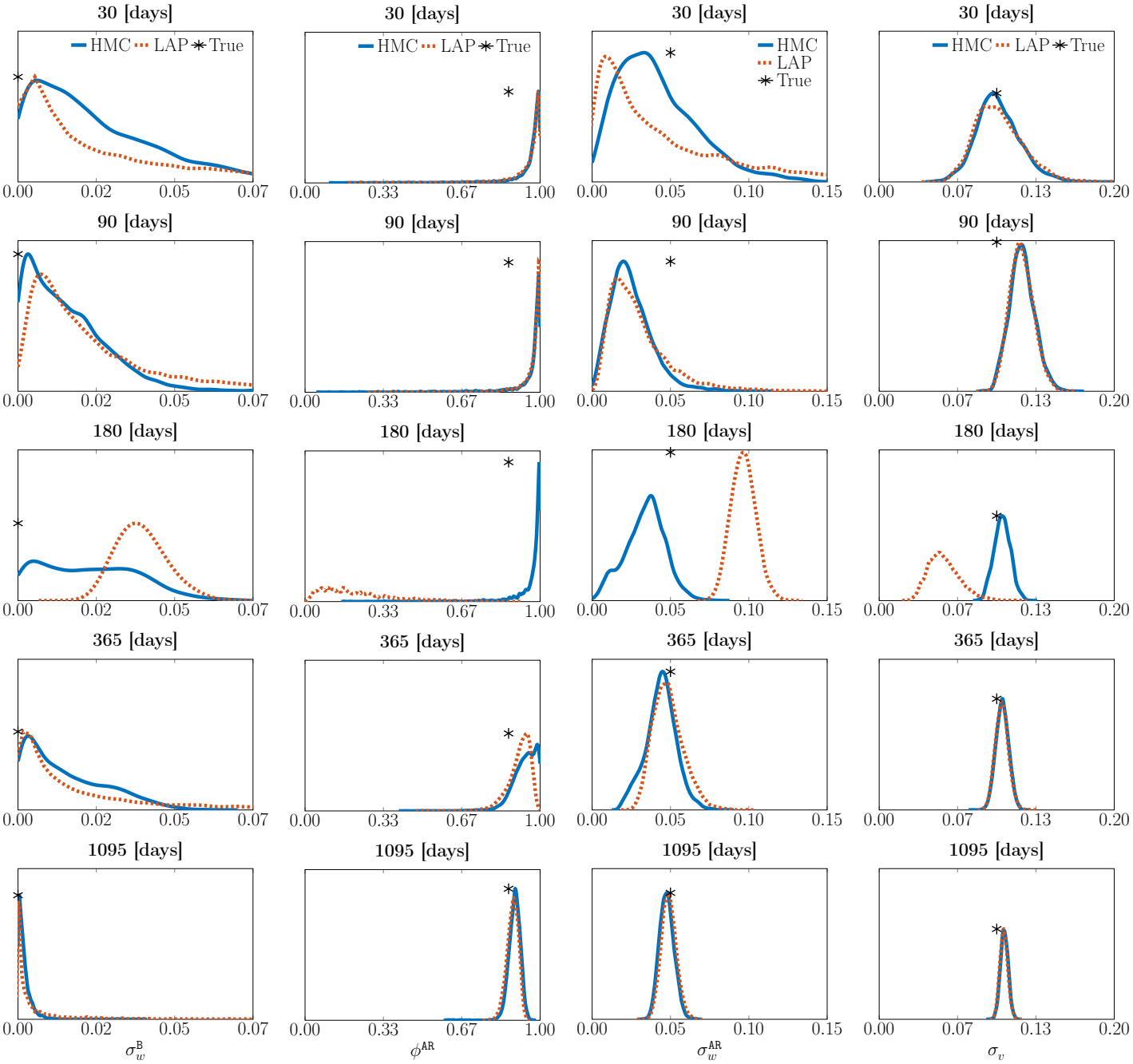
6.1.3 Results

Five tests on the different training-set lengths (TSLs); 30 days , 90 days, 180 days, 365 days, and 1095 days using both LAP-P and HMC-P have been conducted for approximating the posterior PDF of model parameters as well as for the estimation of the hidden state variables. Figure 5 shows the Kernel smoothing function estimate [4] of the PDFs for each parameter according to each TSL. The dashed line and the solid line represent the PDFs obtained using LAP-P and HMC-P, respectively. The *true parameter values* ($\check{\boldsymbol{\theta}}$) are presented by the asterisks.

It can be seen that the PDFs obtained from both procedures concentrate around $\check{\boldsymbol{\theta}}$ as the training dataset length increases. In the case for the TSL of 180 days, where the PDFs approximated using LAP-P are shifted away from those approximated using HMC-P and $\check{\boldsymbol{\theta}}$. This behavior can be explained by the sensitivity to the initial parameter values in LAP-P, leading to a local maximum. Note that in this case, the full Hessian matrix cannot be inverted so that only diagonal terms are used to compute parameter covariance matrix. In addition, the posterior PDF for $\sigma_w^{\mathtt{B}}$ approximated using HMC-P, as presented in Figure 5a, has a larger posterior mass in the tail than the other shorter TSLs. This behavior is contrary to the general intuition that the more data we have, the more posterior mass concentrates around $\check{\boldsymbol{\theta}}$. This behavior justifies that the extraction of information from data depends not only on the size of data, but also on the interaction of the prior with the likelihood function and the parameter being estimated as noted by Gelman et al [13].

The dominance of the prior PDF can be seen on the autocorrelation coefficient $\phi^{\mathtt{AR}}$ presented in Figure 5b for the small dataset size such as $30, 90$ and 180 days. However, the prior becomes dominated by the information from data, as illustrated in the TSLs of 365 and 1095 days. In the TSL of 1095 days, the parameter posterior PDFs obtained with both procedures are identical except the heavy tailed posterior PDF for $\sigma_w^{\mathtt{B}}$ estimated using LAP-P. Overall, HMC-P shows a superior capacity at approximating the parameter posterior PDF over LAP-P.

