Nonlocal Bayesian Modeling of Continuous Spatio-Temporal Dynamics

Jaeyeong Lee[1] **Heeyoung Kim**[*1]

[1]Department of Industrial and Systems Engineering, Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Republic of Korea

Abstract

Real-world spatio-temporal forecasting must handle irregular time points, spatially sparse observations, and the need for uncertainty quantification. This setting is often further compounded by *nonlocal* interactions (long-range spatial coupling). Modeling continuous-space, continuous-time nonlocal dynamics naturally leads to infinite-dimensional integro-differential equations (IDEs), making principled Bayesian inference intractable. We propose the NonLocal Bayesian Spatio-Temporal model (NLBST), a hierarchical Bayesian framework for continuous spatio-temporal fields that learns explicit nonlocal coupling while retaining tractable inference. NLBST represents the latent field via a coordinate-based spatial basis expansion and models the coefficient process with a continuous-time ODE whose learnable linear operator corresponds to a Galerkin reduction of a nonlocal IDE; a Neural ODE residual captures additional nonlinear dynamics. A linear-Gaussian observation model enables Kalman-style sequential updates under missing and irregular observations, while the spatial basis representation enables inductive prediction at unmeasured locations without retraining. Global parameters are learned via variational inference, and uncertainty is handled through a Bayesian hierarchy. Experiments on synthetic and real-world datasets demonstrate strong forecasting and spatial generalization with well-calibrated uncertainty, yielding substantial gains over baselines in strongly nonlocal and partially observed regimes.

1 INTRODUCTION

Spatio-temporal forecasting is central to applications ranging from climate science to public health [Faghmous and Kumar, 2014, Kim et al., 2019, Kumar et al., 2024]. However, real-world sensing imposes practical constraints: observations arrive at irregular time points, are spatially sparse and scattered, and require calibrated uncertainty for downstream decision-making [Rubanova et al., 2019, Jiang, 2018, Pawar et al., 2024]. Addressing these three challenges in a single continuous space-time model with principled uncertainty quantification remains difficult.

Existing approaches address these challenges only partially. Classical statistical models [Cressie, 2015, Wikle et al., 2019, Zammit-Mangion et al., 2022], such as kriging and Gaussian processes, naturally accommodate irregularly located observations and provide principled uncertainty; however, they often rely on separable or weakly structured spatio-temporal covariances, limiting their ability to represent complex, dynamically evolving dependencies. Dynamic spatio-temporal models (DSTMs) [Stroud et al., 2001, Wikle and Hooten, 2010, Koo et al., 2024] explicitly model temporal evolution through state-transition equations, but their discrete-time Markov structure limits applicability under irregular temporal sampling. Deep learning methods [Yu et al., 2017, Che et al., 2018, Jin et al., 2023, Yalavarthi et al., 2024, Zhang et al., 2024, Liu et al., 2026] achieve strong forecasting accuracy at fixed sensor locations, but are often spatially transductive and rely on heuristic uncertainty estimates. Continuous-time neural models [Rubanova et al., 2019, Kidger et al., 2020, Choi et al., 2022, Wu and Chen, 2024], typically based on Neural ODEs [Chen et al., 2018], handle irregular sampling yet lack explicit spatial structure or principled Bayesian inference.

Furthermore, many spatio-temporal systems exhibit *nonlocal* interactions, in which the temporal evolution at one location depends not only on its immediate neighborhood but also on distant regions [Wallace and Gutzler, 1981, Lee and Kim, 2020, Tibau et al., 2022, Jung et al., 2024]. A canoni-

[*]Corresponding author: heeyoungkim@kaist.ac.kr

Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence (UAI 2026), PMLR 337:3355–3375.

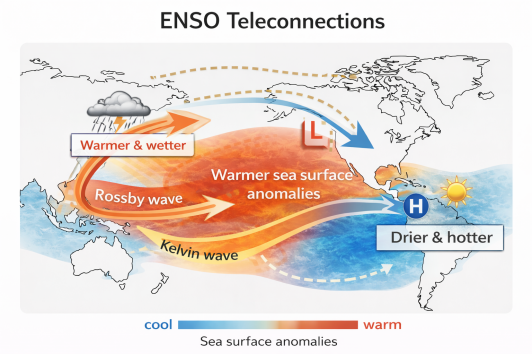

Figure 1: Schematic of ENSO teleconnections illustrating nonlocal spatio-temporal interactions. Tropical Pacific SST anomalies propagate through large-scale wave adjustments (Rossby and Kelvin waves), inducing coherent responses in distant regions and exemplifying long-range dependence beyond local interactions.

cal example is climate teleconnection [Alizadeh, 2024]: sea surface temperature (SST) anomalies in the tropical Pacific associated with the El Niño–Southern Oscillation (ENSO) can reorganize atmospheric circulation and induce coherent responses in remote regions [Yang et al., 2018], including precipitation anomalies over East Africa, storm-track shifts over the North Pacific, and hydroclimate variability over western North America [Yeh et al., 2018] (Figure 1).

A principled formulation of such nonlocal dynamics in continuous space-time is given by the integro-differential equation (IDE) [Laing and Troy, 2003]:

$$\partial_t \mu(x,t) = \int_D G(x,x')\mu(x',t)\,dx' + \mathcal{F}[\mu](x,t), \quad (1)$$

where the kernel $G(x,x')$ encodes how location x' influences the evolution at x, and $\mathcal{F}$ represents additional local dynamics. However, posterior inference is typically intractable: in continuous space, $\mu(\cdot,t)$ is an infinite-dimensional function, and the nonlocal integral term is generally not available in closed form, requiring fine discretization of the domain and expensive numerical evaluation [Stuart, 2010, Dashti and Stuart, 2015].

To address these challenges, we propose a NonLocal Bayesian Spatio-Temporal model (NLBST), a hierarchical Bayesian framework for spatio-temporal fields over continuous space and time that learns nonlocal interactions through a coupled spatio-temporal structure. The proposed model decomposes the latent field into coordinate-based spatial basis functions and a time-varying coefficient vector whose dynamics are governed by an ODE in continuous time. Irregular temporal sampling is handled by propagating the coefficient state across arbitrary time gaps, while spatial prediction at unmeasured locations reduces to evaluating the basis functions at new coordinates.

The coefficient dynamics include an explicit linear coupling term that encodes nonlocal interactions. Conceptually, this can be viewed as a Galerkin reduction of the IDE in Eq.(1), providing a tractable finite-rank representation of long-range dependence, with all inference performed in coefficient space without discretizing the field. A Neural ODE residual [Chen et al., 2018] augments this linear term to capture additional nonlinear dynamics. For uncertainty quantification, a linear-Gaussian observation model admits closed-form Kalman-style updates for sequential Bayesian filtering, while global parameters are learned via variational inference over the Bayesian hierarchy [Wikle et al., 1998].

Our contributions are as follows:

- **Continuous-time Bayesian filtering for irregular and incomplete observations.** We propagate the latent coefficient state via an ODE across arbitrary time gaps. Combined with a linear-Gaussian measurement model, this yields closed-form Kalman-style measurement updates that naturally handle missing entries without the need for imputation.
- **Inductive spatial prediction via spatial basis representation.** Since the learned coefficient dynamics are shared across space, prediction at unmeasured locations requires only evaluating the basis functions at new coordinates, without retraining.
- **Principled uncertainty via Bayesian hierarchy.** Uncertainty is captured through a Bayesian hierarchical model: a linear-Gaussian observation model allows closed-form Kalman-style updates for sequential filtering, while global parameters are learned via variational inference.
- **Learnable nonlocal interactions via a Galerkin reduction.** The linear coupling term in the coefficient ODE induces a finite-rank nonlocal kernel in field space—a Galerkin reduction of a nonlocal IDE—providing an explicit and interpretable pathway for long-range spatial dependence.

We validate NLBST on synthetic dynamics and real-world environmental datasets, demonstrating competitive forecasting accuracy, reliable spatial generalization to unmeasured locations, and well-calibrated predictive uncertainty—particularly excelling in settings with nonlocal interactions and irregular temporal sampling. The code for NLBST is available at `https://github.com/JaeyeongLee1/NLBST`.

2 METHODOLOGY

2.1 PROBLEM DEFINITION

Let $D \subset \mathbb{R}^d$ be a spatial domain. We observe measurements at a set of sensor locations $\mathcal{S}_m = \{x_i^m\}_{i=1}^{S_m} \subset D$ and consider a set of unmeasured locations $\mathcal{S}_u = \{x_j^u\}_{j=1}^{S_u} \subset D$. Let $\{t_k\}_{k=1}^{T}$ denote the observation times (not necessarily equally spaced), and let $\mathbf{y}_k \in \mathbb{R}^{S_m}$ denote the vector of

measurements at time t_k across the measured locations. We allow missing entries in $\mathbf{y}_k$ due to sensor failure, represented by a binary mask $\boldsymbol{\delta}_k \in \{0,1\}^{S_m}$, where $\boldsymbol{\delta}_{k,i} = 1$ indicates that $y_{k,i}$ is missing. Our goals are:

(i) **Forecasting**: predict future values at measured locations for $t > t_T$;
(ii) **Spatial prediction**: infer the signal at $x \in \mathcal{S}_u$ without retraining;
(iii) **Uncertainty quantification**: provide calibrated predictive uncertainty for both tasks.

To achieve these goals, we adopt a three-level Bayesian hierarchy comprising a data model, a process model, and a parameter model.

