Bayesian Inference of Discretization Error Means in ODEs via Ensemble Kalman Filtering

Shoji Toyota — *Department of Advanced Information Technology, Kyushu University, Japan*
Yuto Miyatake — *D3 Center, The University of Osaka, Japan*

Abstract

We propose a Bayesian framework to quantify discretization errors in numerical solutions of ODE models based on observational data. The discretization error is modeled as a random variable, and its mean—referred to as the discretization error mean—is inferred from the observations. By introducing a Markov prior on the temporal evolution of the discretization error mean, we formulate the problem as a state-space model with a linear Gaussian observation process, which enables efficient inference via the Ensemble Kalman Filter. We also propose a specific form of a Markov prior motivated by classical discretization error analysis, in which global errors accumulate from local errors. The proposed prior depends on the step size of a numerical solver, and we establish its convergence rate in probability as the step size tends to zero. Numerical experiments on the pendulum system and the FitzHugh–Nagumo model demonstrate the effectiveness of the proposed approach.

Proceedings of the 2nd International Conference on Probabilistic Numerics (ProbNum) 2026, Lappeenranta, Finland. PMLR Volume 341.

1 Introduction

We consider the problem of finding a solution to a $d_{\mathcal{X}}$-dimensional ordinary differential equation (ODE)

$$\frac{dx(t)}{dt} = f\big(x(t)\big), \tag{1}$$

where the solution $x(t)$ takes values in $\mathbb{R}^{d_{\mathcal{X}}}$ and $f : \mathbb{R}^{d_{\mathcal{X}}} \to \mathbb{R}^{d_{\mathcal{X}}}$ is a continuous map. In the general case where f is nonlinear, a closed-form solution is typically unavailable. A classical approach to this problem is to discretize the ODE using numerical solvers, such as the Euler method or Runge–Kutta methods. Such discretization introduces numerical errors, often referred to as *discretization errors*. Understanding the behavior of these errors is a central topic in numerical analysis (Butcher, 2016; Hairer et al., 1993; Iserles, 2008). The main focus is typically on analyzing its asymptotic behavior as the step size approaches zero, or on deriving bounds on the discretization error.

In contrast, methods for directly quantifying discretization error have become increasingly important. A typical example arises in the estimation of parameters in ODE models from observational data. Practical approaches, such as simulation-based inference (Cranmer et al., 2020) or four-dimensional variational data assimilation (4D-Var) (Asch et al., 2016), often rely on numerical solvers. In these methods, a parameter is chosen so that the resulting numerical solution fits the observations, implicitly assuming sufficient numerical accuracy. However, this assumption may be violated in certain settings, such as chaotic systems or large-scale problems (Conrad et al., 2017). In such cases, discretization error must be explicitly quantified and properly incorporated into the estimation procedure.

Motivated by the need to quantify discretization error, recent research has approached this problem from statistical and machine learning perspectives. Prominent examples include Bayesian ODE solvers (Beck et al., 2024; Kersting et al., 2020; Le Fay et al., 2025; Schmidt et al., 2021; Schober et al., 2018; Tronarp et al., 2022, 2019, 2021; Yao et al., 2025; Krämer, 2025), which formulate ODE discretization as Gaussian process inference. Another important direction is perturbation-based methods (Abdulle and Garegnani, 2020; Conrad et al., 2017; Lie et al., 2022, 2019), where stochastic perturbations are introduced into numerical solutions to represent uncertainty. These developments have led to an interdisciplinary field at the interface of numerical analysis and statistics, known as *Probabilistic Numerics*.

Another line of research on quantifying discretization error in ODEs is known as the *discretization error variance* approach (Marumo et al., 2024; Matsuda and Miyatake, 2021; Miyatake et al., 2025; Toyota and Miyatake, 2025). Within this framework, conventional numerical solvers are used to obtain numerical solutions, while the resulting discretization errors are modeled as random variables. The associated variances, referred to as *discretization error variances*, are regarded as statistical parameters and estimated from observations. In contrast to Bayesian ODE solvers and perturbation-based methods, this approach directly infers the discretization error itself based on observations.

Motivated by the discretization error variance approach, we propose a framework for discretization error quantification, which we term the discretization error *mean* approach. In this setting, the discretization error is modeled as a random variable, and its *mean* is estimated from observational data. By focusing on the mean, the proposed method enables prediction of the direction of the discretization error, whereas approaches based on the variance permit inference only about its absolute magnitude. Inference is carried out within a Bayesian framework, where a prior is placed on the discretization

error mean and the corresponding posterior is derived.

A key component of the proposed method is Bayesian inference of the discretization error mean using the Ensemble Kalman Filter (EnKF), a standard technique in data assimilation (Burgers et al., 1998; Evensen, 2009). By imposing a Markov prior on the discretization error means, the inference problem can be formulated as a state-space model whose observation process is linear Gaussian, to which the EnKF can be applied.

The specification of a Markov prior plays a central role in the effectiveness of the proposed framework. In previous work, Toyota and Miyatake (2025) proposed a Markov prior on discretization error variances based on insights from classical discretization error analysis for ODEs, namely that the global discretization error can be interpreted as the accumulation of local errors generated at each time step. We show that the prior introduced in Toyota and Miyatake (2025) can be naturally extended to the discretization error *mean*. The prior depends explicitly on the step size of the numerical solver; we analyze its asymptotic behavior as the step size tends to zero, following arguments analogous to those developed in Toyota and Miyatake (2025).

The main contributions of this paper are as follows:

- We introduce a statistical framework for quantifying discretization error in ODEs, termed the *discretization error mean* approach, which enables inference on the discretization error from observational data, including its direction in addition to its magnitude.
- We formulate the inference problem as a state-space model by imposing a Markov prior on the time evolution of the discretization error mean, thereby enabling Bayesian inference via the Ensemble Kalman Filter.
- We propose a Markov prior for the discretization error mean by extending the approach of Toyota and Miyatake (2025), and examine its asymptotic behavior as the step size of the numerical solver tends to zero.
- Through numerical experiments, we demonstrate that the proposed method accurately quantifies both the magnitudes and directions of discretization errors.

2 Proposed Method

2.1 Problem Setting

We first describe the problem setting addressed in this paper. For a $d_{\mathcal{X}}$-dimensional ODE (1), let its exact solution be denoted by $x(t)$. In this study, we numerically solve the equation at discrete time points $t_1, \dots, t_N$, and denote the resulting numerical solutions by $x_{t_1}, \dots, x_{t_N}$.

The goal of this study is to quantify the discretization error $x(t_i) - x_{t_i}$. While its behavior has been a major subject of theoretical study in numerical analysis (Butcher, 2016; Hairer et al., 1993; Iserles, 2008), typical analyses only derive bounds on the error or study its asymptotic properties as the step size approaches zero, without providing a direct quantification of the discretization error.

In this paper, we quantify the discretization error $x(t_i) - x_{t_i}$ using time-series observations $y^*_{t_1}, \dots, y^*_{t_N}$ at discrete time points $t_1, \dots, t_N$. These observations are assumed to be generated by the process

$$y^*_{t_i} = Hx(t_i) + \varepsilon_{t_i}, \tag{2}$$

where H is a full-rank linear observation matrix, and the observation noise $\varepsilon_{t_i} \sim \mathcal{N}(0, \Gamma)$ is Gaussian with zero mean and a positive definite covariance matrix Γ.

2.2 Discretization Error Mean

To accomplish our objective, we propose the *discretization error mean* approach. In this framework, we model the discretization error $x(t_i) - x_{t_i}$, the difference between the true solution $x(t_i)$ and the numerical solution x_{t_i}, as a Gaussian random variable:

$$x(t_i) - x_{t_i} \sim \mathcal{N}(\mu_{t_i}, \Sigma), \tag{3}$$

or equivalently,

$$x(t_i) \sim \mathcal{N}(x_{t_i} + \mu_{t_i}, \Sigma), \tag{4}$$

where Σ is a fixed covariance matrix, and μ_{t_i} denotes the mean, which is estimated from observations and referred to as the *discretization error mean*. The discretization error mean is estimated using the likelihood $p(y_{t_1}, ..., y_{t_N} \mid \mu_{t_0}, ..., \mu_{t_N})$. Here, under the models in (2) and (4), the likelihood can be expressed as

$$p(y_{t_1}, ..., y_{t_N} | \mu_{t_0}, ..., \mu_{t_N}) = \prod_{i=1}^{N} p(y_{t_i} | \mu_{t_i}),$$

where $p(y_{t_i} | \mu_{t_i})$ is defined by

$$p(y_{t_i} \mid \mu_{t_i}) = \mathcal{N}\left(y_{t_i};\, H(x_{t_i} + \mu_{t_i}),\, H\Sigma H^\top + \Gamma\right), \tag{5}$$

a normal distribution with its mean and covariance given by $H(x_{t_i} + \mu_{t_i})$ and $H\Sigma H^\top + \Gamma$ respectively. By utilizing this likelihood, we can estimate the discretization error mean μ_{t_i}, enabling us to quantify the discretization error from observations $y^*_{t_1}, ..., y^*_{t_N}$.

Remark 1. Modeling the discretization error as a random variable and estimating its moments from observations was originally proposed by Matsuda and Miyatake (2021) and has been further developed in subsequent works (Miyatake et al., 2025; Marumo et al., 2024; Toyota and Miyatake, 2025). In this framework, the discretization error is modeled as

$$x(t_i) - x_{t_i} \sim \mathcal{N}(0, \Sigma), \tag{6}$$

where the covariance matrix Σ, referred to as the *discretization error variance*, is treated as a statistical quantity to be inferred from observations, while the mean is fixed at zero for all t_i. This formulation allows one to impose a monotonicity constraint on Σ_{t_i} by treating the discretization error variances as time-varying quantities, which may be difficult to incorporate in our approach based on modeling the mean of the discretization error. On the other hand, our discretization error mean approach enables us to infer the direction of the error and thereby adjust the numerical solution toward the exact solution. Another advantage will become apparent in the state-space formulation described in the following subsection.

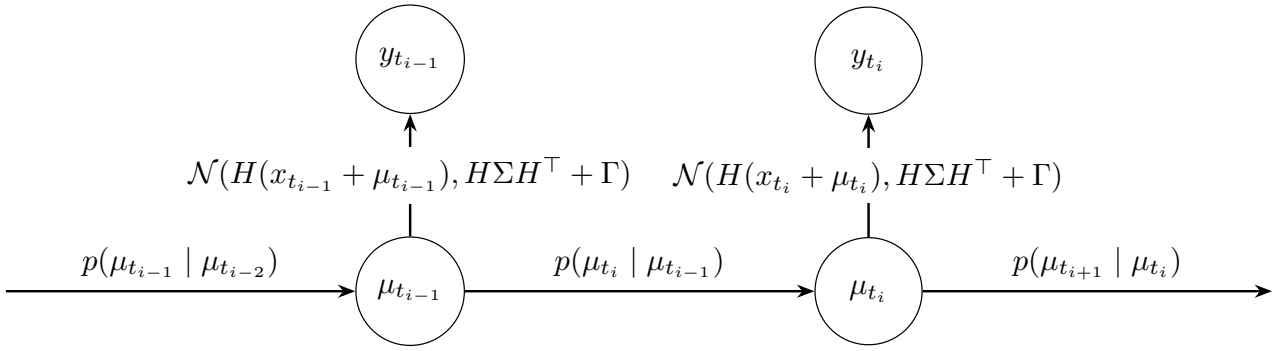


Figure 1: Reformulation of the objective in (7) as a state-space model. The latent transition follows the Markov prior $p(\mu_{t_{i+1}} \mid \mu_{t_i})$, and the observation model is $\mathcal{N}\big(H(x_{t_i} + \mu_{t_i}),\ H\Sigma H^\top + \Gamma\big)$. Equivalently, the observation equation can be written as $y_t = H\mu_t + \varepsilon_t$ with $\varepsilon_t \sim \mathcal{N}\big(Hx_t,\ H\Sigma H^\top + \Gamma\big)$, which allows the use of the Ensemble Kalman Filter.