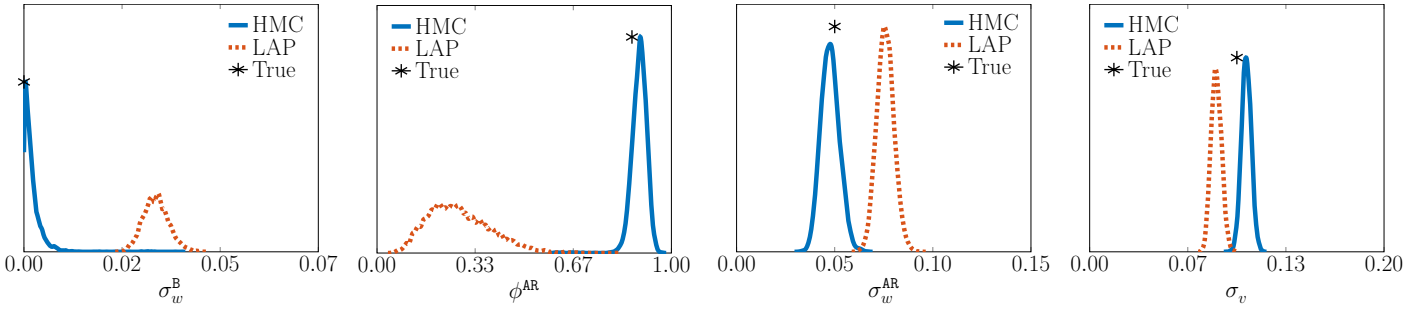
Biased estimation with LAP-P for the TLS of 180 days leads to the question of the robustness of LAP-P with respect to the choice of initial parameter values. To answer this question, an addition test using a different set of initial parameter values, is carried out



(a) $\sigma_w^{\mathtt{B}}$ (b) $\phi^{\mathtt{AR}}$ (c) $\sigma_w^{\mathtt{AR}}$ (d) σ_v

Figure 5: Each column illustrates the Kernel smoothing function estimate of the posterior PDFs for each model parameter $p(\theta_i|\mathbf{y}_{1:\mathtt{T}})$ in the original space with respect to the training-set length of simulated data.

with the TLS of 1095 days. Note that in the previous test, the LAP-P performed well in the approximation of the parameter posterior PDFs, where these PDFs concentrated around the $\check{\boldsymbol{\theta}}$. Figure 6 presents the Kernel smoothing function estimate of the posterior PDFs for each parameter. Instead of concentrating around $\check{\boldsymbol{\theta}}$, the PDFs obtained using LAP-P are



(a) $\sigma_w^{\mathtt{B}}$ (b) $\phi^{\mathtt{AR}}$ (c) $\sigma_w^{\mathtt{AR}}$ (d) σ_v

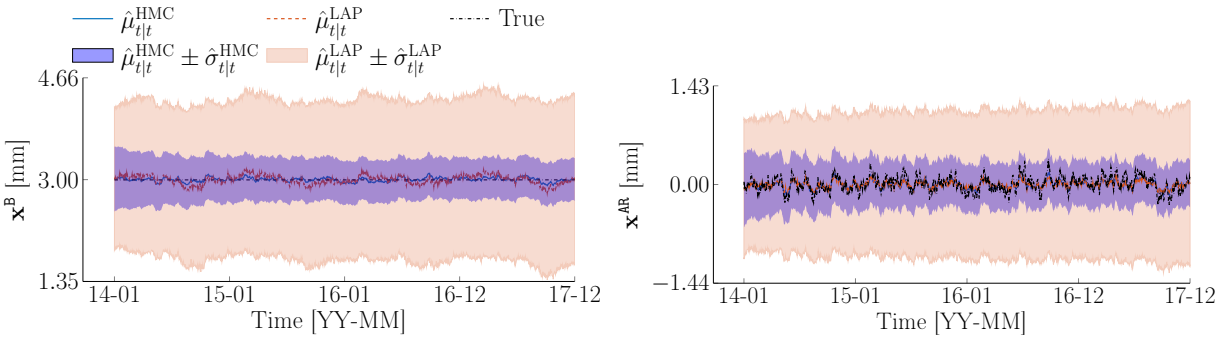
Figure 6: Illustration of the Kernel smoothing function estimate of the posterior PDFs for each model parameter in the original space with the training set of 1095 days.

far from them. It illustrates the effect of poor initial parameter values on the approximation of the parameter posterior PDF. Hence, a careful tuning of the initial parameter values in LAP-P is essential for an accurate estimation. For this purpose, the different sets of initial parameter values should be tested during training.

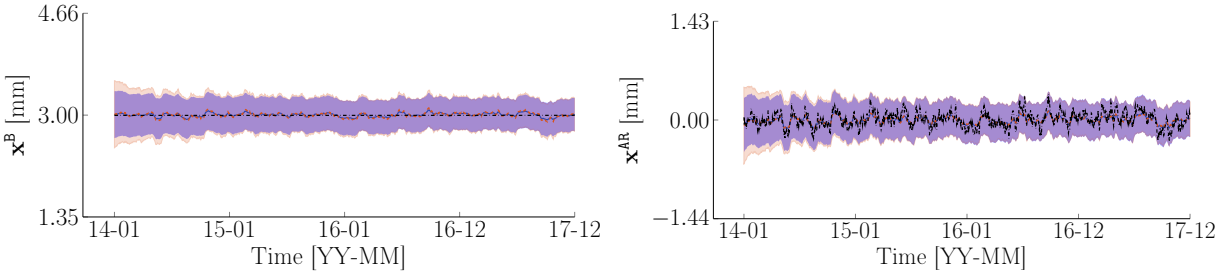
For both procedures, the hidden state variables are estimated using 1000 samples from the joint parameter posterior PDF. This number of samples provides a sufficient accuracy for the estimation of the hidden state variables, because the same results are found for the larger sizes of samples. Figure 7 presents the hidden state variables estimated using Kalman Smoother [30] for the entire dataset with respect to the TSL. We only present the results for the baseline ($\mathbf{x}^{\mathtt{B}}$) and autoregressive ($\mathbf{x}^{\mathtt{AR}}$) components. The mean value and its standard deviation at time t are $\hat{\boldsymbol{\mu}}_{t|t}^{(\cdot)}$ and $\hat{\boldsymbol{\sigma}}_{t|t}^{(\cdot)}$, respectively, where the superscript $(\cdot)$ is associated with either LAP-P or HMC-P being employed for the estimation task. The mean value $\hat{\boldsymbol{\mu}}_{t|t}^{\mathrm{LAP}}$ and its uncertainty bound $\hat{\boldsymbol{\mu}}_{t|t}^{\mathrm{LAP}} \pm \hat{\boldsymbol{\sigma}}_{t|t}^{\mathrm{LAP}}$ at time t are represented by the dashed line and the shaded region delimited by the solid line, respectively. Meanwhile, $\hat{\boldsymbol{\mu}}_{t|t}^{\mathrm{HMC}}$ and $\hat{\boldsymbol{\mu}}_{t|t}^{\mathrm{HMC}} \pm \hat{\boldsymbol{\sigma}}_{t|t}^{\mathrm{HMC}}$ are represented by the solid line and the shaded region, respectively. The true hidden state variables ($\check{\mathbf{x}}_t$) are presented by the dash-dot line.