2.2 LEVEL 1: DATA MODEL

Let $y(x,t)$ be an observation at coordinate $x \in D$ and time t. We assume the observations are noisy measurements of an underlying spatio-temporal latent process:

$$y(x,t) = \mu(x,t) + \epsilon(x,t), \qquad \epsilon(x,t) \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2_{\text{obs}}), \quad (2)$$

where $\mu(x,t)$ is a latent spatio-temporal field and σ^2_{obs} is the observation noise variance.

Vector form at measured locations. At time t_k, let $\mathcal{S}_k \subseteq \mathcal{S}_m$ denote the set of locations with available measurements. Stacking the observations yields

$$\mathbf{y}_k = \boldsymbol{\mu}_k + \boldsymbol{\epsilon}_k, \qquad \boldsymbol{\epsilon}_k \sim \mathcal{N}(\mathbf{0}, \sigma^2_{\text{obs}} I_{n_k}), \quad (3)$$

where $n_k = |\mathcal{S}_k|$ and $\boldsymbol{\mu}_k = \big(\mu(x,t_k)\big)_{x \in \mathcal{S}_k}$.

2.3 LEVEL 2: PROCESS MODEL

We specify the latent process $\mu(x,t)$ through a spatial basis representation with continuous-time coefficient dynamics. This coupled spatio-temporal structure addresses the challenges of inductive spatial prediction and irregular temporal sampling, while naturally enabling a learnable nonlocal interaction mechanism.

2.3.1 Spatial Basis Representation

To enable prediction at arbitrary, unmeasured locations, we represent the latent field as a basis-function expansion over continuous coordinates:

$$\mu(x,t) = \boldsymbol{\phi}(x)^\top \mathbf{z}(t), \quad (4)$$

where $\boldsymbol{\phi}(x) = (\phi_1(x), \ldots, \phi_K(x))^\top \in \mathbb{R}^K$ is a vector of K spatial basis functions evaluated at location $x \in D$, and $\mathbf{z}(t) \in \mathbb{R}^K$ is a time-varying coefficient vector.

2.3.2 Continuous-Time Coefficient Dynamics

To handle irregular temporal sampling, we model the coefficient process in continuous time via an ODE:

$$\frac{d\mathbf{z}(t)}{dt} = A_{\text{NL}}\mathbf{z}(t) + g_\theta(\mathbf{z}(t), t), \quad (5)$$

where $A_{\text{NL}} \in \mathbb{R}^{K\times K}$ is a learnable linear matrix and $g_\theta : \mathbb{R}^K \times \mathbb{R} \to \mathbb{R}^K$ is a neural network parameterized by θ. This continuous-time formulation naturally accommodates irregular observation times: given timestamps $\{t_k\}_{k=1}^T$, we obtain the coefficient state at time t_k by integrating the ODE from the previous state:

$$\hat{\mathbf{z}}_k = \text{ODESolve}(\mathbf{z} \mapsto A_{\text{NL}}\mathbf{z} + g_\theta(\mathbf{z},t),\ \mathbf{z}_{k-1},\ t_{k-1},\ t_k). \quad (6)$$

The integration adapts to the arbitrary time increment $\Delta t_k = t_k - t_{k-1}$ rather than requiring a fixed step size, eliminating the need for imputation. Details of the state update mechanism is provided in Sec. 3.

2.3.3 Induced Nonlocal Coupling Structure

The proposed coupled spatio-temporal modeling induces an explicit nonlocal interaction structure in field space. To make this precise, we express the temporal evolution of the field $\mu(x,t)$ in terms of the coefficient dynamics. From Eq.(4) and Eq.(5), the linear component of the field evolution is:

$$\partial_t \mu(x,t)\big|_{\text{linear}} = \boldsymbol{\phi}(x)^\top A_{\text{NL}}\mathbf{z}(t). \quad (7)$$

Under the spatial basis representation in Eq.(4), the coefficients are given by

$$\mathbf{z}(t) = M^{-1}\int_D \boldsymbol{\phi}(x')\mu(x',t)\,dx', \quad (8)$$

where $M \in \mathbb{R}^{K\times K}$ is the mass matrix with entries $M_{ij} = \int_D \phi_i(x)\phi_j(x)\,dx$. Substituting Eq.(8) into Eq.(7) yields:

$$\begin{aligned}\partial_t \mu(x,t)\big|_{\text{linear}} &= \boldsymbol{\phi}(x)^\top A_{\text{NL}} M^{-1}\int_D \boldsymbol{\phi}(x')\mu(x',t)\,dx' \\ &= \int_D \underbrace{\boldsymbol{\phi}(x)^\top A_{\text{NL}} M^{-1}\boldsymbol{\phi}(x')}_{G(x,x')}\mu(x',t)\,dx'.\end{aligned} \quad (9)$$

This has precisely the form of a nonlocal IDE:

$$\partial_t \mu(x,t) = \int_D G(x,x')\mu(x',t)\,dx' + \mathcal{F}[\mu](x,t), \quad (10)$$

where the induced kernel $G(x,x') = \boldsymbol{\phi}(x)^\top A_{\text{NL}} M^{-1}\boldsymbol{\phi}(x')$ encodes nonlocal spatial interactions, and the neural residual g_θ corresponds to the additional dynamics $\mathcal{F}$. Therefore, learning the coefficient-space matrix A_{NL} is equivalent to learning a finite-rank nonlocal kernel G in field space. From this perspective, the proposed model can be viewed as a Galerkin reduction of a nonlocal IDE.

2.3.4 Stochastic State-Space Formulation

To account for model mismatch and latent process uncertainty, we model the coefficient dynamics as an additive-noise diffusion:

$$d\mathbf{z}(t) = f(\mathbf{z}(t), t)\, dt + \sigma_{\text{proc}}\, d\mathbf{W}_t,$$
$$f(\mathbf{z}, t) = A_{\text{NL}}\mathbf{z} + g_\theta(\mathbf{z}, t), \tag{11}$$

where $\mathbf{W}_t$ is a K-dimensional Wiener process. Given irregular observation times $\{t_k\}$, we use a first-order diffusion discretization to define a tractable Gaussian transition approximation:

$$\hat{\mathbf{z}}_k = \text{ODESolve}\big(\mathbf{z} \mapsto f(\mathbf{z}, t),\, \mathbf{z}_{k-1},\, t_{k-1},\, t_k\big),$$
$$\mathbf{z}_k \mid \mathbf{z}_{k-1} \approx \mathcal{N}\big(\hat{\mathbf{z}}_k,\ \sigma^2_{\text{proc}} \Delta t_k\, I_K\big)\,. \tag{12}$$

The isotropic diffusion covariance $\sigma^2_{\text{proc}}\Delta t_k I_K$ is a simplifying choice that yields a stable and efficient continuous-discrete filtering scheme under irregular sampling. Since σ_{proc} is learned within the variational objective, the effective process uncertainty is calibrated end-to-end to best explain the data likelihood under the chosen variational family.

2.3.5 Spatially Inductive Prediction

The basis representation enables prediction at arbitrary, unmeasured locations $\mathcal{S}_u$. Let $\mathbf{\Phi}(\mathcal{S}_u) \in \mathbb{R}^{S_u \times K}$ be the basis matrix evaluated at $\mathcal{S}_u$. The predictive distribution at $\mathcal{S}_u$ is obtained via linear uncertainty propagation:

$$\mathbb{E}[\boldsymbol{\mu}_u(t) \mid \mathcal{D}] = \mathbf{\Phi}(\mathcal{S}_u)\, \mathbb{E}[\mathbf{z}(t) \mid \mathcal{D}], \tag{13}$$
$$\text{Cov}(\boldsymbol{\mu}_u(t) \mid \mathcal{D}) = \mathbf{\Phi}(\mathcal{S}_u)\, \text{Cov}(\mathbf{z}(t) \mid \mathcal{D})\, \mathbf{\Phi}(\mathcal{S}_u)^\top. \tag{14}$$

Including observation noise, the predictive covariance becomes $\text{Cov}(\mathbf{y}_u(t) \mid \mathcal{D}) = \text{Cov}(\boldsymbol{\mu}_u(t) \mid \mathcal{D}) + \sigma^2_{\text{obs}} I_{S_u}$.

2.3.6 Measurement Model in State-Space Form

At observation time t_k, combining Eq.(3) with Eq.(4) yields the linear-Gaussian measurement model:

$$\mathbf{y}_k = \mathbf{\Phi}_k \mathbf{z}_k + \boldsymbol{\epsilon}_k, \quad \boldsymbol{\epsilon}_k \sim \mathcal{N}(\mathbf{0}, \sigma^2_{\text{obs}} I_{n_k}), \tag{15}$$

where $\mathbf{\Phi}_k := \mathbf{\Phi}(\mathcal{S}_k)$ contains rows $\boldsymbol{\phi}(x)^\top$ for observed locations $x \in \mathcal{S}_k$. Since $\mathbf{y}_k$ depends linearly on $\mathbf{z}_k$ with additive Gaussian noise, the filtering update for $p(\mathbf{z}_k \mid \mathbf{y}_{1:k})$ admits a closed-form Kalman-style conditioning at each observation time, naturally accommodating irregular sampling and missingness. Details of the full predict–update recursion and parameter learning are provided in Sec. 3.