2.3 State-Space Model Interpretation of Discretization Error Mean Estimation

We estimate the discretization error mean μ_{t_i} from observations $y^*_{t_1}, \ldots, y^*_{t_N}$ in a Bayesian manner. Specifically, we place a prior $p(\mu_{t_0}, \ldots, \mu_{t_N})$ on the time series of discretization error means and evaluate the posterior:

$$\begin{aligned} &p(\mu_{t_0}, \ldots, \mu_{t_N} \mid y^*_{t_1}, \ldots, y^*_{t_N}) \\ &\propto p(y^*_{t_1}, \ldots, y^*_{t_N} \mid \mu_{t_0}, \ldots, \mu_{t_N}) p(\mu_{t_0}, \ldots, \mu_{t_N}). \end{aligned} \tag{7}$$

In this study, we adopt a Markov prior for the time series of discretization error means:

$$p(\mu_{t_0}, \ldots, \mu_{t_N}) = p(\mu_{t_0}) \prod_{i=0}^{N-1} p(\mu_{t_{i+1}} \mid \mu_{t_i}). \tag{8}$$

Under the Markov assumption, the discretization error mean μ_{t_i} and the observations $y^*_{t_i}$ form a state-space model (Fig. 1). The dynamics of the latent state are described by the temporal evolution of the discretization error mean, $\mu_{t_i} \to \mu_{t_{i+1}}$, with a stochastic transition governed by $p(\mu_{t_{i+1}} \mid \mu_{t_i})$. The observation process is specified by the likelihood $p(y_{t_i} \mid \mu_{t_i})$, as defined in (5).

Under the state-space formulation, evaluating the posterior (7) reduces to latent state estimation for a given state-space model. Among the available approaches, we employ the Ensemble Kalman Filter (EnKF) (Evensen, 2009). The EnKF can be applied when the observation model of the target state-space model, $p(y_{t_i} \mid \mu_{t_i})$, can be expressed as a linear transformation of μ_{t_i} with additive Gaussian noise. In our case, the observation process can be written as

$$y_{t_i} = H\mu_{t_i} + \varepsilon_{t_i}, \tag{9}$$

where $\varepsilon_{t_i} \sim \mathcal{N}\left(Hx_{t_i},\ H\Sigma H^\top + \Gamma\right)$,[1] enabling us to apply EnKF to our inference.

Remark 2. An advantage of the discretization error mean approach, compared with the variance approach, is that it can be reduced to inference in a state-space model with a *linear Gaussian* observation process (9). In the context of the discretization error variance approach, Toyota and Miyatake (2025) impose a Markov prior on the discretization error variance and perform Bayesian inference for this variance by reducing the problem to a state-space model. However, since the associated observation process cannot be expressed in linear Gaussian form, inference is carried out using a particle filter. It is known that particle filters tend to perform poorly in high-dimensional systems compared with the EnKF, and the EnKF is often preferable in such settings. Therefore, the ability to directly leverage the EnKF constitutes an important advantage of modeling the mean of the discretization error rather than its variance.

Hereafter, the tuples $\{y_{t_1}, \ldots, y_{t_i}\}$ and $\{\mu_{t_0}, \ldots, \mu_{t_i}\}$ are abbreviated as $y_{1:i}$ and $\mu_{0:i}$, respectively. In the EnKF, the distribution $p(\mu_{t_i} \mid y^*_{1:i})$ is approximated by

$$p(\mu_{t_i} \mid y^*_{1:i}) \approx \frac{1}{N_e} \sum_{n=1}^{N_e} \delta_{\hat{\mu}^n_{i|i}},$$

i.e., an ensemble of discrete points $\{\hat{\mu}^n_{i|i}\}_{n=1}^{N_e}$. From the ensemble $\{\hat{\mu}^n_{i|i}\}_{n=1}^{N_e}$ of $p(\mu_{t_i} \mid y^*_{1:i})$, we construct the ensemble $\{\hat{\mu}^n_{i+1|i+1}\}_{n=1}^{N_e}$ at the next time t_{i+1} through the two steps described below, referred to as the *prediction step* (Subsection 2.3.1) and the *correction step* (Subsection 2.3.2). Finally, a smoothing step is applied to obtain the posterior $p(\mu_{0:N} \mid y^*_{1:N})$, incorporating all observations (Subsection 2.3.3).

2.3.1 Prediction Step

Assume that we have an ensemble $\{\hat{\mu}^n_{i|i}\}_{n=1}^{N_e}$ that approximates the distribution $p(\mu_{t_i} \mid y^*_{1:i})$. In the prediction step, we construct a new ensemble $\{\hat{\mu}^n_{i+1|i}\}_{n=1}^{Ne}$ to approximate

$$p(\mu_{t_{i+1}} \mid y^*_{1:i}) = \int p(\mu_{t_{i+1}} | \mu_{t_i}) p(\mu_{t_i} \mid y^*_{1:i}) d\mu_{t_i}.$$

This is achieved by propagating each ensemble member forward through the transition distribution $p(\mu_{i+1} \mid \mu_i)$. Specifically, for $n = 1, \ldots, N_e$, we simulate

$$\hat{\mu}^n_{i+1|i} \sim p\big(\mu_{i+1} \mid \mu_i = \hat{\mu}^n_{i|i}\big).$$

[1] While EnKF typically assumes that the observation noise ε_{t_i} is Gaussian with zero mean, it remains applicable even when the noise has a non-zero mean, without any essential modification, as shown in the following subsections.

Algorithm 1 Ensemble Kalman Filter for Discretization Error Quantification in an ODE Models

Input: Prior $p(\mu_{t_0})\prod_{i=0}^{N-1} p(\mu_{t_{i+1}}|\mu_{t_i})$, observations $\{y^*_{t_i}\}_{i=1}^{N}$, ensemble size N_e, lag length L.
An ODE $\frac{dx(t)}{dt} = f(x(t))$ and a numerical solver.
Generate an initial ensemble $\{\hat{\mu}^n_{0|0}\}_{n=1}^{N_e} \sim p(\mu_{t_0})$
for $i = 0, \ldots, N-1$ **do**
  for $n = 1, \ldots, N_e$ **do** ▷ Prediction Step (Subsection 2.3.1)
    Simulate $\hat{\mu}^n_{i+1|i} \sim p(\mu_{t_{i+1}} \mid \mu_{t_i} = \hat{\mu}^n_{i|i})$
  Compute sample mean $\hat{\mu}_{i+1|i}$ and sample covariance $\hat{\Sigma}_{i+1|i}$ from $\{\hat{\mu}^n_{i+1|i}\}_{n=1}^{N_e}$
  for $j = \max(i+1-L, 0), \ldots, i+1$ **do**
    ▷ Smoothing is applied over $[0,\, i+1]$ if $i+1 \le L$, and over $[i+1-L,\, i+1]$ otherwise.
    Compute cross-covariance $\hat{\Sigma}_{j,i+1|i}$ from $\{\hat{\mu}^n_{i+1|i}\}_{n=1}^{N_e}$ and $\{\hat{\mu}^n_{j|i}\}_{n=1}^{N_e}$.
    Compute Kalman gain $\hat{K}_{j|i} = \hat{\Sigma}_{j,i+1|i}H^\top \left(H\hat{\Sigma}_{i+1,i+1|i}H^\top + H\Sigma H^\top + \Gamma\right)^{-1}$.
    for $n = 1, \ldots, N_e$ **do**
      Draw $\hat{w}^n_j \sim \mathcal{N}(0, H\Sigma H^\top + \Gamma)$
      Update $\hat{\mu}^n_{j|i+1} := \hat{\mu}^n_{j|i} + \hat{K}_{j|i}\left(y^*_{t_{i+1}} + \hat{w}^n_j - H\left(x_{t_{i+1}} + \hat{\mu}^n_{i+1|i}\right)\right)$.
        ▷ Correction Step (Subsection 2.3.2) with fixed lag smoothing (Subsection 2.3.3).
return $\{\hat{\mu}^n_{1|1+L}\}_{n=1}^{N_e}, \{\hat{\mu}^n_{2|2+L}\}_{n=1}^{N_e}, \ldots, \{\hat{\mu}^n_{N-L|N}\}_{n=1}^{N_e}, \{\hat{\mu}^n_{N-L+1|N}\}_{n=1}^{N_e}, \ldots, \{\hat{\mu}^n_{N|N}\}_{n=1}^{N_e}$,
ensembles $p(\mu_1|y^*_{1:1+L}), p(\mu_2|y^*_{1:2+L}), \ldots, p(\mu_{N-L}|y^*_{1:N}), p(\mu_{N-L+1}|y^*_{1:N}), \ldots, p(\mu_N|y^*_{1:N})$.

2.3.2 Correction Step

In this step, we construct an ensemble $\{\hat{\mu}^n_{i+1|i+1}\}_{n=1}^{N_e}$ that approximates the posterior $p(\mu_{t_{i+1}} \mid y^*_{1:i+1})$, given the ensemble $\{\hat{\mu}^n_{i+1|i}\}_{n=1}^{N_e}$ obtained in the prediction step.

In this step, we first compute the covariance matrix $\hat{\Sigma}_{i+1|i}$ from the ensemble $\{\hat{\mu}^n_{i+1|i}\}_{n=1}^{N_e}$ as

$$\hat{\Sigma}_{i+1|i} := \frac{1}{N_e - 1}\sum_{n=1}^{N_e}\left(\hat{\mu}^n_{i+1|i} - \hat{\mu}_{i+1|i}\right)\left(\hat{\mu}^n_{i+1|i} - \hat{\mu}_{i+1|i}\right)^\top, \tag{10}$$

where $\hat{\mu}_{i+1|i} := \frac{1}{N_e}\sum_{n=1}^{N_e}\hat{\mu}^n_{i+1|i}$ denotes the sample mean. With this covariance matrix $\hat{\Sigma}_{i+1|i}$, we compute

$$\hat{K}_{i+1|i} := \hat{\Sigma}_{i+1|i}H^\top\left(H\hat{\Sigma}_{i+1|i}H^\top + H\Sigma H^\top + \Gamma\right)^{-1}, \tag{11}$$

which is referred to as the Kalman gain. Finally, we obtain an ensemble approximating $p(\mu_{t_{i+1}} \mid y^*_{1:i+1})$ by *correcting* each ensemble member $\hat{\mu}^n_{i+1|i}$ using the Kalman gain $\hat{K}_{i+1|i}$ and the new observation $y^*_{t_{i+1}}$:

$$\hat{\mu}^n_{i+1|i+1} := \hat{\mu}^n_{i+1|i} + \hat{K}_{i+1|i}\left(y^*_{t_{i+1}} + \hat{w}^n_{i+1} - H\left(x_{t_{i+1}} + \hat{\mu}^n_{i+1|i}\right)\right), \tag{12}$$

where $\hat{w}^n_{i+1}$ are i.i.d. samples from $\mathcal{N}\big(0,\, H\Sigma H^\top + \Gamma\big)$.

The following theorem ensures that the first and second moments of the updated ensemble (12) coincide with those of $p(\mu_{t_{i+1}} \mid y^*_{1:i+1})$ in the limit as $N_e \to \infty$, under the Gaussian approximation of $p(\mu_{t_{i+1}} \mid y^*_{1:i+1})$. A proof is provided in Appendix A.

Theorem 2.1. *Assume that $p(\mu_{t_{i+1}} \mid y^*_{1:i})$ is a Gaussian distribution $\mathcal{N}(\mu_{i+1|i}, \Sigma_{i+1|i})$ and that $\{\hat{\mu}^n_{i+1|i}\}_{n=1}^{N_e}$ are i.i.d. samples from this distribution. Then, the sample mean $\hat{\mu}_{i+1|i+1}$ and sample covariance $\hat{\Sigma}_{i+1|i+1}$ of the ensemble $\{\hat{\mu}^n_{i+1|i+1}\}_{n=1}^{N_e}$ defined by (12) converge in probability to the true mean $\mu_{i+1|i+1}$ and covariance $\Sigma_{i+1|i+1}$ of $p(\mu_{t_{i+1}} \mid y^*_{1:i+1})$ as $N_e \to \infty$, i.e.,*

$$\hat{\mu}_{i+1|i+1} \xrightarrow{P} \mu_{i+1|i+1}, \quad \hat{\Sigma}_{i+1|i+1} \xrightarrow{P} \Sigma_{i+1|i+1}.$$

2.3.3 Smoothing

The prediction and correction steps yield $p(\mu_{t_i} \mid y^*_{1:i})$ for $0 \le i \le N$. However, our goal is to evaluate the distribution $p(\mu_{t_i} \mid y^*_{1:N})$, the posterior conditioned on all observations at $t_1, \ldots, t_N$. While fixed-interval smoothing and other more advanced smoothing variants (e.g., Krämer (2025)) are also available, we focus on fixed-lag smoothing, where $p(\mu_{t_i} \mid y^*_{1:i+L})$ is used as an approximation of $p(\mu_{t_i} \mid y^*_{1:N})$.