As with the posterior PDFs for model parameters, the estimation for the hidden state variables keeps improving as the amount of the training data increases. The mean values for the hidden state variables tend to $\check{\mathbf{x}}_t$ and their uncertainty bounds narrow down. More importantly, these uncertainty bounds include with $\check{\mathbf{x}}_t$ in almost all TLSs, except 180 days.

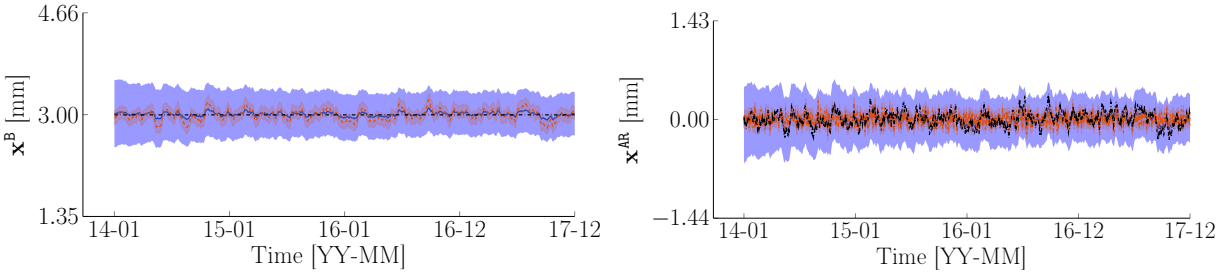
For 180 days, the estimation of the hidden state variables using LAP-P suffers from the biased posterior PDF, as shown in Figure 7c. The baseline and autoregressive components obtained from LAP-P are not well separated as expected, even though their uncertainty bounds are smaller than those obtained from HMC-P. Meanwhile, the uncertainty bounds for the baseline and autoregressive components estimated using HMC-P are larger than



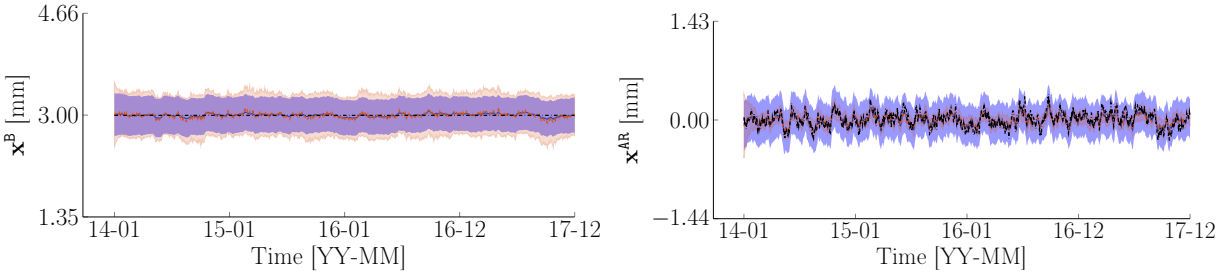
(a) Training set of 30 days (30 data points)



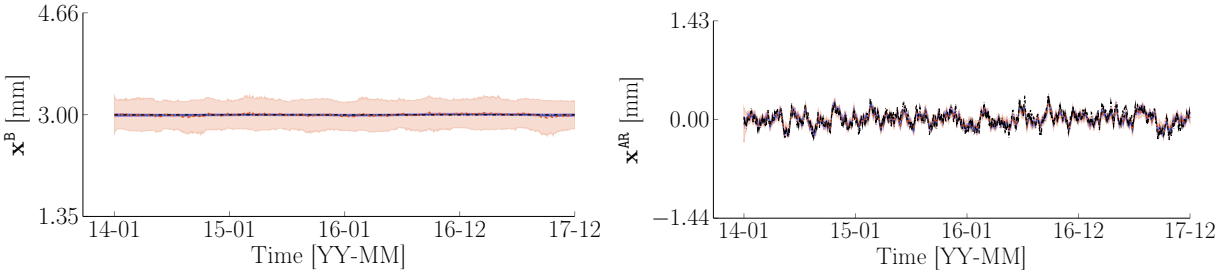
(b) Training set of 90 days (90 data points)



(c) Training set of 180 days (180 data points)



(d) Training set of 365 days (365 data points)



(e) Training set of 1095 days (1095 data points)

Figure 7: Expected value $\hat{\mu}$ and standard deviation $\hat{\sigma}$ for baseline (left) and autoregressive (right) components using LAP-P and HMC-P with respect to the training-set length of the simulated data.

those from the shorter TSL of 90 days.

The uncertainty bounds of the baseline component obtained from LAP-P are larger than those obtained from HMC-P in almost TLSs except for 180 days. The discrepancy between both procedures is clearly observable in the TSLs of 30 and 1095 days, as illustrated in Figures 7a and e. The heavy tailed posterior PDFs for $\sigma_w^{\mathtt{B}}$ obtained from LAP-P (Figure 5a) is to be blamed for such this behavior. For the TSL of 365 days, the uncertainty bounds of the autoregressive component estimated using LAP-P are unexpectedly smaller than those estimated using HMC-P. This is explained by a more precise approximation with LAP-P than with HMC-P in the posterior PDFs for $\phi^{\mathtt{AR}}$ and $\sigma_w^{\mathtt{AR}}$ presented in Figures 5b and c.

These results illustrate a potential impact of the model parameter uncertainty on the estimation of the hidden state variables. Also, the mean values and their uncertainty bounds for the hidden state variables obtained from HMC-P overall are more reliable than those obtained from LAP-P because of the lack of sensitivity test with respect to the initial parameter values in LAP-P.

6.2 A Dam in Canada

In this case-study, the comparison between the LAP-P and the HMC-P is illustrated using the horizontal displacement data collected on a dam located in Canada. The horizontal displacement is measured using an automatic inverted-pendulum along the X-direction. The location of the studied sensor is shown in Figure 8.