2.4 LEVEL 3: PARAMETER MODEL

Let Θ denote all model parameters: the nonlocal coupling matrix $A_{\text{NL}} \in \mathbb{R}^{K \times K}$, Neural ODE weights θ, noise scales σ_{obs} and σ_{proc}, initial state $\mathbf{z}_0$, and spatial basis parameters (centers $\{c_r\}$ and lengthscale ℓ when learnable). The structural parameters A_{NL}, θ, and the spatial basis parameters are point-estimated via an evidence lower bound (ELBO) objective, while we place weakly informative priors on the remaining parameters:

$$\sigma_{\text{obs}} \sim \text{LogNormal}(\mu_{\text{obs}}, \tau^2_{\text{obs}}), \tag{16}$$
$$\sigma_{\text{proc}} \sim \text{LogNormal}(\mu_{\text{proc}}, \tau^2_{\text{proc}}), \tag{17}$$
$$\mathbf{z}_0 \sim \mathcal{N}(\mathbf{0}, \sigma_0^2 I_K). \tag{18}$$

3 INFERENCE

Given observations $\mathcal{D} = \{(t_k, \mathbf{y}_k)\}_{k=1}^T$ at measured locations, with missingness indicators $\{\boldsymbol{\delta}_k\}$, our goal is to approximate the posterior over the latent coefficient states $\{\mathbf{z}_k\}_{k=1}^T$ and learn the model parameters. Our model combines a continuous-time transition—comprising a linear nonlocal component and a nonlinear Neural ODE residual—with a linear-Gaussian measurement model. This structure enables analytic updates in the observation space while retaining flexible nonlinear dynamics.

In practice, we adopt a Kalman-style variational approximation for the latent states:

$$q(\mathbf{z}_{1:T}, \sigma_{\text{obs}}, \sigma_{\text{proc}}) = q(\sigma_{\text{obs}})\, q(\sigma_{\text{proc}}) \times \prod_{k=1}^{T} q(\mathbf{z}_k \mid \mathcal{D}_{1:k}, \sigma_{\text{obs}}, \sigma_{\text{proc}}), \tag{19}$$

where each $q(\mathbf{z}_k \mid \cdot)$ is a Gaussian distribution whose moments are obtained via a closed-form predict–update recursion implied by the linear-Gaussian measurement model in Eq.(15). The global noise scales $(\sigma_{\text{obs}}, \sigma_{\text{proc}})$ are treated as latent variables with variational distributions. The nonlocal kernel matrix A_{NL}, Neural ODE parameters θ, and basis parameters are learned via gradient-based optimization using stochastic variational inference (SVI).

3.1 VARIATIONAL FILTERING FOR LATENT STATES

Let $\mathbf{z}_k := \mathbf{z}(t_k)$ denote the latent coefficient state at time t_k. Conditioned on $(\sigma_{\text{obs}}, \sigma_{\text{proc}})$ and the model parameters,

$$q(\mathbf{z}_k \mid \mathcal{D}_{1:k}) \approx \mathcal{N}(\mathbf{m}_k^+, \mathbf{P}_k^+),$$

where $(\mathbf{m}_k^+, \mathbf{P}_k^+)$ are obtained by alternating a continuous-time prediction step with an analytic observation update. Latent samples $\mathbf{z}_k$ drawn from the filtered posterior are used to evaluate the ELBO (Eq.(28)). Algorithm 1 summarizes the variational filtering procedure with analytic measurement updates.

Predict. Let $(\mathbf{m}_{k-1}^+, \mathbf{P}_{k-1}^+)$ denote the filtered moments at time t_{k-1}. The predicted mean is obtained by integrating the coefficient dynamics (Eq.(5)):

$$\mathbf{m}_k^- = \mathrm{ODESolve}\big(\mathbf{z} \mapsto A_{\mathrm{NL}}\mathbf{z} + g_\theta(\mathbf{z}, t), \mathbf{m}_{k-1}^+,\ t_{k-1},\ t_k\big). \quad (20)$$

To propagate uncertainty under irregular sampling, we adopt a first-order diffusion approximation within the Gaussian variational family:

$$\mathbf{P}_k^- \approx \mathbf{P}_{k-1}^+ + \sigma_{\mathrm{proc}}^2 \Delta t_k\, I_K, \quad (21)$$

where $\Delta t_k = t_k - t_{k-1}$. This approximation treats the deformation of covariance induced by the nonlinear drift as additional process noise; in practice, σ_{proc} is learned end-to-end via the variational objective to best explain the observed data under the chosen variational family.

Update. At time t_k, let $\mathcal{S}_k$ denote the subset of locations with available measurements (determined by the missingness indicator $\boldsymbol{\delta}_k$). Under Eq.(15), the variational update follows the standard Kalman correction:

$$\mathbf{K}_k = \mathbf{P}_k^- \boldsymbol{\Phi}_k^\top \left(\boldsymbol{\Phi}_k \mathbf{P}_k^- \boldsymbol{\Phi}_k^\top + \sigma_{\mathrm{obs}}^2 I_{n_k}\right)^{-1}, \quad (22)$$

$$\mathbf{m}_k^+ = \mathbf{m}_k^- + \mathbf{K}_k \left(\mathbf{y}_k - \boldsymbol{\Phi}_k \mathbf{m}_k^-\right), \quad (23)$$

$$\mathbf{P}_k^+ = \left(I - \mathbf{K}_k \boldsymbol{\Phi}_k\right) \mathbf{P}_k^-. \quad (24)$$

Missing observations are handled naturally by restricting the update to the observed subset $\mathcal{S}_k$.

Sample. After obtaining $(\mathbf{m}_k^+, \mathbf{P}_k^+)$, we draw

$$\mathbf{z}_k \sim q(\mathbf{z}_k \mid \mathcal{D}_{1:k}) = \mathcal{N}(\mathbf{m}_k^+, \mathbf{P}_k^+), \quad (25)$$

to estimate Monte Carlo expectations in the ELBO and to construct posterior predictive distributions (Sec. 3.4).

3.2 INITIALIZATION FROM THE FIRST OBSERVATION

We construct an initial approximate posterior from the first observation via Bayesian linear regression, combining the prior $\mathbf{z}_0 \sim \mathcal{N}(\mathbf{0}, \sigma_0^2 I_K)$ with the first measurement. Given $\mathbf{y}_1$ and $\boldsymbol{\Phi}_1 := \boldsymbol{\Phi}(\mathcal{S}_1)$, the posterior is

$$\mathbf{P}_1^+ = \left(\sigma_0^{-2} I_K + \sigma_{\mathrm{obs}}^{-2} \boldsymbol{\Phi}_1^\top \boldsymbol{\Phi}_1\right)^{-1}, \quad (26)$$

$$\mathbf{m}_1^+ = \sigma_{\mathrm{obs}}^{-2} \mathbf{P}_1^+ \boldsymbol{\Phi}_1^\top \mathbf{y}_1, \quad (27)$$

which corresponds to the standard Gaussian posterior under the linear observation model (Eq.(15)) with prior covariance $\sigma_0^2 I_K$. This initialization is consistent with the Kalman update Eq.(22)–Eq.(24) applied to a prior predictive $\mathcal{N}(\mathbf{0}, \sigma_0^2 I_K)$.

Algorithm 1 Variational filtering with analytic measurement update

Input: Observations $\{(t_k, \mathbf{y}_k)\}_{k=1}^T$; basis matrices $\{\boldsymbol{\Phi}_k\}$; coupling A_{NL}; residual g_θ; noise $(\sigma_{\mathrm{obs}}, \sigma_{\mathrm{proc}})$; prior σ_0^2.
Output: Filtered $\{(\mathbf{m}_k^+, \mathbf{P}_k^+)\}$ and samples $\{\mathbf{z}_k\}$.
(Initialization)
$\mathbf{P}_1^+ \leftarrow \left(\sigma_0^{-2} I_K + \sigma_{\mathrm{obs}}^{-2} \boldsymbol{\Phi}_1^\top \boldsymbol{\Phi}_1\right)^{-1}$
$\mathbf{m}_1^+ \leftarrow \sigma_{\mathrm{obs}}^{-2} \mathbf{P}_1^+ \boldsymbol{\Phi}_1^\top \mathbf{y}_1$
Sample $\mathbf{z}_1 \sim \mathcal{N}(\mathbf{m}_1^+, \mathbf{P}_1^+)$
for $k = 2$ **to** T **do**
  (Predict)
  $\Delta t_k \leftarrow t_k - t_{k-1}$
  $\mathbf{m}_k^- \leftarrow \mathrm{ODESolve}\big(A_{\mathrm{NL}} \cdot + g_\theta,\ \mathbf{m}_{k-1}^+,\ t_{k-1},\ t_k\big)$
  $\mathbf{P}_k^- \leftarrow \mathbf{P}_{k-1}^+ + \sigma_{\mathrm{proc}}^2 \Delta t_k\, I_K$
  (Update)
  $\mathbf{K}_k \leftarrow \mathbf{P}_k^- \boldsymbol{\Phi}_k^\top \left(\boldsymbol{\Phi}_k \mathbf{P}_k^- \boldsymbol{\Phi}_k^\top + \sigma_{\mathrm{obs}}^2 I_{n_k}\right)^{-1}$
  $\mathbf{m}_k^+ \leftarrow \mathbf{m}_k^- + \mathbf{K}_k \left(\mathbf{y}_k - \boldsymbol{\Phi}_k \mathbf{m}_k^-\right)$
  $\mathbf{P}_k^+ \leftarrow \left(I_K - \mathbf{K}_k \boldsymbol{\Phi}_k\right) \mathbf{P}_k^-$
  (Sample)
  Sample $\mathbf{z}_k \sim \mathcal{N}(\mathbf{m}_k^+, \mathbf{P}_k^+)$
end for
Return $\{(\mathbf{m}_k^+, \mathbf{P}_k^+)\}_{k=1}^T$, $\{\mathbf{z}_k\}_{k=1}^T$