In fixed-lag smoothing, we introduce the augmented state

$$\boldsymbol{\mu}_{t_i} = \left(\mu_{t_i}, \mu_{t_{i-1}}, \ldots, \mu_{t_{i-L}}\right)^\top.$$

The dynamics of the augmented state from $\boldsymbol{\mu}_{t_i}$ to

$$\boldsymbol{\mu}_{t_{i+1}} = \left(\mu_{t_{i+1}}, \mu_{t_i}, \ldots, \mu_{t_{i+1-L}}\right)^\top$$

are given as follows: the first component of $\boldsymbol{\mu}_{t_{i+1}}$ is assumed to follow $\mu_{t_{i+1}} \sim p\left(\mu_{t_{i+1}} \mid \mu_{t_i}\right)$, and the remaining components are deterministically shifted from

$$\left(\mu_{t_i}, \mu_{t_{i-1}}, \ldots, \mu_{t_{i+1-L}}\right)^\top,$$

which correspond to the first through L-th components of $\boldsymbol{\mu}_{t_i}$. By defining the observation model as

$$p(y_{t_i} \mid \boldsymbol{\mu}_{t_i}) = p(y_{t_i} \mid \mu_{t_i}),$$

we obtain an augmented state-space model for $\boldsymbol{\mu}_{t_i}$ and y_{t_i}. By solving this augmented state-space model with

the EnKF, we obtain an ensemble approximation of $p(\boldsymbol{\mu}_{t_i} \mid y^*_{1:i})$. By extracting the final component of each ensemble member, we obtain an estimate of $p(\mu_{t_{i-L}} \mid y^*_{1:i})$, where observations up to L time steps ahead are incorporated into the estimation of $\mu_{t_{i-L}}$.

For an ensemble of $p(\boldsymbol{\mu}_{t_{i+1}} \mid y^*_{1:i})$ denoted by

$$\left\{\hat{\boldsymbol{\mu}}^n_{i+1|i}\right\} = \left\{\left(\hat{\mu}^n_{i+1|i}, \ldots, \hat{\mu}^n_{i+1-L|i}\right)^\top\right\}_{n=1}^{N_e},$$

each ensemble member is updated via the fixed-lag smoothing step as

$$\begin{aligned}&\hat{\mu}^n_{j|i+1}\\ &:= \hat{\mu}^n_{j|i} + \hat{K}_{j|i}\left(y^*_{t_{i+1}} + \hat{w}^n_{j|i} - H\left(x_{t_{i+1}} + \hat{\mu}^n_{i+1|i}\right)\right).\end{aligned} \tag{13}$$

The derivation is given in Appendix B. Here, $\hat{\Sigma}_{j,i+1|i}$ denotes the sample cross-covariance between $\{\hat{\mu}^n_{j|i}\}_{n=1}^{N_e}$ and $\{\hat{\mu}^n_{i+1|i}\}_{n=1}^{N_e}$ for $j = i+1-L, ..., i+1$, and the smoothing Kalman gain $\hat{K}_{j|i}$ is given by

$$\hat{K}_{j|i} = \hat{\Sigma}_{j,i+1|i} H^\top \left(H\hat{\Sigma}_{i+1,i+1|i}H^\top + H\Sigma H^\top + \Gamma\right)^{-1}.$$

3 A Markov Prior on Discretization Error Mean

At this stage, we have not specified a Markov prior on the discretization error *mean*. In Toyota and Miyatake (2025), a Markov prior was introduced for the discretization error *variances*, derived from an error propagation analysis in numerical analysis. Here, we adopt the same underlying idea to model the discretization error *mean*. The exposition below follows that of Toyota and Miyatake (2025), with the notation adapted to the current setting for the discretization error mean.

3.1 Local Propagation of the Global Error

In this subsection, we briefly review the standard theory of numerical error analysis. For completeness and to fix notation, we summarize the distinction between the *local error* and the *global error*, and describe the propagation mechanism of the latter.

We make the following assumptions:

- The observation time interval $t_{i+1} - t_i$ is constant.
- The step size used in the time integrator is denoted by h, and $t_{i+1} - t_i = kh$ for some positive integer k.

The first assumption is introduced solely for clarity of exposition. If the second assumption does not hold in practice, the solution at $t = t_i$ can be approximated, for example, by interpolation techniques using neighboring numerical solutions (Hairer et al., 1993).

The numerical solution at $t = t_i + jh$ is denoted by $x_{i,j}$ so that $x_{i+1} = x_{i,k}$. The time-h flow of the ODE is denoted by ϕ_h, and the time-h flow of the numerical solver by ψ_h. The local error is defined as the numerical error induced in a single time step, i.e., $\phi_h(x) - \psi_h(x)$. We define

$$L(t_{i,j}) := \phi_h(x_{i,j}) - \psi_h(x_{i,j}) = \phi_h(x_{i,j}) - x_{i,j+1}.$$

On the other hand, the global error is defined by

$$G(t_{i,j}) = x(t_i + jh) - x_{i,j}.$$

We now examine how the global error propagates. Observe that

$$\begin{aligned}G(t_{i,j+1}) &= x(t_i + (j+1)h) - x_{i,j+1}\\ &= \phi_h(x(t_i + jh)) - \phi_h(x_{i,j}) + \phi_h(x_{i,j}) - x_{i,j+1}\\ &= \phi_h(x(t_i + jh)) - \phi_h(x_{i,j}) + L(t_{i,j}).\end{aligned}$$

The mean-value theorem implies that

$$\begin{aligned}&\phi_h(x(t_i + jh)) - \phi_h(x_{i,j})\\ &= \left(\int_0^1 \nabla_x \phi_h(x_{i,j} + \theta G(t_{i,j}))\, d\theta\right) G(t_{i,j})\\ &\approx \nabla_x \phi_h(x_{i,j}) G(t_{i,j}).\end{aligned}$$

Here, as $\phi_h(x_{i,j}) = x_{i,j} + hf(x_{i,j}) + \mathrm{o}(h)$ by the definition of ODE (1), we see that

$$\nabla_x \phi_h(x_{i,j}) = I + h\nabla_x f(x_{i,j}) + \mathrm{o}(h).$$

By ignoring the term $\mathrm{o}(h)$, the propagation of the global error can be approximated by

$$G(t_{i,j+1}) \approx (I + h\nabla_x f(x_{i,j}))G(t_{i,j}) + L(t_{i,j}). \tag{14}$$

3.2 A Markov Prior

In this section, we propose a Markov prior on discretization error means. Note that the discretization error mean $\mu_{t_{i,j}}$ in our context can be interpreted as a model for the global error $G(t_{i,j})$. Motivated by (14), we arrive at the following Markov prior, in which the discretization error mean $\mu_{t_{i,j}}$ evolves according to the probabilistic transition below as time advances by h:

$$\mu_{t_{i,j+1}} = M_{i,j}\mu_{t_{i,j}} + \tilde{L}(t_{i,j}), \tag{15}$$

for $M_{i,j} \sim P$ $(j = 0, \ldots, k-1)$. Here, P denotes a distribution over matrices $M_{i,j} \in \mathbb{R}^{d_{\mathcal{X}} \times d_{\mathcal{X}}}$. The value of $\tilde{L}(t_{i,j})$ is an approximation of $L(t_{i,j})$. This can be estimated by, for example, either $\tilde{L}(t_{i,j}) = \psi_{h/2} \circ \psi_{h/2}(x_{i,j}) - \psi_h(x_{i,j})$ or $\tilde{L}(t_{i,j}) = \tilde{\psi}_h(x_{i,j}) - \psi_h(x_{i,j})$, where $\tilde{\psi}_h$ denotes a higher order numerical solver. These techniques are commonly used to control the step size during time integration (Hairer et al., 1993).

The proposed Markov prior involves a distribution $M_{i,j} \sim P_\lambda$, which typically depends on a hyperparameter λ. In our experiment, for instance, we set $M_{i,j} = m_{i,j}\, I$, with $m_{i,j}$ drawn from a normal distribution $m_{i,j} \sim \mathcal{N}(\alpha, \beta^2)$; in this case, the hyperparameter $\lambda = (\alpha, \beta)$ must be properly chosen. In this study, we select the hyperparameter vector by maximizing an ensemble-based estimate of the marginal likelihood $p(y^*_{1:N} | \alpha, \beta)$.

As shown in the next section, the empirically optimal parameters often satisfy the condition $\mathbb{E}_{M\sim P_\lambda}[M] \approx I$, as observed in the discretization error variance case (Toyota and Miyatake, 2025). This suggests that it suffices to restrict attention to hyperparameter candidates λ satisfying $\mathbb{E}_{M\sim P_\lambda}[M] = I$. In the next subsection, we provide theoretical support for this empirical finding.

3.3 Asymptotics of the Prior

We study the asymptotics of the prior (15) as the step size h tends to zero. For the sake of theoretical simplicity, we identify $\tilde{L}(t_{i,j})$ with the exact local error $L(t_{i,j})$, although in practice it is approximated using two numerical solvers. To make the dependence on the step size explicit, we denote the prior by

$$\begin{aligned} p_h\left(\mu_{t_{0,0}}, \ldots, \mu_{t_{N,0}}\right) &= p_h\left(\mu_{t_0}, \ldots, \mu_{t_N}\right) \\ &= p_h\left(\mu_{t_0}\right) \prod_{i=0}^{N-1} p_h\left(\mu_{t_{i+1}} \mid \mu_{t_i}\right). \end{aligned} \tag{16}$$

Numerical integrators are generally constructed so that their discretization error vanishes as $h \to 0$. In accordance with this property, the prior $p_h(\mu_{t_i})$ should be specified so that μ_{t_i} converges to 0 in probability as $h \to 0$. The following theorem establishes this property and characterizes the associated convergence rate. The proof is given in Appendix C; the argument closely parallels Theorem 4.1 of Toyota and Miyatake (2025).

Theorem 3.1 (Modification of Toyota and Miyatake (2025, Theorem 4.1))**.** *Assume that there exist constants $C_M, C_L, C_\mu \in \mathbb{R}_{>0}$ satisfying (I) $\mathbb{E}_{M\sim P_\lambda}[\|I - M\|_{op}] \le C_M h$, (II) $L(t_{i,j}) \le C_L h^{a+1}$, (III) $\mathbb{E}[\|\mu_{t_0}\|] \le C_\mu h^b$. Assume also that, for every i and j, (IV) $M_{i,j} \sim P$ is independent of $\mu_{t_{i,j}}$. Then, for all $i \ge 1$,*

$$\mu_{t_i} = \mathcal{O}_p\left(h^{\min(a,b)}\right) \qquad (h \to 0).$$

Condition (I) ensures that the random matrix $M_{i,j}$ is a small perturbation of the identity matrix. This is consistent with the empirical finding $\mathbb{E}_{M\sim P_\lambda}[M] \approx I$ shown in Tables 1 and 2, especially for small step sizes h. Condition (II) defines a given numerical method to be of order a, i.e., the local error is $\mathcal{O}(h^{a+1})$. Condition (III) corresponds to the choice of a prior distribution $p_h(\mu_0)$ at the initial time t_0. This condition can be readily satisfied, since we can specify $p_h(\mu_0)$ flexibly.