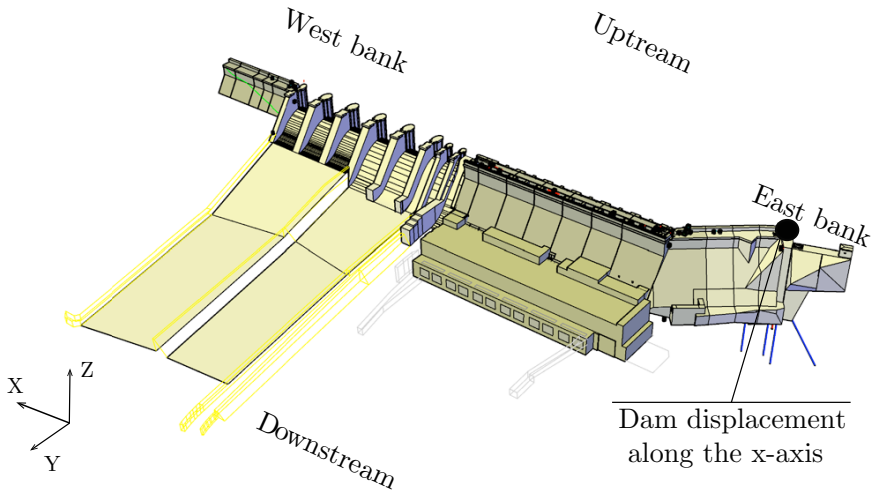


Figure 8: Sensor localization on the dam.

6.2.1 Data Description

The horizontal displacement data are collected over the period of 4 years from 2010 to 2014 with a total of 2679 data points. The entire dataset is shown in Figure 9. A descending trend and a periodic pattern with a period of one year can be observed from the raw data. The periodic pattern reaches its maximum during winter and minimum during

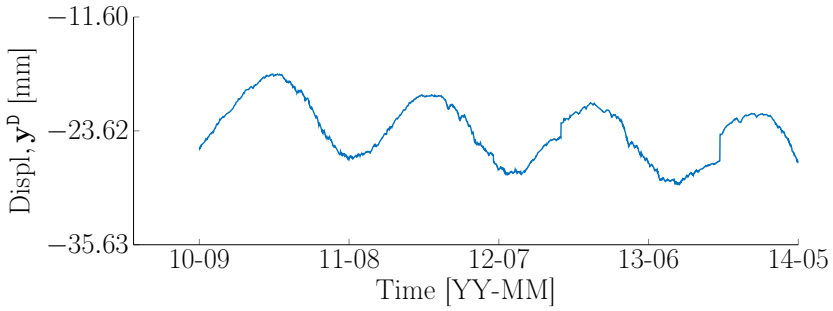


Figure 9: Raw displacement data.

summer. Such a behavior is attributed to the temperature effect. Note that the data are collected with a non-uniform time-step length, as shown in Figure 10. The time-step length varies in the range from 1 to 216 hours, where the most frequent time-step is 12 hours. A reference time-step [15] corresponding to the most frequent time-step according to the studied training-set is selected.

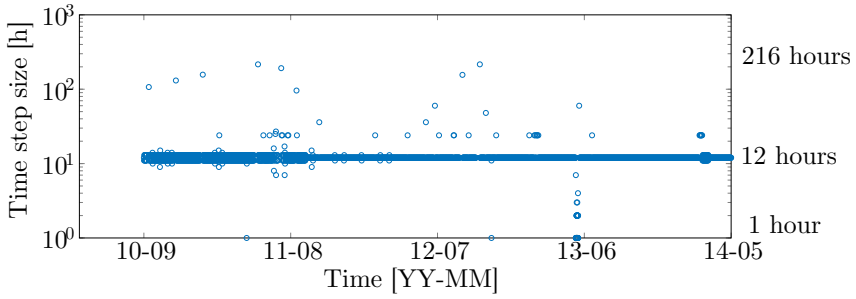


Figure 10: Time step size

6.2.2 Model Construction

Similarly to the previous model detailed in Section 6.1.2, a new model is constructed using the same vector of hidden state variables, with an additional trend component. Because of the descending trend behavior observed from the raw data (Figure 9), the trend component is needed to model the rate of change in the baseline component. Therefore, the vector of hidden state variables is written as

$$\mathbf{x}_t = \left[\underbrace{x_t^{\mathtt{B}}}_{\text{baseline}}, \underbrace{x_t^{\mathtt{T}}}_{\text{trend}}, \underbrace{x_t^{\mathtt{S1,T1}}, x_t^{\mathtt{S2,T1}}}_{\text{cycle, } p=365.24\,\text{days}}, \underbrace{x_t^{\mathtt{AR}}}_{\text{AR}} \right]^{\mathsf{T}}.$$

The parameter vector $\boldsymbol{\theta}$ corresponding to the model is defined following

$$\boldsymbol{\theta} = \left[\sigma_w^{\mathtt{T}}, \phi^{\mathtt{AR}}, \sigma_w^{\mathtt{AR}}, \sigma_v\right]^{\mathsf{T}},$$

where $\sigma_w^{\mathtt{T}}$ is the *trend standard deviation* and the remaining parameters are the same with the bounds as defined in Equation 20. The full matrices for the new model are detailed in Appendix A. The initial parameter values in the original space for the model are

$$\boldsymbol{\theta}^0 = \left[\underbrace{10^{-4}}_{\sigma_w^{\mathtt{T}}}, \underbrace{0.8}_{\phi^{\mathtt{AR}}}, \underbrace{0.02}_{\sigma_w^{\mathtt{AR}}}, \underbrace{0.03}_{\sigma_v}\right]^{\mathsf{T}}.$$