3.3 VARIATIONAL LEARNING OF GLOBAL PARAMETERS

We learn global parameters by maximizing an ELBO via SVI. We use the variational family

$$q(\mathbf{z}_{1:T}, \sigma_{\mathrm{obs}}, \sigma_{\mathrm{proc}}) = q(\sigma_{\mathrm{obs}})\, q(\sigma_{\mathrm{proc}})\, q(\mathbf{z}_{1:T} \mid \sigma_{\mathrm{obs}}, \sigma_{\mathrm{proc}}),$$

where $q(\mathbf{z}_{1:T} \mid \cdot)$ is the structured Gaussian posterior induced by the filtering recursion (Sec. 3.1). The ELBO is

$$\mathcal{L}_{\mathrm{ELBO}} = \mathbb{E}_q\Big[\log p(\mathcal{D}, \mathbf{z}_{1:T}, \sigma) - \log q(\mathbf{z}_{1:T}, \sigma)\Big], \quad (28)$$

where $\sigma{=}(\sigma_{\mathrm{obs}}, \sigma_{\mathrm{proc}})$. The joint model factorizes according to the state-space structure in Eq.(12) and Eq.(15), with likelihoods evaluated only at observed entries.

3.4 POSTERIOR PREDICTIVE DISTRIBUTION

Temporal forecasting. For forecasting at measured locations, we first run the filtering procedure over a historical context window to obtain the filtered posterior moments $(\mathbf{m}_T^+, \mathbf{P}_T^+)$ at the final observed time. We then propagate the latent state forward by integrating the dynamics (Eq.(5)) and injecting process noise according to Eq.(12), without measurement updates. Predictive uncertainty is obtained by Monte Carlo sampling: we draw $(\sigma_{\mathrm{obs}}, \sigma_{\mathrm{proc}})$ from the variational posterior, sample $\mathbf{z}(T)$ from $\mathcal{N}(\mathbf{m}_T^+, \mathbf{P}_T^+)$, roll

forward the latent dynamics via $A_{\mathrm{NL}}\mathbf{z} + g_\theta(\mathbf{z}, t)$, and decode the resulting latent trajectories into the observation space.

Spatial prediction. For unmeasured locations $\mathcal{S}_u$, we evaluate the basis matrix $\mathbf{\Phi}(\mathcal{S}_u)$ and decode the same latent coefficients as in Eq.(13). Predictive moments at $\mathcal{S}_u$ are then obtained via linear uncertainty propagation (Eq.(13)–Eq.(14)).

4 THEORETICAL GUARANTEES

Continuum view of NLBST. We interpret NLBST as a Galerkin approximation of the $\mathcal{H}{=}L^2(D)$-valued stochastic nonlocal evolution:

$$d\mu(t) = \big(\mathcal{G}\mu(t) + \mathcal{F}_\theta(t, \mu(t))\big)\, dt + \Sigma(t, \mu(t))\, dW_t, \quad (29)$$

where $(\mathcal{G}u)(x){=}\int_D G(x, x')u(x')dx'$ is a Hilbert–Schmidt integral operator, $\mathcal{F}_\theta$ is a pointwise nonlinear drift, and W_t is a Q-Wiener process on $\mathcal{H}$. Projecting Eq.(29) onto $V_K{=}\mathrm{span}\{\phi_1, \ldots, \phi_K\}$ yields the coefficient dynamics in Eq.(11).

Assumptions. We assume globally Lipschitz drift and diffusion with linear growth, $\mathrm{Tr}(Q){<}\infty$, and a square-integrable initial condition (Assumptions 1–4 in Appendix A).

Theorem 1 (Well-posedness)**.** *Under Assumptions 1–4, Eq.*(29) *admits a unique mild solution* $\mu \in L^2(\Omega; C([0, T]; \mathcal{H}))$*, where* Ω *is the underlying probability space.*

Theorem 2 (Stability)**.** *Under Assumptions 1–4, for any two mild solutions* $\mu, \tilde{\mu}$ *to Eq.*(29) *with initial conditions* $\mu_0, \tilde{\mu}_0$*, there exists* $C_T > 0$ *such that*

$$\mathbb{E}\Big[\sup_{t\in[0,T]} \|\mu(t) - \tilde{\mu}(t)\|^2\Big] \le C_T\, \mathbb{E}\big[\|\mu_0 - \tilde{\mu}_0\|^2\big].$$

Theorem 3 (Galerkin convergence)**.** *Assume additionally that* $\overline{\cup_{K\ge1} V_K}{=}\mathcal{H}$ *(Assumption 5). Let* μ_K *denote the Galerkin solution (Eq.*(31) *in Appendix A). Then*

$$\lim_{K\to\infty} \mathbb{E}\Big[\sup_{t\in[0,T]} \|\mu(t) - \mu_K(t)\|^2\Big] = 0.$$

Theorem 3 implies that increasing K systematically improves NLBST's finite-dimensional approximation, which we verify empirically in Sec. 6 (Table 5). Detailed analysis and proofs are provided in Appendix A.

5 RELATED WORK

Statistical spatio-temporal modeling. Statistical spatio-temporal models provide principled uncertainty quantification and naturally accommodate sparse spatial measurements [Cressie, 2015, Wikle et al., 2019]. Dynamic spatio-temporal models (DSTMs) cast spatio-temporal fields in Markovian state-space form and admit exact Kalman filtering under a linear-Gaussian measurement model [Wikle and Hooten, 2010], while SPDE-based methods link Gaussian fields to stochastic PDEs for scalable inference [Lindgren et al., 2011, Sigrist et al., 2015, Clarotto et al., 2024]. These frameworks often rely on discrete-time transitions or prespecified (typically local) operators, which can limit expressivity under irregular timestamps and long-range coupling.

Deep spatio-temporal forecasting and neural operators. Deep spatio-temporal forecasting models [Yu et al., 2017, Guo et al., 2019, Liu et al., 2022, Zhang et al., 2024] typically couple temporal modules with spatial encoders (e.g., graph convolutions or attention) and achieve strong accuracy when the sensor set and its connectivity are fixed [Jin et al., 2023]. However, many such approaches are spatially transductive: the spatial component is trained for a specific set of nodes/edges and does not directly support prediction at unseen coordinates without re-parameterization or retraining. Moreover, uncertainty is often quantified via heuristic post-hoc techniques (e.g., ensembles or dropout) rather than through a principled Bayesian treatment. Neural operators learn nonlocal mappings between function spaces and excel in surrogate modeling under dense field supervision [Li et al., 2020b, Kovachki et al., 2023], but they are not designed for sequential data assimilation from sparse, intermittently missing sensor measurements.

Irregular multivariate time series and continuous-time models. Irregular multivariate time series (IMTS) forecasting methods [Che et al., 2018, Rubanova et al., 2019, De Brouwer et al., 2019, Scholz et al., 2023, Yalavarthi et al., 2024, Zhang et al., 2024, Liu et al., 2026] handle missingness and irregular sampling via masking, decay mechanisms, learned time embeddings, or structure-aware aggregation. GraFITi [Yalavarthi et al., 2024] reformulates irregularly sampled multivariate series with missing values as a sparse bipartite graph and performs prediction via GNN-based edge inference, while APN [Liu et al., 2026] introduces a lightweight patch-based framework with time-aware adaptive patch aggregation to regularize irregular histories before decoding. Continuous-time latent dynamics models such as Neural ODEs and CDEs [Chen et al., 2018, Kidger et al., 2020] naturally accommodate irregular timestamps, and latent ODE/SDE frameworks [Rubanova et al., 2019, Li et al., 2020a] combine these dynamics with variational inference.

6 EXPERIMENTS

We evaluate NLBST on synthetic PDE benchmarks and a real-world environmental monitoring dataset. Our experiments are organized around four research questions:

- **(RQ1) Forecasting under partial observations:** How robust is NLBST to randomly missing measurements at varying rates?

- **(RQ2) Spatially inductive prediction:** How well does NLBST generalize to unseen spatial coordinates without retraining?
- **(RQ3) Uncertainty quantification:** Are predictive uncertainties well calibrated?
- **(RQ4) Nonlocal dynamics:** Does NLBST improve performance for dynamics dominated by nonlocal coupling, and does performance improve with basis dimension K consistent with Galerkin convergence (Theorem 3)?

6.1 DATASETS

We use two synthetic PDE datasets that provide dense ground-truth fields, along with one real-world environmental monitoring dataset. For the synthetic datasets, we sample noisy observations at S sensor coordinates and optionally apply missing-at-random masks at rate ρ_m.

(D1) Advection–diffusion (synthetic). We simulate a scalar field $u(x,t)$ over a 2D periodic domain $D = [0,10]^2$ governed by $\frac{\partial u}{\partial t} + \mathbf{v} \cdot \nabla u \;=\; \kappa\,\Delta u + F(x,t)$.