This theorem shows that the proposed prior has a natural convergence rate that reflects insights from numerical analysis. For many numerical methods, if the local error satisfies $L(t_{i,j}) = \mathcal{O}(h^{a+1})$, then the global error $G(t_{i,j})$ is $\mathcal{O}(h^a)$, provided that the vector field of the underlying ODE is Lipschitz continuous (Butcher, 2016; Hairer et al., 1993; Iserles, 2008). Since the mean of the discretization error in this study corresponds to the global error, the proposed prior is expected to be $\mathcal{O}_p(h^a)$ under Condition (II). This theorem shows that this objective is achieved by constructing P and $p(\mu_{t_0})$ to satisfy Conditions (I) and (III), respectively.

Remark 3. The same set of conditions and the corresponding convergence rate were established in Toyota and Miyatake (2025) (Theorem 4.1) for the discretization error *variance*. Theorem 3.1 demonstrates that the analogous asymptotic property derived in Toyota and Miyatake (2025) extends to the discretization error *mean* setting considered in the present study. We also note that the nonnegativity assumption on M imposed in Toyota and Miyatake (2025) is relaxed here, thereby allowing for a broader class of prior constructions.

4 Experiment

We evaluate the proposed method on the pendulum system and the FitzHugh–Nagumo model, with an additional Lorenz-96 experiment presented in Appendix E. The ODEs are solved using the explicit Euler method, which is of order one, and the associated discretization errors are evaluated. For local error estimation, we employ the Runge method, which is of order two [2]. A distribution $M_{i,j} \sim P$ in the proposed prior is assumed to be $m_{i,j}I$, where $m_{i,j}$ follows a univariate normal distribution $\mathcal{N}(\alpha, \beta^2)$. The covariance matrix Σ representing the discretization error (3) is assumed to be γI, where $\gamma \in \mathbb{R}$. These parameters $(\alpha, \beta, \gamma) \in \mathbb{R}^3$ are selected by maximizing the marginal log-likelihood $\log p(y^*_{1:N} \mid \alpha, \beta, \gamma)$. The covariance matrix of the observation process (2) is fixed to $\Gamma = \mathrm{diag}(1.0^2, 1.0^2)$, and the ensemble size N_e and the smoothing lag L are 100 and 10, respectively. We assess the proposed method with a single observation trajectory, comparing the inferred discretization errors with the exact ones (more precisely, the error between a highly accurate reference solution and the numerical solution).

4.1 Pendulum System

We consider the second order ODE $\frac{dx^2(t)}{dt^2} = -\frac{g}{l}\sin x(t)$. In this section, we solve the equivalent first order system

$$\frac{d}{dt}\begin{bmatrix} x^1(t) \\ x^2(t) \end{bmatrix} = \begin{bmatrix} x^2(t) \\ -\frac{g}{l}\sin(x^1(t)) \end{bmatrix}, \tag{17}$$

where l is fixed to 3.0. We use the Euler method with a step size of 0.05 to numerically solve the system. The observation operator H is defined as

$$H = \begin{bmatrix} 1.0 & 2.0 \\ 2.0 & 1.0 \end{bmatrix}.$$

This operator is not one that is typically used in practice; rather, it was introduced to examine the performance of the method beyond the diagonal case. We assume that observations are available at $t = 1, 2, \ldots, 40$, with a fixed covariance matrix $\Gamma = \mathrm{diag}(1.0^2, 1.0^2)$. We also set $p(\mu_{t_0})$ to be a 2-dimensional standard normal distribution, $p(\mu_{t_0}) = \mathcal{N}(0, I)$.

Table 1 lists the top 10 parameters (α, β, γ) with corresponding log-likelihoods, among the candidates $\alpha \in \{-1.4, -1.2, \ldots, 1.4\}$, $\beta \in \{0.05, 0.10, \ldots, 0.5\}$ and $\gamma \in$

[2] Depending on the baseline solver, using a solver of the same or lower order may be sufficient. The key requirement is to compare numerical solutions from two different solvers.

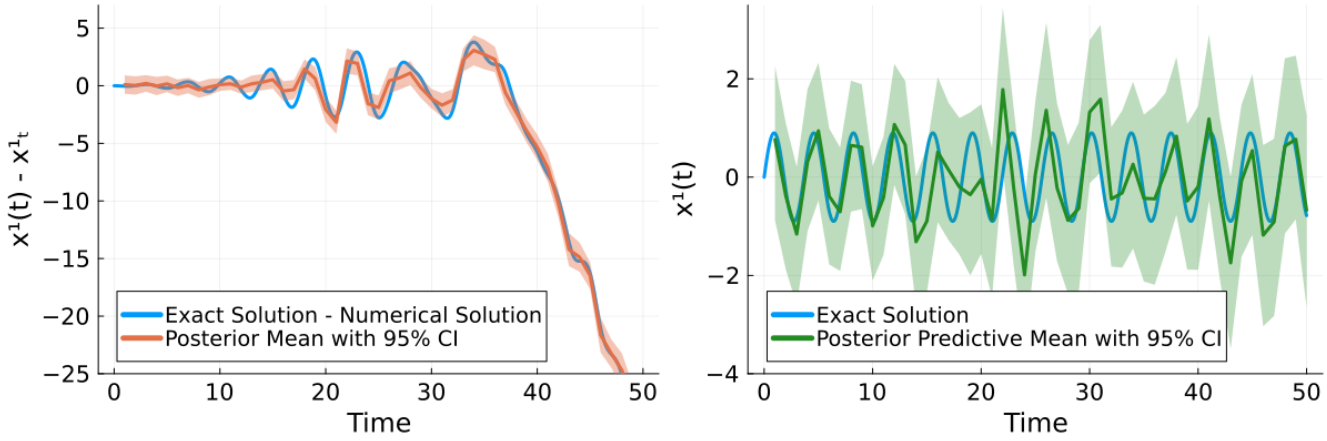


Figure 2: Discretization error quantification results for $x^1(t)$ in the pendulum system. **Left:** The mean and 95% credible interval of the posterior $p(\mu^1_{t_i} \mid y^*_{1:40})$. **Right:** The mean and 95% credible interval of the posterior predictive distribution $p(x^1(t) \mid y^*_{1:40})$.

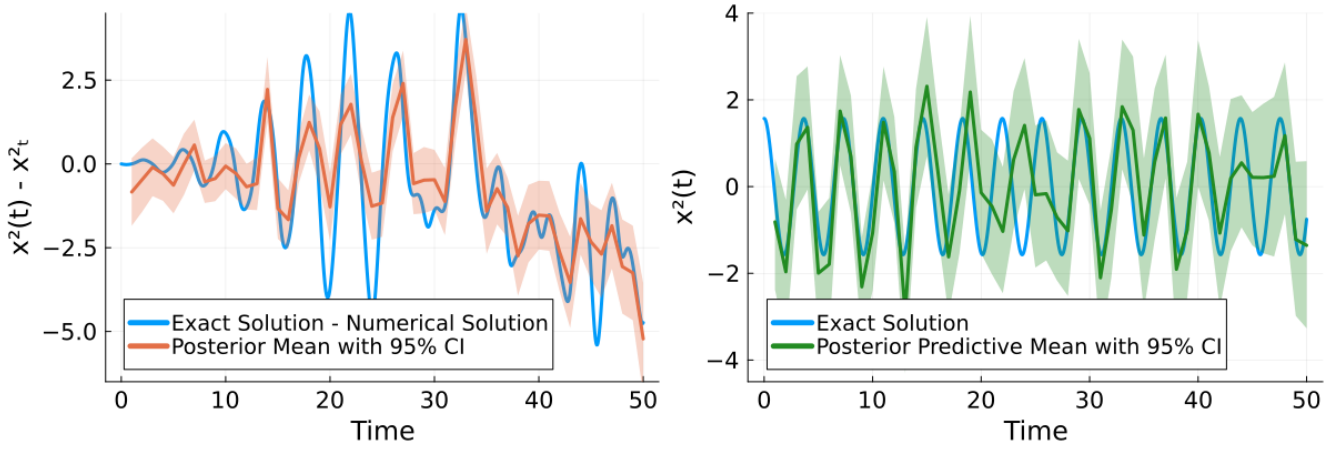


Figure 3: Discretization error quantification results for $x^2(t)$ in the pendulum system. **Left:** The mean and 95% credible interval of the posterior $p(\mu^2_{t_i} \mid y^*_{1:40})$. **Right:** The mean and 95% credible interval of the posterior predictive distribution $p(x^2(t) \mid y^*_{1:40})$.

Table 1: Top 10 hyperparameters (α, β, γ) with corresponding log-likelihoods $\log p(y^*_{1:40} \mid \alpha, \beta, \gamma)$ in the pendulum system.

Rank	(α, β, γ)	Log-likelihood
1	(1.000, 0.300, 0.500)	−231.731995
2	(1.000, 0.350, 0.500)	−233.595016
3	(1.000, 0.250, 0.500)	−237.323069
4	(1.000, 0.400, 0.500)	−241.765176
5	(1.000, 0.350, 1.000)	−241.830257
6	(1.000, 0.300, 1.000)	−244.841849
7	(−1.000, 0.400, 0.500)	−245.983714
8	(1.000, 0.250, 1.500)	−246.298006
9	(1.000, 0.400, 1.000)	−247.887706
10	(−1.000, 0.350, 0.500)	−248.311574

$\{0.5, 1.0, \ldots, 3.0\}$. We observe that high-likelihood parameter combinations tend to have $\alpha = 1$, with a few cases at $\alpha = -1$. This is consistent with the theoretical implication of Theorem 3.1. The standard-deviation parameter β ranges from 0.25 to 0.40, while γ takes 0.50 and 1.00. In the following numerical experiments, we use the choice $(\alpha, \beta, \gamma) = (1.0, 0.3, 0.5)$, which achieved the highest log-likelihood among all candidates.

The left panels in Figures 2 and 3 show the estimated posterior on μ_{t_i} with the true discretization error. We confirm that the proposed method successfully captures the true discretization error. The right panels in Figures 2 and 3 show the posterior predictive mean and the 95% credible intervals (CIs), computed from 100 samples drawn from the posterior predictive distribution $p(x(t) \mid y^*_{1:40})$. Specifically, we randomly draw samples $\hat{\mu}_{1:40}$ from the posterior $p(\mu_{1:40} \mid y^*_{1:40})$, and then generate samples of the discretization error as $r_{t_i} \sim \mathcal{N}(\hat{\mu}_{t_i}, \Sigma)$, following the definition of the discretization error mean in (3). Unlike in the left panels, we add the numerical solution x_{t_i} to r_{t_i} and compare $x_{t_i} + r_{t_i}$ with the exact solution $x(t_i)$. We observe that, while the estimated solution occasionally deviates from the exact solution, particularly in the second dimension around $t = 20$–30 and $t = 40$–50, the resulting trajectory follows the exact solution, indicating that the proposed method effectively reconstructs it.

4.2 FitzHugh-Nagumo model

We consider the FitzHugh-Nagumo model given by

$$\frac{d}{dt}\begin{bmatrix} x^1(t) \\ x^2(t) \end{bmatrix} = \begin{bmatrix} c\left(x^1(t) - \frac{\left(x^1(t)\right)^3}{3} + x^2(t)\right) \\ -\frac{1}{c}\left(x^1(t) - a + bx^2(t)\right) \end{bmatrix}$$

with $(a, b, c) = (0.5, -0.2, 1.0)$. We solve the equation using the Euler method with a step size of 0.2. The observation operator is set to $H = \mathrm{diag}(3.0, 3.0)$,

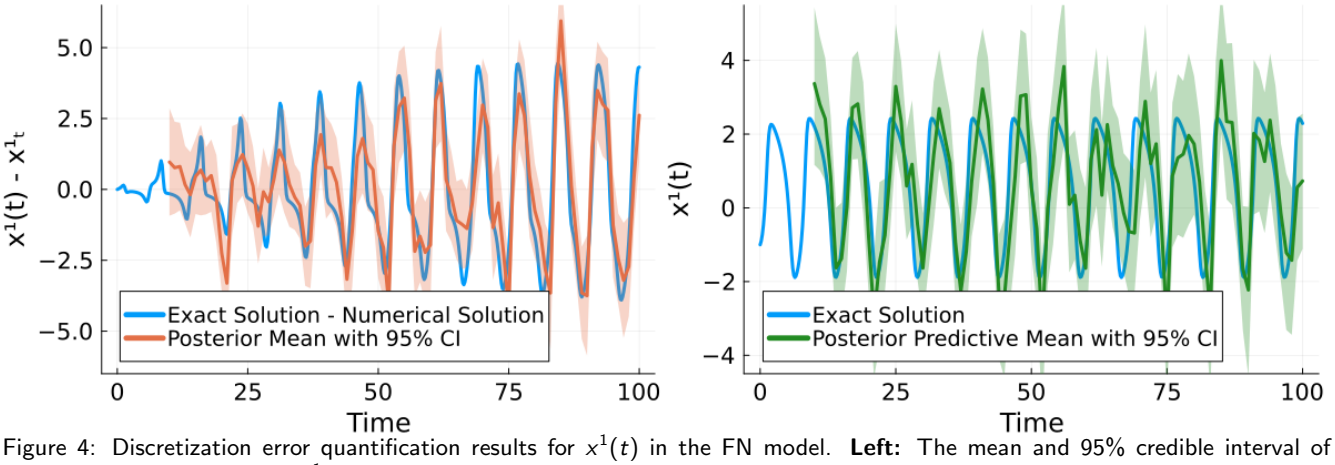


Figure 4: Discretization error quantification results for $x^1(t)$ in the FN model. **Left:** The mean and 95% credible interval of the posterior distribution $p(\mu^1_{t_i} \mid y^*_{10:50})$. **Right:** The mean and 95% credible interval of the posterior predictive distribution $p(x^1(t) \mid y^*_{10:50})$.