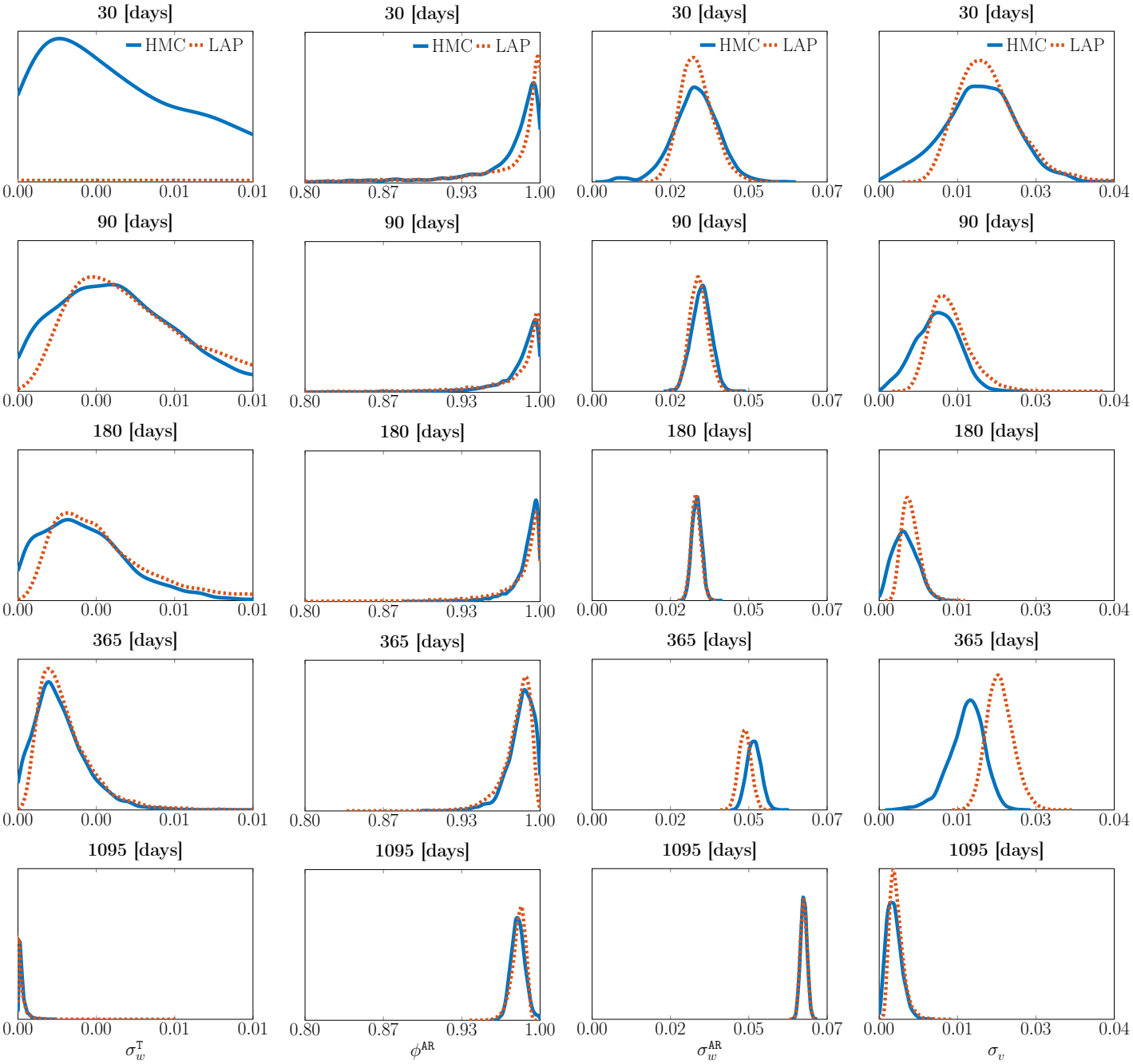
The other settings such as the transformation functions and the prior PDFs related to LAP-P and HMC-P in the new model remain identical as the previous model.

6.2.3 Results

As with the case involving simulated data, five tests with different training-set lengths (TSLs) have been carried out using both LAP-P and HMC-P for this case-study. The amount of data in each TSL is $30, 90, 180, 365$ and 1095 days correspond to $51, 161, 328, 651$ and 2142 data points, respectively. The convention for the figure remains identical as the simulated case presented in Section 6.1.3. The dashed and solid lines represent the Kernel smoothing function estimate of parameter posterior PDF obtained from LAP-P and HMC-P, respectively. For hidden state variables, the dashed line and the shaded region delimited by the solid line represent respectively the mean value $\hat{\boldsymbol{\mu}}_{t|t}^{\text{LAP}}$ and its uncertainty bound $\hat{\boldsymbol{\mu}}_{t|t}^{\text{LAP}} \pm \hat{\boldsymbol{\sigma}}_{t|t}^{\text{LAP}}$ at time t, whereas $\hat{\boldsymbol{\mu}}_{t|t}^{\text{HMC}}$ and $\hat{\boldsymbol{\mu}}_{t|t}^{\text{HMC}} \pm \hat{\boldsymbol{\sigma}}_{t|t}^{\text{HMC}}$ are illustrated by the solid line and the shaded region, correspondingly.

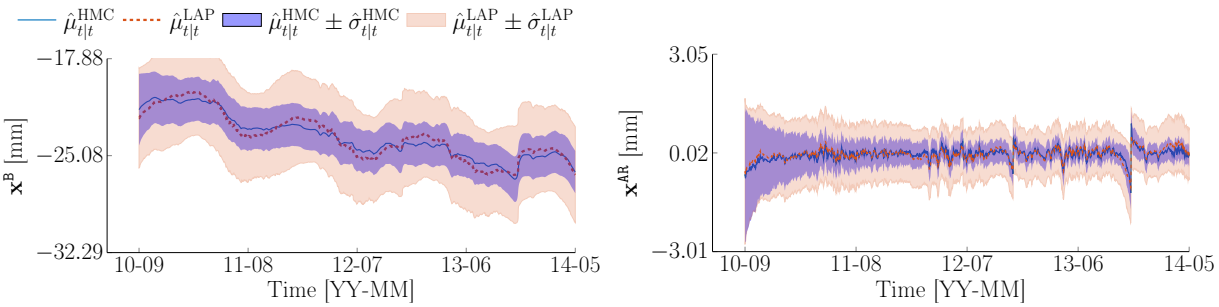
Figures 11 and 12 show the posterior PDFs for each parameter and the hidden state variables estimated using Kalman Smoother, respectively. A common remark is that the posterior PDF of model parameters approximated using both LAP-P and HMC-P tend to concentrate to the same values with the increasing size of the training data. For the TSL of 30 days, the posterior PDF for $\sigma_w^{\mathtt{T}}$ obtained from LAP-P, as illustrated in Figure 11a, is uniform for the range while those obtained from HMC-P yields reasonable inferences given this short TSL. An interesting behavior identified in the posterior PDFs for $\sigma_w^{\mathtt{AR}}$ and σ_v, is that there is a lack of consistency between these posterior PDFs regarding the TSL. Furthermore, the expected values of these posterior PDFs slightly change with respect to TLS. The time varying model parameters might be an explanation for this change. The posterior PDFs for $\phi^{\mathtt{AR}}$ presented in Figure 11b, again shows the dominance of the likelihood function over the prior for large dataset sizes.