(D2) Nonlocal IDE (synthetic). We generate fields from the nonlocal IDE

$$\partial_t u(x,t) = \int_D G_\star(x,x')\,u(x',t)\,dx' + \kappa\,\Delta u + f(x,t),$$

where $G_\star(x,x') = \sum_{p=1}^{R_\star} \lambda_p \psi_p(x)\psi_p(x')$ is a rank-$R_\star$ kernel designed for long-range spatial coupling; we set $R_\star$=4.

(D3) EPA $PM_{2.5}$ daily (real-world). We use daily mean $PM_{2.5}$ concentrations from the U.S. Environmental Protection Agency (EPA) Air Quality System (AQS) database. We select 50 monitoring stations in the northeastern United States covering 366 days in 2024, applying a $\log(1+y)$ transformation to stabilize variance. The first 300 days are used for training and the remaining 66 days for evaluation.

6.2 EVALUATION PROTOCOL

Rolling-origin forecasting. Given a context window length L and forecast horizon H, we perform rolling-origin evaluation [Tashman, 2000] over the test period with stride r. For each window, we run filtering over the context to obtain the terminal belief $(\mathbf{m}_L^+, \mathbf{P}_L^+)$, then propagate the latent state forward for H steps without measurement updates.

Missingness setup (RQ1, RQ3). On synthetic datasets, we apply i.i.d. random missing masks at rates $\rho_m \in \{10,20,30\}\%$. On **D3**, we use ρ_m=10% in addition to the natural $\sim$2% missing rate.

Spatial generalization setup (RQ2). We designate a fraction of sensor locations as unobserved during training and evaluate predictions at these held-out coordinates without retraining. On **D1**, we hold out $\{10,20,30\}\%$ of sensors with ρ_m=0. On **D3**, we use a 20% holdout.

Nonlocal dynamics setup (RQ4). On the nonlocal IDE benchmark (**D2**), we compare forecasting accuracy under the same rolling-origin protocol and sweep the basis dimension $K \in \{8,12,16,24\}$ to examine how approximation rank affects performance.

Evaluation metrics. We report RMSE and CRPS (Appendix C.2). For **RQ3**, we plot the empirical coverage of 90% prediction intervals and calibration curves. We draw N=100 Monte Carlo samples, and results are averaged over 10 seeds.

6.3 BASELINES

We compare NLBST against baselines spanning discrete-time recurrent models, irregular time-series forecasters, continuous-time latent dynamics, and neural operators:

- **Linear-DSTM** [Wikle and Hooten, 2010]**:** A linear state-space model in the same coefficient space as NLBST (shared spatial basis), but with a linear transition matrix replacing the nonlinear Neural ODE dynamics.
- **GRU-D** [Che et al., 2018]**:** A GRU with trainable exponential decay to handle missing values.
- **Latent ODE** [Rubanova et al., 2019]**:** Continuous-time latent dynamics with a VAE-style encoder; uncertainty is obtained via posterior sampling.
- **FNO** [Li et al., 2020b]**:** A neural operator using spectral convolution along the time axis.
- **GraFITi** [Yalavarthi et al., 2024]**:** An irregular multivariate time series forecaster that converts partially observed series into a sparse bipartite graph and performs forecasting as an edge-weight prediction task using a graph neural network.
- **APN** [Liu et al., 2026]**:** A lightweight patch-based model for irregular time series that forms adaptive patches and aggregates them with time-aware patch aggregation.

For spatial prediction at unmeasured locations, NLBST and LINEAR-DSTM use the shared spatial basis expansion. For the remaining baselines (GRU-D, Latent ODE, FNO, APN, and GraFITi), we apply spatial Gaussian process interpolation to $\mathcal{S}_u$. For uncertainty quantification, we use MC Dropout [Gal and Ghahramani, 2016] for deterministic neural baselines.

6.4 RESULTS

We organize results by research question, integrating findings from the synthetic datasets (**D1**–**D2**) and the real-world $PM_{2.5}$ dataset (**D3**).

RQ1: Forecasting under partial observations. On **D1**, NLBST achieves the lowest RMSE and CRPS across all missing rates, with stable performance as ρ_m increases from 10% to 30% (Table 1). This robustness stems from continuous-time ODE dynamics that bridge missing obser-

vations without imputation, combined with Kalman-style updates that restrict conditioning to the observed subset at each time step. In contrast, FNO degrades sharply with increasing missingness (RMSE $1.65 \rightarrow 3.09$), as its spectral convolutions assume a regular temporal grid and are less suited to missing entries. Linear-DSTM exhibits high variance at ρ_m=10% (std 2.05), with occasional instability in a subset of random seeds. On **D3** (Table 2), NLBST achieves the lowest RMSE and CRPS at both measured and unmeasured locations under 20% spatial holdout and 10% missingness, improving over Linear-DSTM by $\sim$8% in RMSE at both measured (0.48 vs. 0.52) and unmeasured stations (0.49 vs. 0.53). GraFITi is competitive at measured locations (RMSE 0.50) but degrades at unmeasured ones (0.58), indicating limited spatial generalization. The improvement over Linear-DSTM—which shares the same spatial basis and Kalman updates—isolates the contribution of nonlinear continuous-time dynamics via the Neural ODE residual.

Table 1: (**RQ1**, **RQ3**) Forecasting under missing-at-random observations on **D1** (advection–diffusion). Mean (std) over 10 seeds. Best in **bold**; second-best underlined.

Method	$\rho_m = 10\%$	$\rho_m = 20\%$	$\rho_m = 30\%$
RMSE			
NLBST (Ours)	**0.589 (0.108)**	**0.620 (0.087)**	**0.613 (0.103)**
Linear-DSTM	2.543 (2.053)	0.795 (0.382)	0.740 (0.358)
GRU-D	1.076 (0.137)	0.880 (0.112)	0.754 (0.090)
Latent ODE	4.280 (0.245)	4.331 (0.268)	4.229 (0.305)
FNO	1.651 (0.803)	2.746 (0.581)	3.089 (0.551)
GraFITi	0.702 (0.304)	0.713 (0.320)	0.703 (0.329)
APN	0.648 (0.062)	0.655 (0.056)	0.654 (0.062)
CRPS			
NLBST (Ours)	**0.369 (0.055)**	**0.383 (0.046)**	**0.384 (0.051)**
Linear-DSTM	1.706 (1.549)	0.470 (0.204)	0.434 (0.201)
GRU-D	0.743 (0.118)	0.569 (0.089)	0.473 (0.069)
Latent ODE	3.325 (0.251)	3.368 (0.261)	3.271 (0.299)
FNO	1.048 (0.601)	1.957 (0.576)	2.343 (0.586)
GraFITi	0.439 (0.178)	0.447 (0.191)	0.442 (0.206)
APN	0.424 (0.147)	0.415 (0.136)	0.419 (0.150)

Table 2: (**RQ1**, **RQ2**, **RQ3**) $PM_{2.5}$ daily forecasting (**D3**): mean (std) over 10 seeds with 20% spatial holdout and ρ_m=10%. Best in **bold**; second-best underlined.

	Measured		**Unmeasured**	
Model	RMSE	CRPS	RMSE	CRPS
NLBST (Ours)	**0.48** (0.01)	**0.28** (0.01)	**0.49** (0.02)	**0.29** (0.01)
Linear-DSTM	0.52 (0.03)	0.29 (0.01)	0.53 (0.02)	0.30 (0.01)
GRU-D	1.12 (0.03)	0.72 (0.02)	1.03 (0.04)	0.78 (0.04)
Latent ODE	1.20 (0.07)	0.72 (0.06)	1.09 (0.07)	0.78 (0.08)
FNO	1.12 (0.08)	0.73 (0.06)	0.90 (0.09)	0.65 (0.08)
GraFITi	0.50 (0.01)	0.34 (0.01)	0.58 (0.11)	0.39 (0.07)
APN	0.63 (0.21)	0.44 (0.19)	0.70 (0.21)	0.50 (0.19)

RQ2: Spatially inductive prediction. On **D1**, NLBST achieves the lowest RMSE at measured locations across all holdout levels and also the lowest RMSE at held-out, unmeasured locations (Table 3), confirming inductive prediction via the spatial basis $\mu(x,t) = \boldsymbol{\phi}(x)^\top \mathbf{z}(t)$ without retraining. On **D3** (Table 2), the unmeasured-location RMSE (0.49) is comparable to the measured-location RMSE (0.48), indicating that the learned coefficient dynamics generalize to unseen coordinates through basis evaluation alone. By contrast, the deep learning baselines rely on post-hoc spatial interpolation for held-out stations and incur substantially larger errors at unmeasured locations.

RQ3: Uncertainty quantification. NLBST achieves the lowest CRPS across various datasets and settings (Tables 1, 2, 4), indicating strong probabilistic forecasting performance. On **D3**, predictive intervals exhibit sensible time-varying behavior, widening during volatile periods (Figure 2). To assess NLBST's uncertainty quantification, we construct Gaussian prediction intervals at each nominal level using the posterior mean and standard deviation from Monte Carlo samples, and compute the fraction of test observations contained within each interval as the empirical coverage. Across nominal levels from 10% to 99%, the empirical coverage closely aligns with the nominal level (Figure 3), demonstrating that the three-level Bayesian hierarchy yields well-calibrated uncertainty.

Table 3: (**RQ2**) Spatially inductive prediction on **D1** (advection–diffusion). RMSE at measured (m) and held-out unmeasured (u) locations. Mean (std) over 10 seeds; **bold** = best, underlined = second best per column.