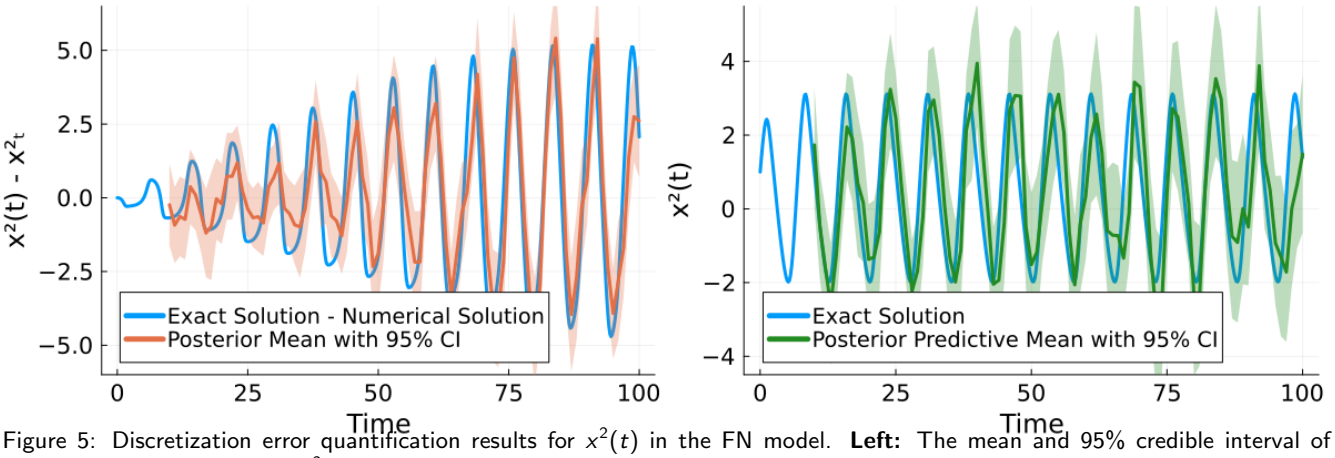


Figure 5: Discretization error quantification results for $x^2(t)$ in the FN model. **Left:** The mean and 95% credible interval of the posterior distribution $p(\mu^2_{t_i} \mid y^*_{10:50})$. **Right:** The mean and 95% credible interval of the posterior predictive distribution $p(x^2(t) \mid y^*_{10:50})$.

and observations are available at $t = 10, \ldots, 50$. We set $p(\mu_{t_0})$ to be a two dimensional normal distribution, $p(\mu_{t_0}) = \mathcal{N}(0, I)$.

Table 2 lists the top 10 hyperparameters (α, β, γ) and their corresponding log-likelihoods among the candidates $\alpha \in \{-1.4, -1.2, \ldots, 1.4\}$, $\beta \in \{0.1, 0.2, \ldots, 1.0\}$, and $\gamma \in \{0.5, 1.0, \ldots, 4.0\}$. We observe that all selected values of α are 1.0, consistent with the tendency observed in the pendulum system. The standard-deviation parameter β takes values 0.3 and 0.4, while no clear tendency is observed for γ. Henceforth, we adopt $(\alpha, \beta, \gamma) = (1.0, 0.3, 0.4)$, which achieved the highest log-likelihood among all candidates.

The left panels of Figs. 4 and 5 show the estimated posterior distributions in comparison with the true discretization errors. We confirm that the proposed method generally captures the trajectory of the true discretization errors. The right panels of the same figures present the posterior predictive distributions of the exact solutions. The resulting trajectories match the exact solutions, indicating that the proposed method effectively corrects the numerical solutions toward the exact ones.

Table 2: Top 10 parameter triplets (α, β, γ) with corresponding log-likelihoods $\log p(y^*_{10:50} \mid \alpha, \beta, \gamma)$ for the FN model.

Rank	(α, β, γ)	Log-likelihood
1	(1.000, 0.300, 4.000)	−556.734585
2	(1.000, 0.300, 2.500)	−557.755289
3	(1.000, 0.300, 1.500)	−558.155779
4	(1.000, 0.300, 3.500)	−558.311221
5	(1.000, 0.300, 2.000)	−558.649372
6	(1.000, 0.300, 1.000)	−559.566540
7	(1.000, 0.300, 3.000)	−561.452090
8	(1.000, 0.300, 0.500)	−564.933778
9	(1.000, 0.400, 4.000)	−568.680713
10	(1.000, 0.400, 3.500)	−571.341663

5 Conclusion

We propose a Bayesian framework to quantify discretization errors in ODEs through the *discretization error mean*. Our method enables inference of both the magnitudes and directions of the errors, allowing the numerical solution to be adjusted toward the exact solution. By introducing a Markov prior on the error mean motivated by classical error analysis, we perform inference using the Ensemble Kalman Filter. Numerical experiments demonstrate that the proposed method effectively captures discretization errors. Our approach relies on the assumption that the underlying model is consistent with the observations; handling model misspecification is an important further work. Jointly modeling the mean and variance of discretization errors would be another interesting extension, allowing for a more flexible representation of numerical uncertainty.

Acknowledgements

We thank Dr. Shin'ya Nakano and Dr. Keisuke Yano at the Institute of Statistical Mathematics for valuable discussions. In particular, we thank Dr. Shin'ya Nakano for insightful comments regarding the marginal likelihood computation for the EnKF. This work is supported by JSPS KAKENHI Grant Numbers 24K02951, 24K00540, 25H00449, 24K20750, 25K21806, JP25H01454 and JST ACT-X, Japan, Grant Number JPMJAX25CH. Miyatake is also supported by JST (Moonshot R&D Program) Japan Grant Number JPMJMS24A3 for the modeling and analysis parts.

References

A. Abdulle and G. Garegnani. Random time step probabilistic methods for uncertainty quantification in chaotic and geometric numerical integration. *Statistics and Computing*, 30(4):907–932, 2020.

M. Asch, M. Bocquet, and M. Nodet. *Data Assimilation*. Society for Industrial and Applied Mathematics, 2016.

J. Beck, N. Bosch, M. Deistler, K. L. Kadhim, J. H. Macke, P. Hennig, and P. Berens. Diffusion tempering improves parameter estimation with probabilistic integrators for ordinary differential equations. In *The 41st International Conference on Machine Learning*, volume 235, 2024.

G. Burgers, P. Jan van Leeuwen, and G. Evensen. Analysis scheme in the ensemble Kalman filter. *Monthly weather review*, 126(6):1719–1724, 1998.

J. C. Butcher. *Numerical Methods for Ordinary Differential Equations*. John Wiley & Sons, third edition, 2016.

J. Cockayne, C. J. Oates, T. J. Sullivan, and M. Girolami. Bayesian probabilistic numerical methods. *SIAM Review*, 61(4):756–789, 2019.

P. R. Conrad, M. Girolami, S. Särkkä, A. Stuart, and K. Zygalakis. Statistical analysis of differential equations: introducing probability measures on numerical solutions. *Statistics and Computing*, 27(4):1065–1082, 2017.

K. Cranmer, J. Brehmer, and G. Louppe. The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences*, 117(48):30055–30062, 2020.

G. Evensen. *Data Assimilation: the Ensemble Kalman filter*. Springer, 2009.

G. H. Golub and C. F. Van Loan. *Matrix computations*. JHU press, 2013.

E. Hairer, S. Nørsett, and G. Wanner. *Solving Ordinary Differential Equations I: Nonstiff Problems*. Springer, second edition, 1993.

P. Hennig, M. A. Osborne, and H. P. Kersting. *Probabilistic Numerics: Computation as Machine Learning*. Cambridge University Press, 2022.

A. Iserles. *A First Course in the Numerical Analysis of Differential Equations*. Cambridge University Press, USA, 2nd edition, 2008. ISBN 0521734908.

M. Katzfuss, J. R. Stroud, and C. K. Wikle. Understanding the ensemble Kalman filter. *The American Statistician*, 70(4):350–357, 2016.

H. Kersting, T. J. Sullivan, and P. Hennig. Convergence rates of Gaussian ODE filters. *Statistics and Computing*, 30(6):1791–1816, 2020.

N. Krämer. Adaptive probabilistic ODE solvers without adaptive memory requirements. In *Proceedings of the First International Conference on Probabilistic Numerics*, 2025.

N. Krämer. Numerically robust fixed-point smoothing without state augmentation. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=LVQ8BEL5n3.

Y. Le Fay, S. Särkkä, and A. Corenflos. Modelling pathwise uncertainty of stochastic differential equations samplers via probabilistic numerics. *Bayesian Analysis*, 1(1):1–24, 2025.

H. C. Lie, A. M. Stuart, and T. J. Sullivan. Strong convergence rates of probabilistic integrators for ordinary differential equations. *Statistics and Computing*, 29 (6):1265–1283, 2019.

H. C. Lie, M. Stahn, and T. J. Sullivan. Randomised one-step time integration methods for deterministic operator differential equations. *Calcolo*, 59(1):13, 2022.

N. Marumo, T. Matsuda, and Y. Miyatake. Modelling the discretization error of initial value problems using the Wishart distribution. *Applied Mathematics Letters*, 147:108833, 2024.

T. Matsuda and Y. Miyatake. Estimation of ordinary differential equation models with discretization error quantification. *SIAM/ASA Journal on Uncertainty Quantification*, 9(1):302–331, 2021.

Y. Miyatake, K. Irie, and T. Matsuda. Quantifying uncertainty in the numerical integration of evolution equations based on Bayesian isotonic regression. *Japan Journal of Industrial and Applied Mathematics*, June 2025.

S. Särkkä and L. Svensson. *Bayesian Filtering and Smoothing*, volume 17. Cambridge university press, 2023.

J. Schmidt, N. Krämer, and P. Hennig. A probabilistic state space model for joint inference from differential equations and data. In *Advances in Neural Information Processing Systems*, volume 34. Curran Associates, Inc., 2021.

M. Schober, S. Särkkä, and P. Hennig. A probabilistic model for the numerical solution of initial value problems. *Statistics and Computing*, 29(1):99–122, 2018.

S. Toyota and Y. Miyatake. Joint Bayesian inference of parameter and discretization error uncertainties in ODE models. *arXiv preprint arXiv:2511.23010*, 2025.

F. Tronarp, H. Kersting, S. Särkkä, and P. Hennig. Probabilistic solutions to ordinary differential equations as non-linear Bayesian filtering: A new perspective. *Statistics and Computing*, 29(6):1297–1315, 2019.

F. Tronarp, S. Särkkä, and P. Hennig. Bayesian ODE solvers: the maximum a posteriori estimate. *Statistics and Computing*, 31(3):23, 2021.

F. Tronarp, N. Bosch, and P. Hennig. Fenrir: Physics-enhanced regression for initial value problems. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, 2022.