Despite the discrepancy of the posterior PDFs for $\sigma_w^{\mathtt{AR}}$ and σ_v across training-set lengths, the inferences for the hidden state variables from both procedures are well behaved as expected. The autoregressive component $\mathbf{x}^{\mathtt{AR}}$ shows a stationary behavior with a small amplitude even though an abnormal peak with high amplitude is identified at the end of the year 2013. This jump is likely to be caused by the presence of a malfunction in the measurement sensor. The estimation accuracy improves with the dataset size. Their expected values tend to the same values and their uncertainty bounds are reduced as the TSL increases. The estimation results for the TSL of 1095 days presented in Figure 12e, outperform the others in the remaining TSLs. It confirms that the model parameter and state estimate uncertainties can be reduced through an increase in dataset size.

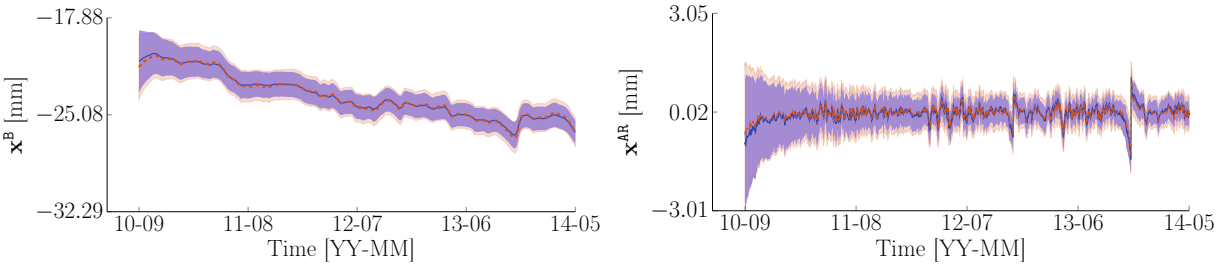


(a) $\sigma_w^{\mathtt{T}}$ (b) $\phi^{\mathtt{AR}}$ (c) $\sigma_w^{\mathtt{AR}}$ (d) σ_v

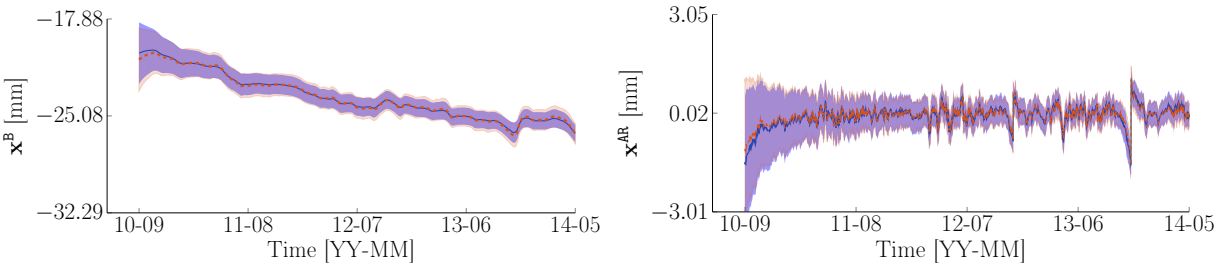
Figure 11: Illustration of the Kernel smoothing function estimate the posterior PDFs for each model parameter $p(\theta_i|\mathbf{y}_{1:\mathtt{T}})$ in the original space with respect to the training-set length. The data are collected on a dam in Canada.



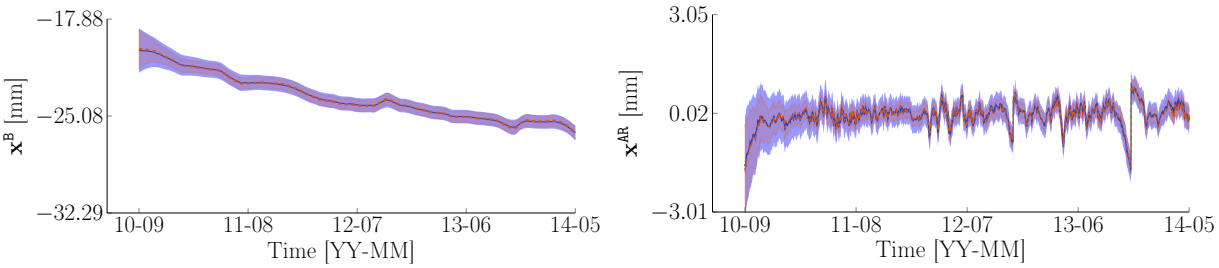
(a) Training set of 30 days (51 data points)



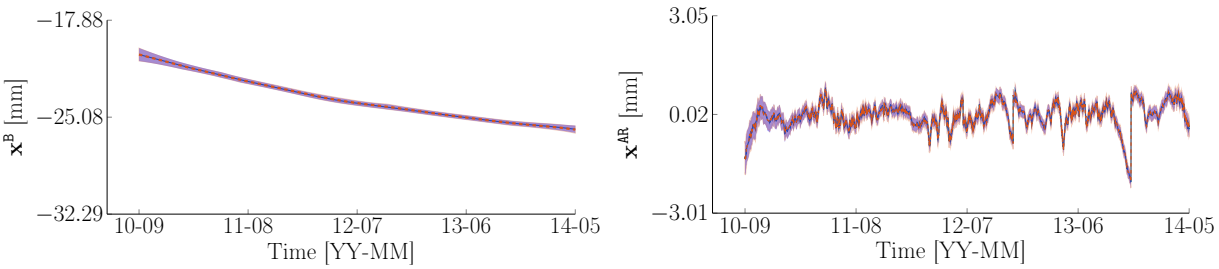
(b) Training set of 90 days (161 data points)



(c) Training set of 180 days (328 data points)



(d) Training set of 365 days (651 data points)



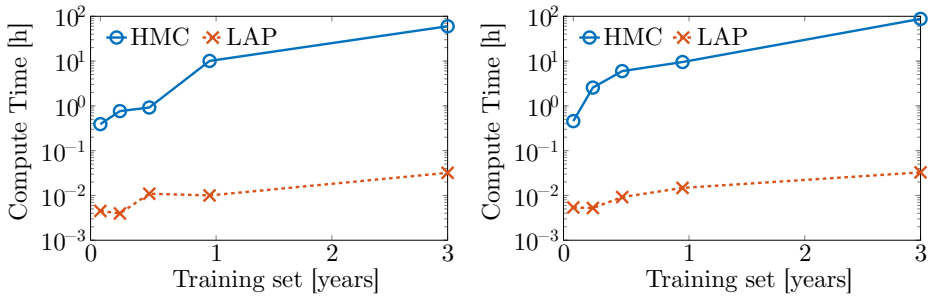
(e) Training set of 1095 days (2142 data points)

Figure 12: Expected value $\hat{\mu}$ and standard deviation $\hat{\sigma}$ for baseline (left) and autoregressive (right) components using LAP-P and HMC-P with respect to the training-set length. The data are collected on a dam in Canada.