	5 held out		10 held out		15 held out	
Model	m	u	m	u	m	u
NLBST (Ours)	**0.53** *(0.01)*	**0.71** *(0.28)*	**0.52** *(0.04)*	**0.78** *(0.14)*	**0.55** *(0.03)*	**0.76** *(0.12)*
Linear-DSTM	1.04 *(0.35)*	2.39 *(0.81)*	1.27 *(0.43)*	3.12 *(0.62)*	0.84 *(0.25)*	2.02 *(0.47)*
GRU-D	1.93 *(0.27)*	2.01 *(0.58)*	1.80 *(0.13)*	1.99 *(0.16)*	1.96 *(0.17)*	2.10 *(0.27)*
Latent ODE	4.21 *(0.25)*	4.36 *(0.22)*	4.39 *(0.18)*	4.56 *(0.17)*	4.32 *(0.08)*	4.48 *(0.16)*
FNO	2.27 *(0.58)*	2.37 *(0.23)*	2.50 *(0.52)*	2.59 *(0.60)*	2.37 *(0.43)*	2.45 *(0.38)*
GraFITi	0.71 *(0.28)*	1.02 *(0.54)*	0.70 *(0.28)*	0.94 *(0.44)*	0.80 *(0.36)*	1.01 *(0.46)*
APN	0.71 *(0.20)*	0.95 *(0.40)*	0.63 *(0.11)*	0.88 *(0.31)*	0.70 *(0.11)*	0.94 *(0.30)*

Table 4: (**RQ3**, **RQ4**) Forecasting on the nonlocal IDE benchmark (**D2**). Mean (std) over 10 seeds.

Model	RMSE	CRPS
NLBST (Ours)	**9.54** (3.36)	**6.84** (2.26)
Linear-DSTM	14.42 (0.95)	10.10 (1.22)
GRU-D	20.15 (4.72)	16.79 (5.05)
Latent ODE	28.31 (9.27)	24.04 (9.55)
FNO	29.60 (10.69)	25.25 (10.62)
GraFITi	15.29 (4.40)	10.48 (3.64)
APN	20.88 (4.78)	17.28 (5.14)

RQ4: Nonlocal dynamics. On **D2**, NLBST substantially outperforms all baselines, reducing the RMSE of the best competitor by over 33% (Table 4: 9.54 vs. Linear-DSTM 14.42; CRPS 6.84 vs. 10.10). These gains highlight that explicitly learning long-range coupling via A_{NL} is particularly beneficial when the underlying dynamics are dominated by nonlocal interactions. Notably, FNO—a neural-operator baseline designed to model nonlocal operator structure—still performs poorly on **D2** (RMSE 29.60), suggesting that capturing nonlocal structure alone is insufficient in this sparse-sensor sequential forecasting setting without principled temporal dynamics. The basis-dimension sweep (Table 5) shows monotonic improvement as K increases ($11.70 \to 9.54$), consistent with the Galerkin convergence theory (Theorem 3).

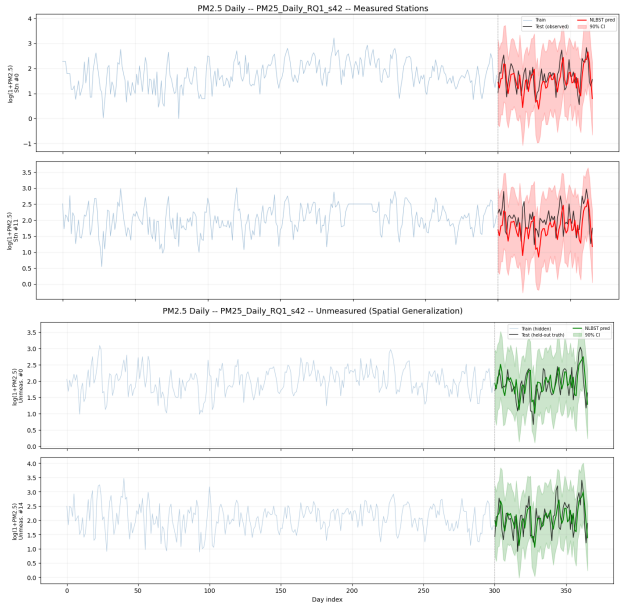

Figure 2: (**RQ3**) EPA $PM_{2.5}$ predictive trajectories with 90% intervals. **Top:** the upper two subplots (red-shaded), corresponding to measured stations. **Bottom:** the lower two subplots (green-shaded), corresponding to unmeasured stations. Shaded bands denote 90% predictive intervals.

Table 5: (**RQ4**) Basis dimension sweep on **D2**. RMSE decreases monotonically with K, consistent with Galerkin convergence theory (Theorem 3).

	K=8	K=12	K=16	K=24
RMSE	11.70 (3.69)	10.70 (3.66)	10.08 (3.38)	**9.54** (3.36)

Computational cost and effect of K. Table 6 reports accuracy and per-epoch cost as a function of the basis dimension K on **D3**. Accuracy improves up to $K \approx 24$ and then saturates (measured RMSE $0.527 \to 0.494$ from K=8 to 24, with $< 1\%$ further gain up to K=64), supporting K=24 as a sensible default. Although the Kalman update has $\mathcal{O}(K^3)$ worst-case complexity, per-epoch runtime

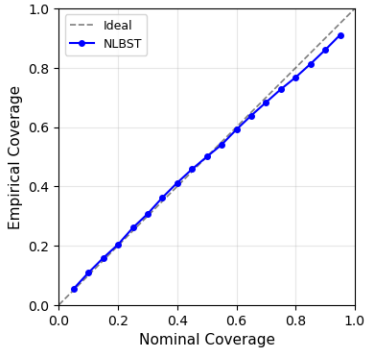

Figure 3: (**RQ3**) Calibration plot for EPA $PM_{2.5}$: nominal vs. empirical coverage. The dashed line indicates ideal calibration, and the solid curve (NLBST) tracks it closely.

Table 6: (K**-sensitivity & Computational Cost**) Accuracy and per-epoch cost vs. basis dimension K on **D3**. Accuracy saturates around K=24 while runtime and memory stay nearly constant. RMSE: mean (std) over seeds; cost: mean (std) over epochs.

K	RMSE (m)	RMSE (u)	Time/ep. (s)	Mem (GB)
8	0.527 (0.005)	0.521 (0.006)	6.71 (0.13)	0.03
16	0.499 (0.004)	0.485 (0.013)	6.55 (0.03)	0.03
24	0.494 (0.006)	0.482 (0.009)	6.63 (0.09)	0.04
32	0.494 (0.009)	0.479 (0.013)	6.57 (0.06)	0.04
64	0.490 (0.010)	0.480 (0.018)	6.58 (0.05)	0.06

stays near-constant ($\sim$6.6 s) and memory grows only mildly ($0.03 \to 0.06$ GB) as K grows by a factor of eight, indicating that fixed overhead dominates the $\mathcal{O}(K^3)$ cost within the tested range. NLBST thus attains strong accuracy at low computational cost.

7 CONCLUSION

We introduced NLBST, a continuous-space, continuous-time hierarchical Bayesian model that learns explicit nonlocal coupling while retaining tractable inference and Kalman-style measurement updates. NLBST combines a coordinate-based spatial basis with an ODE-driven latent process and variational learning of global uncertainty parameters, enabling forecasting, inductive prediction at unseen coordinates, and calibrated uncertainty in a unified framework. By jointly learning nonlocal interactions and continuous-time dynamics within a Bayesian hierarchy, NLBST bridges the gap between expressive neural dynamics and principled uncertainty quantification, which prior methods typically treat in isolation. Experiments on synthetic dynamics and real-world EPA $PM_{2.5}$ data show consistently strong performance, especially in strongly nonlocal and partially observed regimes. The future work includes structured regularization of A_{NL} and learnable basis-dimension selection.

Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (2023R1A2C2005453, RS-2023-00218913).

References

Omid Alizadeh. A review of enso teleconnections at present and under future global warming. *Wiley Interdisciplinary Reviews: Climate Change*, 15(1):e861, 2024.

Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. Recurrent neural networks for multivariate time series with missing values. *Scientific reports*, 8(1):6085, 2018.

Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. *Advances in neural information processing systems*, 31, 2018.

Jeongwhan Choi, Hwangyong Choi, Jeehyun Hwang, and Noseong Park. Graph neural controlled differential equations for traffic forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 6367–6374, 2022.

Lucia Clarotto, Denis Allard, Thomas Romary, and Nicolas Desassis. The spde approach for spatio-temporal datasets with advection and diffusion. *Spatial Statistics*, 62:100847, 2024.

Noel Cressie. *Statistics for spatial data*. John Wiley & Sons, 2015.

Masoumeh Dashti and Andrew M Stuart. The bayesian approach to inverse problems. In *Handbook of uncertainty quantification*, pages 1–118. Springer, 2015.

Edward De Brouwer, Jaak Simm, Adam Arany, and Yves Moreau. Gru-ode-bayes: Continuous modeling of sporadically-observed time series. *Advances in neural information processing systems*, 32, 2019.

James H Faghmous and Vipin Kumar. Spatio-temporal data mining for climate data: Advances, challenges, and opportunities. *Data mining and knowledge discovery for big data: Methodologies, challenge and opportunities*, pages 83–116, 2014.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016.