D. Yao, F. Tronarp, and N. Bosch. Propagating model uncertainty through filtering-based probabilistic numerical ODE solvers. In *Proceedings of the First International Conference on Probabilistic Numerics*, 2025.

Supplementary Materials to "Bayesian Inference of Discretization Error Means in ODEs via Ensemble Kalman Filtering"

A Proof of Theorem 2.1

We prove Theorem 2.1. This result is essentially established in the classical EnKF (Burgers et al., 1998; Evensen, 2009), with the only difference being the non-zero mean in the Gaussian observation process.

We first state two lemmas for the proof.

Lemma A.1 (Woodbury Matrix Identity (Golub and Van Loan, 2013)). *Let $A \in \mathbb{R}^{n\times n}$ be an invertible matrix, $U \in \mathbb{R}^{n\times k}$, $C \in \mathbb{R}^{k\times k}$, and $V \in \mathbb{R}^{k\times n}$. Assume that C is invertible and that $C^{-1} + VA^{-1}U$ is invertible. Then we have:*

$$(A+UCV)^{-1} = A^{-1} - A^{-1}U\left(C^{-1}+VA^{-1}U\right)^{-1}VA^{-1}.$$

Lemma A.2. *Let $x \sim \mathcal{N}(\mu_1, \Sigma_1)$, and suppose that $p(y|x)$ is given by*

$$y = Hx + \varepsilon$$

where $\varepsilon \sim \mathcal{N}(\mu_2, \Sigma_2)$ is independent of x. Then, $p(x|y)$ is given by

$$\mathcal{N}\left(\mu_1 + K\left(y - H\mu_1 - \mu_2\right), \left(I - KH\right)\Sigma_1\right),$$

where

$$K := \Sigma_1 H^\top (H\Sigma_1 H^\top + \Sigma_2)^{-1}.$$

Proof. By Bayes' rule, we have

$$\begin{aligned} p(x|y) &\propto p(y|x)p(x) \\ &\propto \exp\Big\{-\frac{1}{2}\big[(y-Hx-\mu_2)^\top\Sigma_2^{-1}(y-Hx-\mu_2) + (x-\mu_1)^\top\Sigma_1^{-1}(x-\mu_1)\big]\Big\}. \end{aligned}$$

Expanding the quadratic form in x, we have

$$\begin{aligned} &(y-Hx-\mu_2)^\top\Sigma_2^{-1}(y-Hx-\mu_2) + (x-\mu_1)^\top\Sigma_1^{-1}(x-\mu_1) \\ &= \big((y-\mu_2)-Hx\big)^\top\Sigma_2^{-1}\big((y-\mu_2)-Hx\big) + (x-\mu_1)^\top\Sigma_1^{-1}(x-\mu_1) \\ &= x^\top H^\top\Sigma_2^{-1}Hx - 2(y-\mu_2)^\top\Sigma_2^{-1}Hx + (x-\mu_1)^\top\Sigma_1^{-1}(x-\mu_1) + \text{const} \\ &= x^\top(\Sigma_1^{-1} + H^\top\Sigma_2^{-1}H)x - 2\big(H^\top\Sigma_2^{-1}(y-\mu_2) + \Sigma_1^{-1}\mu_1\big)^\top x + \text{const} \\ &= \Big(x - (\Sigma_1^{-1} + H^\top\Sigma_2^{-1}H)^{-1}\big(H^\top\Sigma_2^{-1}(y-\mu_2) + \Sigma_1^{-1}\mu_1\big)\Big)^\top \\ &\quad (\Sigma_1^{-1} + H^\top\Sigma_2^{-1}H) \\ &\quad \Big(x - (\Sigma_1^{-1} + H^\top\Sigma_2^{-1}H)^{-1}\big(H^\top\Sigma_2^{-1}(y-\mu_2) + \Sigma_1^{-1}\mu_1\big)\Big) + \text{const}. \end{aligned}$$

Here, the term "const" represents a quantity that does not depend on x. Therefore the posterior mean and covariance matrix are given by

$$(\Sigma_1^{-1} + H^\top\Sigma_2^{-1}H)^{-1}(H^\top\Sigma_2^{-1}(y-\mu_2) + \Sigma_1^{-1}\mu_1), \tag{18}$$

and

$$(\Sigma_1^{-1} + H^\top\Sigma_2^{-1}H)^{-1} \tag{19}$$

respectively. By applying Lemma A.1 with $A = \Sigma_1^{-1}$, $U = H^\top$, $C = \Sigma_2^{-1}$ and $V = H$, (19) can be written as

$$\begin{aligned} (19) &= \Sigma_1 - \Sigma_1 H^\top(\Sigma_2 + H\Sigma_1 H^\top)^{-1}H\Sigma_1 \\ &= \Sigma_1 - KH\Sigma_1 = (I-KH)\Sigma_1. \end{aligned}$$

Here, K is defined by $K := \Sigma_1 H^\top (\Sigma_2 + H\Sigma_1 H^\top)^{-1}$. Moreover, applying Lemma A.1 once again, we obtain

$$\begin{aligned}(18) &= (\Sigma_1 - \Sigma_1 H^\top (H\Sigma_1 H^\top + \Sigma_2)^{-1} H\Sigma_1)\big(H^\top \Sigma_2^{-1}(y-\mu_2) + \Sigma_1^{-1}\mu_1\big) \\ &= (\Sigma_1 - KH\Sigma_1)\big(H^\top \Sigma_2^{-1}(y-\mu_2) + \Sigma_1^{-1}\mu_1\big) \\ &= \mu_1 - KH\mu_1 + (\Sigma_1 - KH\Sigma_1)H^\top \Sigma_2^{-1}(y-\mu_2). \end{aligned} \tag{20}$$

We note that

$$\begin{aligned}(\Sigma_1 - KH\Sigma_1)H^\top\Sigma_2^{-1} &= \Sigma_1 H^\top \Sigma_2^{-1} - \Sigma_1 H^\top (H\Sigma_1 H^\top + \Sigma_2)^{-1} H\Sigma_1 H^\top \Sigma_2^{-1} \\ &= \Sigma_1 H^\top \left[\Sigma_2^{-1} - (H\Sigma_1 H^\top + \Sigma_2)^{-1} H\Sigma_1 H^\top \Sigma_2^{-1}\right].\end{aligned}$$

Using the identity

$$\begin{aligned}\Sigma_2^{-1} - (H\Sigma_1 H^\top + \Sigma_2)^{-1} H\Sigma_1 H^\top \Sigma_2^{-1} &= (H\Sigma_1 H^\top + \Sigma_2)^{-1}\left(\left(H\Sigma_1 H^\top + \Sigma_2\right)\Sigma_2^{-1} - H\Sigma_1 H^\top \Sigma_2^{-1}\right) \\ &= (H\Sigma_1 H^\top + \Sigma_2)^{-1} I = (H\Sigma_1 H^\top + \Sigma_2)^{-1},\end{aligned}$$

we obtain

$$(\Sigma_1 - KH\Sigma_1)H^\top\Sigma_2^{-1} = \Sigma_1 H^\top (H\Sigma_1 H^\top + \Sigma_2)^{-1} = K.$$

Therefore,

$$(20) = \mu_1 - KH\mu_1 + K(y-\mu_2) = \mu_1 + K(y - H\mu_1 - \mu_2).$$

This concludes the proof. □

Proof of Theorem 2.1 We first prove

$$\hat{\mu}_{i+1|i+1} \xrightarrow{P} \mu_{i+1|i+1}. \tag{21}$$

By the definition of $\hat{\mu}_{i+1|i+1}$, we have

$$\begin{aligned}\hat{\mu}_{i+1|i+1} &= \frac{1}{N_e}\sum_{n=1}^{N_e} \hat{\mu}^n_{i+1|i+1} = \frac{1}{N_e}\sum_{n=1}^{N_e} \hat{\mu}^n_{i+1|i} + \frac{1}{N_e}\sum_{n=1}^{N_e} \hat{K}_{i+1|i}\left(y^*_{t_{i+1}} + \hat{w}^n_{i+1} - H\left(x_{t_{i+1}} + \hat{\mu}^n_{i+1|i}\right)\right) \\ &= \hat{\mu}_{i+1|i} + \hat{K}_{i+1|i}\left(y^*_{t_{i+1}} + \hat{w}_{i+1} - H\left(x_{t_{i+1}} + \hat{\mu}_{i+1|i}\right)\right),\end{aligned} \tag{22}$$

where

$$\hat{w}_{i+1} := \frac{1}{N_e}\sum_{n=1}^{N_e} \hat{w}^n_{i+1}, \quad \hat{\mu}_{i+1|i} = \frac{1}{N_e}\sum_{n=1}^{N_e} \hat{\mu}^n_{i+1|i}.$$

Since $\hat{w}^n_{i+1} \sim \mathcal{N}(0, H\Sigma H^\top + \Gamma)$, we have $\hat{w}_{i+1} \xrightarrow{P} 0$. Therefore, it follows that

$$(22) \xrightarrow{P} \mu_{i+1|i} + K_{i+1|i}\left(y^*_{t_{i+1}} - H\left(x_{t_{i+1}} + \mu_{i+1|i}\right)\right), \tag{23}$$

where

$$K_{i+1|i} := \Sigma_{i+1|i} H^\top \left(H\Sigma_{i+1|i} H^\top + H\Sigma H^\top + \Gamma\right)^{-1}. \tag{24}$$

By Lemma A.2, (23) also coincides with $\mu_{i+1|i+1}$, which proves (21).

We next prove

$$\hat{\Sigma}_{i+1|i+1} \xrightarrow{P} \Sigma_{i+1|i+1}.$$

We first rewrite $\hat{\mu}_{i+1|i+1}$ as

$$\begin{aligned}\hat{\mu}^n_{i+1|i+1} - \hat{\mu}_{i+1|i+1} &= \hat{\mu}^n_{i+1|i} + \hat{K}_{i+1|i}\left(y^*_{t_{i+1}} + \hat{w}^n_{i+1} - H\left(x_{t_{i+1}} + \hat{\mu}^n_{i+1|i}\right)\right) \\ &\quad - \hat{\mu}_{i+1|i} - \hat{K}_{i+1|i}\left(y^*_{t_{i+1}} + \hat{w}_{i+1} - H\left(x_{t_{i+1}} + \hat{\mu}_{i+1|i}\right)\right) \\ &= (I - \hat{K}_{i+1|i}H)(\hat{\mu}^n_{i+1|i} - \hat{\mu}_{i+1|i}) + \hat{K}_{i+1|i}(\hat{w}^n_{i+1} - \hat{w}_{i+1}).\end{aligned}$$

Therefore, $\hat{\Sigma}_{i+1|i+1}$ can be written as

$$\begin{aligned}
\hat{\Sigma}_{i+1|i+1} &:= \frac{1}{N_e-1}\sum_{n=1}^{N_e}\left(\hat{\mu}^n_{i+1|i+1}-\hat{\mu}_{i+1|i+1}\right)\left(\hat{\mu}^n_{i+1|i+1}-\hat{\mu}_{i+1|i+1}\right)^\top \\
&= \frac{1}{N_e-1}\sum_{n=1}^{N_e}\left\{(I-\hat{K}_{i+1|i}H)(\hat{\mu}^n_{i+1|i}-\hat{\mu}_{i+1|i})+\hat{K}_{i+1|i}(\hat{w}^n_{i+1}-\hat{w}_{i+1})\right\} \\
&\qquad\qquad \left\{(I-\hat{K}_{i+1|i}H)(\hat{\mu}^n_{i+1|i}-\hat{\mu}_{i+1|i})+\hat{K}_{i+1|i}(\hat{w}^n_{i+1}-\hat{w}_{i+1})\right\}^\top \\
&= \underbrace{\frac{1}{N_e-1}\sum_{n=1}^{N_e}\{(I-\hat{K}_{i+1|i}H)(\hat{\mu}^n_{i+1|i}-\hat{\mu}_{i+1|i})(\hat{\mu}^n_{i+1|i}-\hat{\mu}_{i+1|i})^\top(I-\hat{K}_{i+1|i}H)^\top\}}_{\text{(A)}} \\
&\qquad\qquad + \underbrace{\frac{1}{N_e-1}\sum_{n=1}^{N_e}\left\{\hat{K}_{i+1|i}(\hat{w}^n_{i+1}-\hat{w}_{i+1})(\hat{w}^n_{i+1}-\hat{w}_{i+1})^\top\hat{K}^\top_{i+1|i}\right\}}_{\text{(B)}}+(*),
\end{aligned}$$

where

$$\begin{aligned}
(*) = &\frac{1}{N_e-1}\sum_{n=1}^{N_e}\left\{(I-\hat{K}_{i+1|i}H)(\hat{\mu}^n_{i+1|i}-\hat{\mu}_{i+1|i})(\hat{w}^n_{i+1}-\hat{w}_{i+1})^\top\hat{K}^\top_{i+1|i}\right\} \\
&+\frac{1}{N_e-1}\sum_{n=1}^{N_e}\left\{\hat{K}_{i+1|i}(\hat{w}^n_{i+1}-\hat{w}_{i+1})(\hat{\mu}^n_{i+1|i}-\hat{\mu}_{i+1|i})^\top(I-\hat{K}_{i+1|i}H)^\top\right\}.
\end{aligned}$$