HMC-P again shows its more reliable estimation capacity over LAP-P for the small dataset size such as 30 and 90, as presented in Figures 12a and b. Yet, the differences in the estimation between both LAP-P and HMC-P become unnoticeable for the TSLs of 180, 365, and 1095 days, as shown in Figures 12c, d and e.

7 Discussion

From the comparative studies, we are able to expose the advantages as well as the limitations of both LAP-P and HMC-P. More specifically, the LAP-P provides a fast method for the approximation of parameter posterior PDFs, yet it is prone to be trapped in a local maximum due to its sensitivity towards the selection of initial parameter values. In addition, the evaluation of the parameter covariance matrix depends not only on the structure of the log-posterior but also on the parameter being estimated. The accuracy and feasibility of such evaluation become challenging for either high-dimensional parameter spaces, or for a small dataset size. To ensure a reliable approximation, the model must be (a) trained with a large amount of data and (b) tested with the different sets of the initial parameter values. In the other hand, the HMC-P is less sensitive to the initial parameter values and provides more reliable estimation than LAP-P, especially when the amount of data in the training set is limited. However, the computational cost is much higher than the LAP-P. Figure 13 presents the compute time of both procedures for approximating the parameter posterior PDF in the simulated and real datasets.



(a) Simulated dataset (b) Real dataset

Figure 13: The compute time of LAP-P and HMC-P for approximating the parameter posterior PDF is presented in logarithmic scale.

Because LAP-P is a fast method for the estimation task and providing an equivalent performance as HMC-P in the case of the large size of the training data. The LAP-P is well suited for approximating the posterior PDF of model parameters as well as for the estimation of the hidden state variables during the model development. HMC-P should then be used for verifying the resulting estimations obtained from LAP-P. If the results estimated using both procedures are different from each other, the model construction must be carefully revised for consistency.

An interesting behavior identified in the case-study of the real dataset is that the model

parameters might be changed over time. Future work will investigate the possibility of estimating time varying model parameters.

8 Conclusion

This paper proposes the LAP and HMC-based procedures for quantifying the uncertainty for the model parameters as well as the hidden state variables in the existing BDLM framework. A comparative study has been conducted using a simulated dataset and a real dataset collected on a dam in Canada. The results show that the LAP-P is able to provide a reasonable estimation without requiring a high computation cost, yet it is prone to be trapped in a local maximum. Meanwhile, the estimation of HMC-P is more reliable than those of LAP-P, but it is computationally demanding. The estimation results obtained from both LAP-P and HMC-P converge to the same values when the training dataset is large. Therefore, we suggest a deployment of both LAP-P and HMC-P in order to ensure a reliable estimation. More specifically, LAP-P is first employed for the model development. HMC-P should then be used to verify the resulting estimation obtained using LAP-P.

Acknowledgements

This project is funded by the Natural Sciences and Engineering Research Council of Canada (NSERC, RGPIN-2016-06405). The authors acknowledge the contribution of Hydro-Quebec who provided the dataset employed in this research. More specifically, the authors thank Benjamin Miquel and Patrice Côté from Hydro-Quebec for their help in the project.

References

[1] S.-K. Au. *Bayesian Inference*, chapter 8, pages 265–289. Springer Singapore, 2017.

[2] J. L. Beck. Bayesian system identification based on probability logic. *Structural Control and Health Monitoring*, 17(7):825–847, 2010.

[3] M. Betancourt. A conceptual introduction to hamiltonian monte carlo. *arXiv preprint arXiv:1701.02434*, 2017.

[4] A. W. Bowman and A. Azzalini. *Applied smoothing techniques for data analysis: the kernel approach with S-Plus illustrations*, volume 18. OUP Oxford, New York, 1997.

[5] S. Brooks, A. Gelman, G. Jones, and X.-L. Meng. *Handbook of Markov Chain Monte Carlo*. CRC press, 2011.

[6] O. Cappé, E. Moulines, and T. Ryden. *Inference in Hidden Markov Models (Springer Series in Statistics)*. Springer-Verlag New York, Inc., 2005.

[7] S. H. Cheung and J. L. Beck. Bayesian model updating using hybrid monte carlo simulation with application to structural dynamic models with many uncertain parameters. *Journal of Engineering Mechanics*, 135(4):243–255, 2009.

[8] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Series B Methodological*, 39(1):1–38, 1977.

[9] S. Duane, A. D. Kennedy, B. J. Pendleton, and D. Roweth. Hybrid monte carlo. *Physics letters B*, 195(2):216–222, 1987.

[10] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian data analysis*. CRC Press, 3 edition, 2014.

[11] A. Gelman, D. Lee, and J. Guo. Stan: A probabilistic programming language for bayesian inference and optimization. *Journal of Educational and Behavioral Statistics*, 40(5):530–543, 2015.

[12] A. Gelman and D. B. Rubin. Inference from iterative simulation using multiple sequences. *Statistical science*, 7(4):457–472, 1992.

[13] A. Gelman, D. Simpson, and M. Betancourt. The prior can often only be understood in the context of the likelihood. *Entropy*, 19(10):555, 2017.

[14] J. Geweke and H. Tanizaki. Bayesian estimation of state-space models using the Metropolis-Hastings algorithm within Gibbs sampling. *Computational Statistics & Data Analysis*, 37(2):151 – 170, 2001.

[15] J.-A. Goulet. Bayesian dynamic linear models for structural health monitoring. *Structural Control and Health Monitoring*, 24:e2035–n/a, 2017.