Shengnan Guo, Youfang Lin, Ning Feng, Chao Song, and Huaiyu Wan. Attention based spatial-temporal graph convolutional networks for traffic flow forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 922–929, 2019.

Zhe Jiang. A survey on spatial prediction methods. *IEEE transactions on knowledge and Data Engineering*, 31(9): 1645–1664, 2018.

Guangyin Jin, Yuxuan Liang, Yuchen Fang, Zezhi Shao, Jincai Huang, Junbo Zhang, and Yu Zheng. Spatio-temporal graph neural networks for predictive learning in urban computing: A survey. *IEEE transactions on knowledge and data engineering*, 36(10):5388–5408, 2023.

Junsub Jung, Sungil Kim, and Heeyoung Kim. Spatially-correlated time series clustering using location-dependent dirichlet process mixture model. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 17 (1):e11649, 2024.

Patrick Kidger, James Morrill, James Foster, and Terry Lyons. Neural controlled differential equations for irregular time series. *Advances in neural information processing systems*, 33:6696–6707, 2020.

Keunseo Kim, Vinnam Kim, and Heeyoung Kim. Spatiotemporal auto-regressive model for origin–destination air passenger flows. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 182(3):1003–1016, 2019.

Wonmo Koo, Eun-Yeol Ma, and Heeyoung Kim. Deep latent factor model for spatio-temporal forecasting. *Technometrics*, 66(3):470–482, 2024.

Nikola Kovachki, Zongyi Li, Burigede Liu, Kamyar Azizzadenesheli, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Neural operator: Learning maps between function spaces with applications to pdes. *Journal of Machine Learning Research*, 24(89):1–97, 2023.

Rahul Kumar, Manish Bhanu, João Mendes-Moreira, and Joydeep Chandra. Spatio-temporal predictive modeling techniques for different domains: a survey. *ACM Computing Surveys*, 57(2):1–42, 2024.

Carlo R Laing and William C Troy. Pde methods for nonlocal models. *SIAM Journal on Applied Dynamical Systems*, 2(3):487–516, 2003.

Youngmin Lee and Heeyoung Kim. Bayesian nonparametric joint mixture model for clustering spatially correlated time series. *Technometrics*, 62(3):313–329, 2020.

Xuechen Li, Ting-Kam Leonard Wong, Ricky TQ Chen, and David Duvenaud. Scalable gradients for stochastic differential equations. In *International conference on artificial intelligence and statistics*, pages 3870–3882. PMLR, 2020a.

Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Fourier neural operator for parametric partial differential equations. *arXiv preprint arXiv:2010.08895*, 2020b.

Finn Lindgren, Håvard Rue, and Johan Lindström. An explicit link between gaussian fields and gaussian markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 73(4):423–498, 2011.

Dachuan Liu, Jin Wang, Shuo Shang, and Peng Han. Msdr: Multi-step dependency relation networks for spatial temporal forecasting. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, pages 1042–1050, 2022.

Xvyuan Liu, Xiangfei Qiu, Xingjian Wu, Zhengyu Li, Chenjuan Guo, Jilin Hu, and Bin Yang. Rethinking irregular time series forecasting: A simple yet effective baseline. In *AAAI*, 2026.

Shivaji D Pawar, Varsha S Pawar, and Satheesh Abimannan. Handling uncertainty in spatiotemporal data. In *Spatiotemporal Data Analytics and Modeling: Techniques and Applications*, pages 69–87. Springer, 2024.

Yulia Rubanova, Ricky TQ Chen, and David K Duvenaud. Latent ordinary differential equations for irregularly-sampled time series. *Advances in neural information processing systems*, 32, 2019.

Randolf Scholz, Stefan Born, Nghia Duong-Trung, Mariano Nicolas Cruz-Bournazou, and Lars Schmidt-Thieme. Latent linear odes with neural kalman filtering for irregular time series forecasting. 2023.

Fabio Sigrist, Hans R Künsch, and Werner A Stahel. Stochastic partial differential equation based modelling of large space–time data sets. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 77(1): 3–33, 2015.

Jonathan R Stroud, Peter Müller, and Bruno Sansó. Dynamic models for spatiotemporal data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63 (4):673–689, 2001.

Andrew M Stuart. Inverse problems: a bayesian perspective. *Acta numerica*, 19:451–559, 2010.

Leonard J Tashman. Out-of-sample tests of forecasting accuracy: an analysis and review. *International journal of forecasting*, 16(4):437–450, 2000.

Xavier-Andoni Tibau, Christian Reimers, Andreas Gerhardus, Joachim Denzler, Veronika Eyring, and Jakob Runge. A spatiotemporal stochastic climate model for benchmarking causal discovery methods for teleconnections. *Environmental Data Science*, 1:e12, 2022.

John M Wallace and David S Gutzler. Teleconnections in the geopotential height field during the northern hemisphere winter. *Monthly weather review*, 109(4):784–812, 1981.

Christopher K Wikle and Mevin B Hooten. A general science-based framework for dynamical spatio-temporal models. *Test*, 19(3):417–451, 2010.

Christopher K Wikle, L Mark Berliner, and Noel Cressie. Hierarchical bayesian space-time models. *Environmental and ecological statistics*, 5(2):117–154, 1998.

Christopher K Wikle, Andrew Zammit-Mangion, and Noel Cressie. *Spatio-temporal statistics with R*. Chapman and Hall/CRC, 2019.

Jiajia Wu and Ling Chen. Continuously evolving graph neural controlled differential equations for traffic forecasting. *arXiv preprint arXiv:2401.14695*, 2024.

Vijaya Krishna Yalavarthi, Kiran Madhusudhanan, Randolf Scholz, Nourhan Ahmed, Johannes Burchert, Shayan Jawed, Stefan Born, and Lars Schmidt-Thieme. Grafiti: Graphs for forecasting irregularly sampled time series. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 16255–16263, 2024.

Song Yang, Zhenning Li, Jin-Yi Yu, Xiaoming Hu, Wenjie Dong, and Shan He. El niño–southern oscillation and its impact in the changing climate. *National Science Review*, 5(6):840–857, 2018.

Sang-Wook Yeh, Wenju Cai, Seung-Ki Min, Michael J McPhaden, Dietmar Dommenget, Boris Dewitte, Matthew Collins, Karumuri Ashok, Soon-Il An, Bo-Young Yim, et al. Enso atmospheric teleconnections and their response to greenhouse gas forcing. *Reviews of Geophysics*, 56(1):185–206, 2018.

Bing Yu, Haoteng Yin, and Zhanxing Zhu. Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. *arXiv preprint arXiv:1709.04875*, 2017.

Andrew Zammit-Mangion, Tin Lok James Ng, Quan Vu, and Maurizio Filippone. Deep compositional spatial models. *Journal of the American Statistical Association*, 117 (540):1787–1808, 2022.

Weijia Zhang, Le Zhang, Jindong Han, Hao Liu, Yanjie Fu, Jingbo Zhou, Yu Mei, and Hui Xiong. Irregular traffic time series forecasting based on asynchronous spatio-temporal graph convolutional networks. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 4302–4313, 2024.

Nonlocal Bayesian Modeling of Continuous Spatio-Temporal Dynamics (Supplementary Material)

Jaeyeong Lee[1] **Heeyoung Kim**[*1]

[1]Department of Industrial and Systems Engineering, Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Republic of Korea

A THEORETICAL ANALYSIS

This section provides the detailed analysis and proofs for the theoretical guarantees stated in Sec. 4, (Theorems 1–3).

A.1 PROBLEM SETUP

Nonlocal stochastic evolution equations. Let $\mathcal{H} = L^2(D)$ with norm $\|\cdot\|$ and inner product $\langle\cdot,\cdot\rangle$. A nonlocal kernel $G(x,x')$ induces the integral operator $\mathcal{G}:\mathcal{H}\to\mathcal{H}$,

$$(\mathcal{G}u)(x) \;=\; \int_D G(x,x')\,u(x')\,dx'.$$

We consider the $\mathcal{H}$-valued nonlocal stochastic evolution of the latent field $\mu(t)\in\mathcal{H}$:

$$d\mu(t) = \big(\mathcal{G}\mu(t) + \mathcal{F}_\theta(t,\mu(t))\big)\,dt + \Sigma(t,\mu(t))\,dW_t, \tag{30}$$

where W is a Q-Wiener process on $\mathcal{H}$. The drift and diffusion are defined by applying time-dependent scalar-valued functions $f_\theta(t,\cdot):\mathbb{R}\to\mathbb{R}$ and $\sigma(t,\cdot):\mathbb{R}\to\mathbb{R}$ pointwise to the field:

$$(\mathcal{F}_\theta(t,u))(x) = f_\theta(t,u(x)), \qquad (\Sigma(t,u)v)(x) = \sigma(t,u(x))\,v(x), \quad u,v\in\mathcal{H}.$$

Galerkin projection. Let $\{\phi_i\}_{i\ge1}\subset\mathcal{H}$ and $V_K = \operatorname{span}\{\phi_1,\dots,\phi_K\}$. For simplicity, let $P_K:\mathcal{H}\to V_K$ denote the orthogonal projection (non-orthonormal bases can be handled via the Gram/mass matrix). The Galerkin projected dynamics on V_K are defined by

$$d\mu_K(t) = P_K\big(\mathcal{G}\mu_K(t) + \mathcal{F}_\theta(t,\mu_K(t))\big)\,dt + P_K\Sigma(t,\mu_K(t))\,dW_t, \quad \mu_K(0) = P_K\mu_0. \tag{31}$$

A.2 PRELIMINARIES

Semigroups of linear operators. Let H be a Hilbert space.