Since $\hat{w}^n_{i+1}$ and $\hat{\mu}^n_{i+1|i}$ are independent, $(*)$ converges in probability to the zero matrix O:

$$(*) \xrightarrow{P} O. \tag{25}$$

Since $\hat{\mu}^n_{i+1|i}$ and $\hat{w}^n_{i+1}$ are drawn i.i.d. from $p(\mu_{t_{i+1}}|y_{1:i})$ and $\mathcal{N}(0, H\Sigma H^\top+\Gamma)$, we see that

$$(A) \xrightarrow{P} (I-K_{i+1|i}H)\Sigma_{i+1|i}(I-K_{i+1|i}H)^\top, \tag{26}$$

$$(B) \xrightarrow{P} K_{i+1|i}(H\Sigma H^\top+\Gamma)K^\top_{i+1|i}. \tag{27}$$

Equations (25), (26), (27) lead to

$$\begin{aligned}
\hat{\Sigma}_{i+1|i+1} = (\mathrm{A})+(\mathrm{B})+(*) &\xrightarrow{P} (I-K_{i+1|i}H)\Sigma_{i+1|i}(I-K_{i+1|i}H)^\top + K_{i+1|i}(H\Sigma H^\top+\Gamma)K^\top_{i+1|i}+O \\
&= (I-K_{i+1|i}H)\Sigma_{i+1|i}(I-K_{i+1|i}H)^\top + K_{i+1|i}(H\Sigma H^\top+\Gamma)K^\top_{i+1|i}.
\end{aligned}$$

Finally, we see that

$$\begin{aligned}
&(I-K_{i+1|i}H)\Sigma_{i+1|i}(I-K_{i+1|i}H)^\top + K_{i+1|i}(H\Sigma H^\top+\Gamma)K^\top_{i+1|i} \\
&= (I-K_{i+1|i}H)\Sigma_{i+1|i} - \Sigma_{i+1|i}H^\top K^\top_{i+1|i} + K_{i+1|i}H\Sigma_{i+1|i}H^\top K^\top_{i+1|i} + K_{i+1|i}(H\Sigma H^\top+\Gamma)K^\top_{i+1|i} \\
&= (I-K_{i+1|i}H)\Sigma_{i+1|i} - \Sigma_{i+1|i}H^\top K^\top_{i+1|i} + K_{i+1|i}(H\Sigma_{i+1|i}H^\top + H\Sigma H^\top+\Gamma)K^\top_{i+1|i} \\
&= (I-K_{i+1|i}H)\Sigma_{i+1|i} - \Sigma_{i+1|i}H^\top K^\top_{i+1|i} \\
&\quad + \Sigma_{i+1|i}H^\top\left(H\Sigma_{i+1|i}H^\top + H\Sigma H^\top+\Gamma\right)^{-1}(H\Sigma_{i+1|i}H^\top + H\Sigma H^\top+\Gamma)K^\top_{i+1|i} \\
&= (I-K_{i+1|i}H)\Sigma_{i+1|i} - \Sigma_{i+1|i}H^\top K^\top_{i+1|i} + \Sigma_{i+1|i}H^\top K^\top_{i+1|i} \\
&= (I-K_{i+1|i}H)\Sigma_{i+1|i}.
\end{aligned}$$

Here, the third equation is derived from the definition (24) of $K_{i+1|i}$. This concludes the proof, with the use of Lemma A.2.

B Derivation of Eq. (13).

We first note that, under the augmented state-space model, the observation operator $\tilde{\boldsymbol{H}}$ is given by

$$\tilde{\boldsymbol{H}} = (H, O, \ldots, O), \tag{28}$$

where O denotes a $d_{\mathcal{Y}} \times d_{\mathcal{X}}$ zero matrix. Assume that we have an ensemble $\{\hat{\boldsymbol{\mu}}^n_{i+1|i}\}_{n=1}^{N_e} = \left\{ \left(\hat{\mu}^n_{i+1|i}, \ldots, \hat{\mu}^n_{i+1-L|i} \right)^\top \right\}_{n=1}^{N_e}$. of $p(\boldsymbol{\mu}_{i+1}|y_{1:i})$. Let $\hat{\boldsymbol{\Sigma}}_{i+1|i}$ be a sample covariance matrix of the ensemble. Then, the correction step for the augmented state-space model updates the ensemble $\{\hat{\boldsymbol{\mu}}^n_{i+1|i}\}_{n=1}^{N_e}$ by

$$\hat{\boldsymbol{\mu}}^n_{i+1|i+1} := \hat{\boldsymbol{\mu}}^n_{i+1|i} + \hat{\boldsymbol{\Sigma}}_{i+1|i}\tilde{\boldsymbol{H}}^\top \left(\tilde{\boldsymbol{H}}\hat{\boldsymbol{\Sigma}}_{i+1|i}\tilde{\boldsymbol{H}}^\top + H\Sigma H^\top + \Gamma \right)^{-1} \left(y^*_{t_{i+1}} + \hat{w}^n_{i+1} - Hx_{t_{i+1}} - \tilde{\boldsymbol{H}}\hat{\boldsymbol{\mu}}^n_{i+1|i} \right). \tag{29}$$

It follows from the definition of $\tilde{\boldsymbol{H}}$ that

$$\tilde{\boldsymbol{H}}\hat{\boldsymbol{\mu}}^n_{i+1|i} = (H, O, \ldots, O)(\hat{\mu}^n_{i+1|i}, \hat{\mu}^n_{i|i}, \ldots, \hat{\mu}^n_{i-L+1|i})^\top = H\hat{\mu}^n_{i+1|i}. \tag{30}$$

Noting that

$$\hat{\boldsymbol{\Sigma}}_{i+1|i} = \begin{pmatrix} \hat{\Sigma}_{i+1,i+1|i} & \hat{\Sigma}_{i+1,i|i} & \cdots & \hat{\Sigma}_{i+1,i-L+1|i} \\ \hat{\Sigma}_{i,i+1|i} & \hat{\Sigma}_{i,i|i} & \cdots & \hat{\Sigma}_{i,i-L+1|i} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\Sigma}_{i-L+1,i+1|i} & \hat{\Sigma}_{i-L+1,i|i} & \cdots & \hat{\Sigma}_{i-L+1,i-L+1|i} \end{pmatrix},$$

where $\hat{\Sigma}_{j,k|i} := \frac{1}{N_e - 1}\sum_{n=1}^{N_e} \left(\hat{\mu}^n_{j|i} - \hat{\mu}_{j|i} \right) \left(\hat{\mu}^n_{k|i} - \hat{\mu}_{k|i} \right)^\top$ denotes the sample cross-covariance between $\{\hat{\mu}^n_{j|i}\}_{n=1}^{N_e}$ and $\{\hat{\mu}^n_{k|i}\}_{n=1}^{N_e}$, we have

$$\hat{\boldsymbol{\Sigma}}_{i+1|i}\tilde{\boldsymbol{H}}^\top = (\hat{\Sigma}_{i+1,i+1|i}, \hat{\Sigma}_{i,i+1|i}, ..., \hat{\Sigma}_{i-L+1,i+1|i})^\top H^\top, \tag{31}$$

$$\tilde{\boldsymbol{H}}\hat{\boldsymbol{\Sigma}}_{i+1|i}\tilde{\boldsymbol{H}}^\top = H\hat{\Sigma}_{i+1,i+1|i}H^\top. \tag{32}$$

Applying (30), (31) and (32) to (29) concludes the proof.

C Proof of Theorem 3.1

In this appendix, we prove Theorem 3.1. The proof proceeds in essentially the same manner as that of Theorem 4.1 in Toyota and Miyatake (2025).

While (16) defines the Markov prior only at the observation points $t_0, \ldots, t_N$, the recursive construction in (15) naturally allows us to extend this definition to finer time grids as follows:

$$p_h(\mu_{t_{0,0}}, \mu_{t_{0,1}}, ..., \mu_{t_{0,k-1}}, \mu_{t_{1,0}} \ldots, \mu_{t_{N,0}}) = p_h(\mu_{t_{0,0}}) \prod_{i=0}^{N-1} \prod_{j=0}^{k-1} p_h(\mu_{t_{i,j+1}} \mid \mu_{t_{i,j}}).$$

Proof of Theorem 3.1. Let us recall that random variables $\{X_h\}_{h>0}$ and $\{Y_h\}_{h>0}$ indexed by h satisfy $X_h = \mathcal{O}_p(Y_h)$ if X_h/Y_h is uniformly tight:

$$\lim_{\epsilon \to \infty} \sup_h \mathbb{P}\left(\left| \frac{X_h}{Y_h} \right| \geq \epsilon \right) = 0.$$

By Markov inequality, we have

$$\sup_h \mathbb{P}\left(\left\| \frac{\mu_{t_i}}{h^{\min(a,b)}} \right\| \geq \epsilon \right) \leq \sup_h \frac{1}{\epsilon h^{\min(a,b)}} \mathbb{E}_{\mu_{t_i} \sim p_h(\mu_{t_i})}[\|\mu_{t_i}\|]. \tag{33}$$

Hence, it suffices to prove that

$$\mathbb{E}_{p_h(\mu_{t_i})}[\|\mu_{t_i}\|] = \mathcal{O}(h^{\min(a,b)})$$

for $i \geq 1$.