[16] J.-A. Goulet and K. Koo. Empirical validation of bayesian dynamic linear models in the context of structural health monitoring. *Journal of Bridge Engineering*, 23(2):05017017, 2018.

[17] P. L. Green, K. Worden, M. Street, and S. Sheffield. Bayesian and Markov chain Monte Carlo methods for identifying nonlinear systems in the presence of uncertainty. *Philosophical Transactions Royal Society A*, 373, 2015.

[18] H. Haario, E. Saksman, and J. Tamminen. An adaptive Metropolis algorithm. *Bernoulli*, 7(2):223–242, 2001.

[19] W.K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97, 1970.

[20] M. D. Hoffman and A. Gelman. The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. *Journal of Machine Learning Research*, 15(1):1593–1623, 2014.

[21] S. S. Jang and R. B. Gopaluni. Parameter estimation in nonlinear chemical and biological processes with unmeasured variables from small data sets. *Chemical Engineering Science*, 66(12):2774 – 2787, 2011.

[22] A. Kramer, B. Calderhead, and N. Radde. Hamiltonian monte carlo methods for efficient parameter estimation in steady state dynamical systems. *BMC Bioinformatics*, 15(1), 2014.

[23] H.-F. Lam, H. Yang, J, and S.-K. Au. Markov chain monte carlo-based bayesian method for structural model updating and damage detection. *Structural Control and Health Monitoring*, pages e2140–n/a, 2018.

[24] S. L. Lauritzen. *Graphical models*, volume 17. Clarendon Press, 1996.

[25] L. Le Cam. Maximum Likelihood: An introduction. *Internat. Statist. Rev.*, 58:153–171, 1990.

[26] I. S. Mbalawata, S. Särkkä, and H. Haario. Parameter estimation in stochastic differential equations with Markov chain Monte Carlo and non-linear Kalman filtering. *Computational Statistics*, 28(3):1195–1223, 2013.

[27] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *J. Chem. Phys.*, 21:1087–1091, 1953.

[28] N. Metropolis and S. Ulam. The Monte Carlo method. *Journal of the American Statistical Association*, 44(247):335–341, 1949.

[29] K. P. Murphy. Switching kalman filters. Technical report, Citeseer, 1998.

[30] K. P. Murphy. *Machine learning: a probabilistic perspective*. The MIT Press, 2012.

[31] R. M. Neal et al. Mcmc using hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, 2(11):113–160, 2011.

[32] Y. Nesterov. Primal-dual subgradient methods for convex problems. *Mathematical programming*, 120(1):221–259, 2009.

[33] L. H. Nguyen and J.-A. Goulet. Anomaly detection with the switching kalman filter for structural health monitoring. *Structural Control and Health Monitoring*, pages e2136–n/a, 2018.

[34] B. Ninness and S. Henriksen. Bayesian system identification via Markov chain Monte Carlo techniques. *Automatica*, 46(1):40–51, 2010.

[35] B. Rannala. Identifiability of parameters in mcmc bayesian inference of phylogeny. *Systematic Biology*, 51(5):754–760, 2002.

[36] J. S. Rosenthal. Parallel computing and Monte Carlo algorithms. *Far East Journal of Theoretical Statistics*, 4:207–236, 2000.

[37] S. Särkkä. *Bayesian Filtering and Smoothing*. Cambridge University Press, 2013.

[38] R. H. Shumway and D. S. Stoffer. An approach to time series smoothing and forecasting using the em algorithm. *Journal of Time Series Analysis*, 3(4):253–264, 1982.

[39] D. Simon. *Optimal state estimation: Kalman, H infinity, and nonlinear approaches*. John Wiley & Sons Inc., New Jersey, 2006.

[40] M. West and J. Harrison. *Bayesian Forecasting and Dynamic Models*. Springer Series in Statistics. Springer New York, 1999.

[41] K. Yuan, M. Girolami, and Niranjan M. Markov Chain Monte Carlo Methods for State-Space Models with Point Process Observations. *Neural Computation*, 24(6):1462–1486, 2012.

[42] Z. Y. Zhao, M. Xie, and M. West. Dynamic dependence networks: Financial time series forecasting and portfolio decisions. *Appl. Stoch. Model. Bus. Ind.*, 32(3):311–332, May 2016.

Appendix A

The *transition matrix* ($\mathbf{A}_t$), the *observation matrix* ($\mathbf{C}_t$), the *observation error covariance matrix* ($\mathbf{R}_t$), and the *model error covariance matrix* ($\mathbf{Q}_t$) for the simulated-dataset model class and the real-dataset model are defined following

Simulated-dataset model

$$\begin{aligned}
\mathbf{A}_t &= \text{block diag}\left(1, \begin{bmatrix} \cos\omega^{\mathtt{T1}} & \sin\omega^{\mathtt{T1}} \\ -\sin\omega^{\mathtt{T1}} & \cos\omega^{\mathtt{T1}} \end{bmatrix}, \phi^{\mathtt{AR}}\right) \\
\mathbf{C}_t &= [1, 1, 0, 1] \\
\mathbf{R}_t &= \left[(\sigma_v)^2\right] \\
\mathbf{Q}_t &= \text{block diag}\left(\left(\sigma_w^{\mathtt{B}}\right)^2, \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}, \left(\sigma_w^{\mathtt{AR}}\right)^2\right)
\end{aligned}$$

Real-dataset model

$$\begin{aligned}
\mathbf{A}_t &= \text{block diag}\left(\begin{bmatrix} 1 & \Delta t \\ 0 & 1 \end{bmatrix}, \begin{bmatrix} \cos\omega^{\mathtt{T1}} & \sin\omega^{\mathtt{T1}} \\ -\sin\omega^{\mathtt{T1}} & \cos\omega^{\mathtt{T1}} \end{bmatrix}, \phi^{\mathtt{AR}}\right) \\
\mathbf{C}_t &= [1, 0, 1, 0, 1] \\
\mathbf{R}_t &= \left[(\sigma_v)^2\right] \\
\mathbf{Q}_t &= \text{block diag}\left(\left(\sigma_w^{\mathtt{T}}\right)^2 \begin{bmatrix} \frac{\Delta t^3}{3} & \frac{\Delta t^2}{2} \\ \frac{\Delta t^2}{2} & \Delta t \end{bmatrix}, \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}, \left(\sigma_w^{\mathtt{AR}}\right)^2\right)
\end{aligned}$$

where Δt is the time step at the time t.