Define

$$e_{i,j} := \mathbb{E}_{p_h(\mu_{t_{i,j}})}[\|\mu_{t_{i,j}}\|]$$

to simplify the notation. By the definition (15), it holds that for $0 \leq i \leq (N-1)$ and $0 \leq j \leq k-1$,

$$\mu_{t_{i,j+1}} = M_{i,j}\mu_{t_{i,j}} + L\left(t_{i,j}\right).$$

This leads to the inequality

$$
\begin{aligned}
e_{i,j+1} &= \mathbb{E}_{p_h(\mu_{t_{i,j+1}})}[\|\mu_{t_{i,j+1}}\|] \\
&= \mathbb{E}_{p_h(\mu_{t_{i,j}}),P_\lambda}[\|M_{i,j}\mu_{t_{i,j}} + L(t_{i,j})\|] \\
&\leq \mathbb{E}_{p_h(\mu_{t_{i,j}}),P_\lambda}[\|M_{i,j}\mu_{t_{i,j}}\|] + \|L(t_{i,j})\| \\
&\leq \mathbb{E}_{p_h(\mu_{t_{i,j}}),P_\lambda}[\|M_{i,j}\|_{\mathrm{op}}\|\mu_{t_{i,j}}\|] + \|L(t_{i,j})\| \\
&= \mathbb{E}_{P_\lambda}[\|M_{i,j}\|_{\mathrm{op}}]\mathbb{E}_{p_h(\mu_{t_{i,j}})}[\|\mu_{t_{i,j}}\|] + \|L(t_{i,j})\| \qquad (34) \\
&= \mathbb{E}_{P_\lambda}[\|I - (I - M_{i,j})\|_{\mathrm{op}}]e_{i,j} + \|L(t_{i,j})\| \\
&\leq \{\mathbb{E}_{P_\lambda}[\|I\|_{\mathrm{op}}] + \mathbb{E}_{P_\lambda}[\|I - M_{i,j}\|_{\mathrm{op}}]\}\, e_{i,j} + \|L(t_{i,j})\| \\
&\leq (1 + C_M h)e_{i,j} + C_L h^{a+1}. \qquad (35)
\end{aligned}
$$

Here, $\mathbb{E}_{p_h(\mu_{t_{i,j}}),\, P_\lambda}$ denotes the expectation with respect to $\mu_{t_{i,j}} \sim p_h(\mu_{t_{i,j}})$ and $M_{i,j} \sim P_\lambda$. The equality (34) is derived by Condition (IV). The inequality (35) is derived from Conditions (I) and (II). Applying (35) recursively yields

$$
\begin{aligned}
e_{i,j+1} &\leq (1 + C_M h)^{j+1+ik} e_{0,0} + C_L h^{a+1} \sum_{l=1}^{j+ik+1} (1 + C_M h)^{l-1} \\
&= (1 + C_M h)^{j+1+ik} e_{0,0} + C_L h^a \frac{(1 + C_M h)^{j+1+ik} - 1}{L} \\
&\leq \exp\big(C_M(t_{i,j+1} - t_0)\big) e_{0,0} + C_L h^a \frac{\exp\big(C_M(t_{i,j+1} - t_0)\big) - 1}{C_M}. \qquad (36)
\end{aligned}
$$

Since this holds for $j = k - 1$ and $0 \leq i \leq N - 1$, we obtain

$$
\begin{aligned}
\mathbb{E}_{p_h(\mu_{t_i})}[\|\mu_{t_i}\|] &= e_{i-1,k} \\
&\leq \exp\big(C_M(t_i - t_0)\big) e_{0,0} + C_L h^a \frac{\exp\big(C_M(t_i - t_0)\big) - 1}{C_M} \\
&\leq \exp\big(C_M(t_i - t_0)\big) C_\mu h^b + C_L h^a \frac{\exp\big(C_M(t_i - t_0)\big) - 1}{C_M} \\
&\leq \left\{ \exp\big(C_M(t_i - t_0)\big) C_\mu + C_L \frac{\exp\big(C_M(t_i - t_0)\big) - 1}{C_M} \right\} h^{\min(a,b)}
\end{aligned}
$$

for $h \leq 1$, which concludes the proof. Here, the final inequality is derived from Condition (III). □

D Related Work

In this section, we briefly review uncertainty quantification methods for differential equations and related techniques based on the Ensemble Kalman Filter.

Probabilistic Numerics For the quantification of discretization error in differential equations, a recently well-established paradigm is known as probabilistic numerics (PN)(Cockayne et al., 2019; Hennig et al., 2022). By viewing numerical computation as a statistical inference problem, PN provides numerical solutions together with principled uncertainty quantification of discretization errors. As discussed in Section 1, Bayesian ODE solvers (Beck et al., 2024; Kersting et al., 2020; Le Fay et al., 2025; Schmidt et al., 2021; Schober et al., 2018; Tronarp et al., 2022, 2019, 2021; Yao et al., 2025; Krämer, 2025) are among the most developed approaches within PN for ordinary differential equations. These methods place a prior distribution on the solution and its derivatives, and treat the constraint that the solution satisfies the differential equation as an observation (referred to as an information operator in the PN literature (Cockayne et al., 2019)). Conditioning on these observations yields a posterior distribution over the numerical solution. An alternative approach is the perturbation-based method (Abdulle and Garegnani, 2020; Conrad et al., 2017; Lie et al., 2022, 2019). Rather than placing a prior directly on the solution, these methods intentionally introduce stochastic perturbations into the numerical solver and quantify uncertainty through the resulting distribution of numerical trajectories.

These approaches differ from the present work in two important respects. First, Bayesian ODE solvers place a Markov prior directly on the solution of the differential equation, whereas in this study the Markov prior is imposed

on the discretization-error trajectory. By placing the prior on the discretization error itself, we can incorporate insights from classical error analysis into the prior model, which is difficult when the prior is defined on the solution. Second, unlike other PN approaches, our method infers the discretization error directly from observations. This perspective enables the simultaneous updating of the discretization error, alongside the standard targets of ODE inverse problems (e.g., model parameters), using observational data.

Extensions of Ensemble Kalman Filters Although our implementation of the EnKF assumes a linear observation process, it is important to note that a variety of extensions have been developed to accommodate nonlinear observation and state-transition models. A widely used approach is the Extended Kalman Filter (EKF) (Särkkä and Svensson, 2023), which locally linearizes nonlinear mappings via a first-order Taylor expansion based on the Jacobian matrix. However, this procedure introduces approximation errors inherent to the linearization. To this end, alternative methods have been developed that use only the outputs of the observation process generated from ensemble inputs, without requiring an explicit analytical representation of the observation operator (Katzfuss et al., 2016; Evensen, 2009).

In our experiments, we employ fixed-lag smoothing. Since the primary objective of the present study is to demonstrate that estimating the discretization error mean can be reduced to a problem setting in which the EnKF can be directly applied without approximation, we adopt the classical fixed-lag smoothing approach. We would like to emphasize, however, that other smoothing strategies may also improve performance. In particular, fixed-interval smoothing (Särkkä and Svensson, 2023), which targets the full time series, is a natural alternative for implementing our framework. By exploiting all available observations, it enables inference that more fully utilizes observational information. Recent results have also shown that fixed-point smoothing can be implemented with low computational cost while retaining numerical stability and efficiency (Krämer, 2025). These techniques therefore provide another computationally efficient option.

E Additional Experiment on the Lorenz–96 Model

To evaluate the effectiveness of the proposed method in a high-dimensional setting, we consider the Lorenz–96 model, a standard benchmark system for data assimilation and chaotic dynamics. The system is defined by the following set of coupled ordinary differential equations:

$$\frac{dx_i(t)}{dt} = \big(x_{i+1}(t) - x_{i-2}(t)\big)x_{i-1}(t) - x_i(t) + F, \quad i = 1, \ldots, d_{\mathcal{X}}, \tag{37}$$

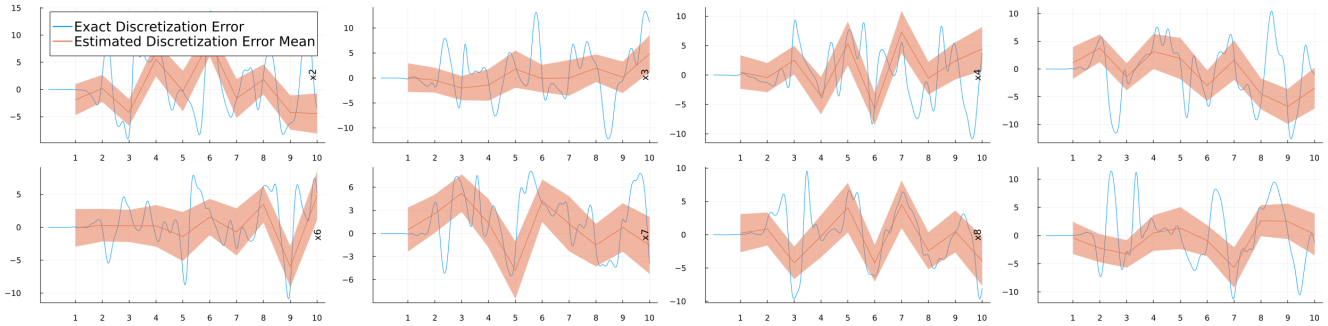
with periodic indexing. In this experiment, we set the system dimension to $d_{\mathcal{X}} = 8$ and the forcing term to $F = 8.0$. The system is numerically solved using the Euler method with a step size of 0.01. The observation operator H is defined as the identity matrix, and observations are assumed to be available at $t = 1.0, 2.0, \ldots, 10.0$, with a fixed covariance matrix $\Gamma = \mathrm{diag}(1.0^2, 1.0^2, \ldots, 1.0^2)$.

The initial prior distribution is given by a multivariate Gaussian: $p(\mu_{t_0}) = \mathcal{N}(0, I)$. The prior distribution is constructed following the same procedure as in Section 4. The hyperparameters (α, β, γ) are selected by maximizing the marginal likelihood over the candidate sets $\alpha \in \{-1.4, -1.2, \ldots, 1.4\}$, $\beta \in \{0.1, 0.2, \ldots, 1.0\}$, $\gamma \in \{0.5, 1.0, \ldots, 4.0\}$. We use an ensemble size of 100 for the Ensemble Kalman Filter (EnKF). For comparison, we implement a particle filter (PF) with 100,000 particles. We attempted to select the hyperparameters using the same candidate sets and marginal-likelihood criterion as in the EnKF experiments. However, the computed log-marginal likelihood overflowed to infinity for all candidate parameter combinations. A possible explanation is severe particle degeneracy, whereby the particles failed to adequately represent the true state and the total particle weight became numerically close to zero. Accordingly, we fix the hyperparameters at $(\alpha, \beta, \gamma) = (1.0, 1.0, 1.0)$.

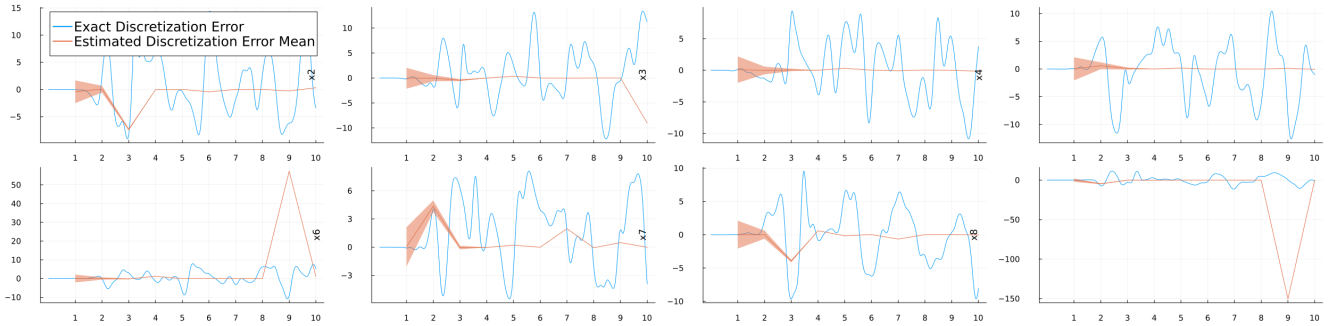
Method	$t=1$	$t=2$	$t=3$	$t=4$	$t=5$	$t=6$	$t=7$	$t=8$	$t=9$	$t=10$
EnKF	0.7329	0.5500	0.6625	0.4909	0.7585	0.7389	0.6655	0.5939	1.0142	1.3573
PF	0.1320	0.3055	0.9591	2.5171	2.2915	2.2845	3.5753	3.1893	1.2671	2.6477

Table 3: Estimation error comparison between EnKF and PF over time steps.

Table 3 reports the performance of EnKF and PF at each observation time. Their performance is evaluated using the absolute difference between the true discretization error and the ensemble mean estimate in EnKF, and between the true discretization error and the particle mean estimate in PF. The results show that the performance of PF deteriorates rapidly after $t \approx 3$. This degradation is likely due to particle degeneracy, which becomes increasingly severe over time and prevents the PF from accurately representing the underlying posterior distribution. In contrast, EnKF maintains relatively stable performance and consistently provides more accurate estimates of the discretization error at later observation times. These results suggest that the proposed EnKF-based approach is more robust than PF for high-dimensional dynamical systems such as the Lorenz–96 model.



(a) EnKF-based posterior on discretization error mean and true discretization error



(b) PF-based posterior on discretization error mean and true discretization error

Figure 6: Discretization error quantification results for the 8-dimensional Lorenz–96 model. The posterior mean and 95% credible interval for the mean discretization error are reported and compared with the true discretization error.

Figure 6 compares the true discretization error with the estimated discretization error means obtained using EnKF and PF. The figure further demonstrates that the EnKF estimate closely follows the trajectory of the true discretization error, whereas PF fails to accurately capture the evolution of the discretization error in the eight-dimensional setting, even when 100,000 particles are used.