Definition 1 (Semigroup)**.** *A family $\{S(t)\}_{t\ge0}$ of bounded linear operators $S(t):H\to H$ is a* semigroup *if*

1. $S(0) = I$,
2. $S(t+s) = S(t)S(s)$ *for all* $s,t\ge0$.

Definition 2 (Infinitesimal generator)**.** *The* infinitesimal generator *of a semigroup $\{S(t)\}_{t\ge0}$ is the operator A defined by*

$$Au \;:=\; \lim_{h\downarrow0}\frac{S(h)u-u}{h},$$

on the domain $D(A) = \{u\in H : \text{the limit exists in } H\}$.

If A is a bounded operator on H, then $S(t) = e^{tA}$ is a *uniformly continuous* semigroup satisfying

$$\|S(t)\|_{\mathcal{L}(H)} \;\le\; e^{t\|A\|}. \tag{32}$$

Mild solutions (variation of constants). Consider a semilinear stochastic evolution on a Hilbert space H of the form

$$du(t) = \big(Au(t) + F(t, u(t))\big)\, dt + \Sigma(t, u(t))\, dW_t, \qquad u(0) = u_0 \in H, \tag{33}$$

where A generates a semigroup $S(t)$ on H and W is a Q-Wiener process. A process u is called a *mild solution* if it satisfies the variation-of-constants formula

$$u(t) = S(t)\, u_0 + \int_0^t S(t-s)\, F(s, u(s))\, ds + \int_0^t S(t-s)\, \Sigma(s, u(s))\, dW_s. \tag{34}$$

In our setting Eq.(30), we take $H = \mathcal{H}$, $A = \mathcal{G}$, and $F = \mathcal{F}_\theta$.

Banach fixed-point theorem. Let (X, d) be a complete metric space and $J : X \to X$ satisfy $d(Ju, Jv) \le \gamma\, d(u, v)$ for some $\gamma \in [0, 1)$. Then J has a unique fixed point in X.

Grönwall's inequality. If $z(t) \le \alpha + \beta \int_0^t z(s)\, ds$ with $\alpha, \beta \ge 0$, then $z(t) \le \alpha e^{\beta t}$ for all $t \in [0, T]$.

Burkholder-Davis-Gundy inequality (BDG inequality). For a predictable process $B(t)$ with $B(t)Q^{1/2}$ Hilbert–Schmidt,

$$\mathbb{E}\Big[\sup_{t\in[0,T]} \Big\| \int_0^t B(s)\, dW_s \Big\|^2\Big] \le C_{\mathrm{BDG}}\, \mathbb{E} \int_0^T \|B(s)Q^{1/2}\|_{\mathrm{HS}}^2\, ds. \tag{35}$$

A.3 ASSUMPTIONS

Assumption 1 (Bounded nonlocal operator)**.** *The kernel G is square-integrable on $D \times D$:*

$$\int_D \int_D |G(x, x')|^2\, dx\, dx' < \infty,$$

so $\mathcal{G}$ is Hilbert–Schmidt (hence bounded) on $\mathcal{H}$.

Assumption 2 (Lipschitz and linear growth)**.** *There exist constants $L, C > 0$ such that for all $t \in [0, T]$ and $u, v \in \mathbb{R}$,*

$$\begin{aligned} |f_\theta(t, u) - f_\theta(t, v)| &\le L|u - v|, \\ |\sigma(t, u) - \sigma(t, v)| &\le L|u - v|, \\ |f_\theta(t, u)|^2 + |\sigma(t, u)|^2 &\le C(1 + |u|^2). \end{aligned}$$

Assumption 3 (Q-Wiener noise regularity)**.** *Let W be a Q-Wiener process on $\mathcal{H}$ with $\mathrm{Tr}(Q) < \infty$. Assume $\Sigma(t, u)Q^{1/2}$ is Hilbert–Schmidt for all $t \in [0, T]$ and $u \in \mathcal{H}$, and there exist $L_\Sigma, C_\Sigma > 0$ such that for all $u, v \in \mathcal{H}$,*

$$\|\Sigma(t, u)Q^{1/2} - \Sigma(t, v)Q^{1/2}\|_{\mathrm{HS}} \le L_\Sigma \|u - v\|, \qquad \|\Sigma(t, u)Q^{1/2}\|_{\mathrm{HS}}^2 \le C_\Sigma(1 + \|u\|^2).$$

Assumption 4 (Initial condition)**.** *The initial condition $\mu_0 \in L^2(\Omega; \mathcal{H})$ is independent of W and satisfies $\mathbb{E}[\|\mu_0\|^2] < \infty$.*

Assumption 5 (Approximation property)**.** $\overline{\bigcup_{K\ge1} V_K} = \mathcal{H}$.

A.4 PROOF OF THEOREM 1 (WELL-POSEDNESS)

Proof. Let $S(t) = e^{t\mathcal{G}}$. Since $\mathcal{G}$ is bounded (Assumption 1), Eq.(32) with $A = \mathcal{G}$ implies

$$\|S(t)\|_{\mathcal{L}(\mathcal{H})} \le e^{t\|\mathcal{G}\|} \le M_T := e^{T\|\mathcal{G}\|}, \qquad t \in [0, T].$$

We first note that the pointwise drift map is Lipschitz on $\mathcal{H} = L^2(D)$. By Assumption 2, for all $t \in [0, T]$ and $u, v \in L^2(D)$,

$$\begin{aligned} \|\mathcal{F}_\theta(t, u) - \mathcal{F}_\theta(t, v)\|_{L^2(D)}^2 &= \int_D \big|f_\theta(t, u(x)) - f_\theta(t, v(x))\big|^2\, dx \\ &\le \int_D \big(L|u(x) - v(x)|\big)^2\, dx \\ &= L^2 \|u - v\|_{L^2(D)}^2, \end{aligned}$$

and hence $\|\mathcal{F}_\theta(t, u) - \mathcal{F}_\theta(t, v)\|_{L^2(D)} \le L\|u - v\|_{L^2(D)}$. Similarly, Assumption 3 gives for $u, v \in \mathcal{H}$,

$$\|\big(\Sigma(t, u) - \Sigma(t, v)\big)Q^{1/2}\|_{\mathrm{HS}} \le L_\Sigma \|u - v\|.$$

Existence and uniqueness on a short interval. Fix a horizon $\tau \in (0, T]$ and consider the Banach space $\mathcal{X}_\tau := L^2(\Omega; C([0,\tau]; \mathcal{H}))$ with norm $\|u\|_{\mathcal{X}_\tau}^2 := \mathbb{E}\big[\sup_{t\in[0,\tau]} \|u(t)\|^2\big]$. Define the mild-map $\mathcal{T}$ by

$$(\mathcal{T}u)(t) := S(t)\mu_0 + \int_0^t S(t-s)\mathcal{F}_\theta(s, u(s))\, ds + \int_0^t S(t-s)\Sigma(s, u(s))\, dW_s.$$

A fixed point of $\mathcal{T}$ is a mild solution (cf. Eq.(34) with $A = \mathcal{G}$ and $F = \mathcal{F}_\theta$).

We use the BDG inequality in $\mathcal{H}$: for predictable $B(t)$ with $B(t)Q^{1/2}$ Hilbert–Schmidt,

$$\mathbb{E}\Big[\sup_{t\in[0,\tau]} \Big\| \int_0^t B(s)\, dW_s \Big\|^2\Big] \le C_{\mathrm{BDG}} \mathbb{E} \int_0^\tau \|B(s)Q^{1/2}\|_{\mathrm{HS}}^2\, ds.$$

For $u, v \in \mathcal{X}_\tau$, let $\Delta F(s) := \mathcal{F}_\theta(s, u(s)) - \mathcal{F}_\theta(s, v(s))$. Then

$$\begin{aligned}
&\mathbb{E}\left[\sup_{t\in[0,\tau]} \Big\| \int_0^t S(t-s)\Delta F(s)\, ds\Big\|^2\right] \\
&\le \mathbb{E}\left[\sup_{t\in[0,\tau]} \Big(\int_0^t \|S(t-s)\Delta F(s)\|\, ds\Big)^2\right] \qquad \text{(Minkowski: } \|\textstyle\int g\| \le \int \|g\|\text{)} \\
&\le \mathbb{E}\left[\sup_{t\in[0,\tau]} \Big(\int_0^t \|S(t-s)\|_{\mathcal{L}(\mathcal{H})}\, \|\Delta F(s)\|\, ds\Big)^2\right] \qquad \text{(sub-multiplicativity)} \\
&\le \mathbb{E}\left[\sup_{t\in[0,\tau]} \Big(\int_0^t M_T\, \|\Delta F(s)\|\, ds\Big)^2\right] \qquad \text{(semigroup bound)} \\
&= M_T^2\, \mathbb{E}\left[\sup_{t\in[0,\tau]} \Big(\int_0^t \|\Delta F(s)\|\, ds\Big)^2\right] \\
&\le M_T^2\, \mathbb{E}\left[\sup_{t\in[0,\tau]} \Big(t \int_0^t \|\Delta F(s)\|^2\, ds\Big)\right] \qquad \text{(Cauchy–Schwarz: } (\textstyle\int_0^t a)^2 \le t\int_0^t a^2\text{)} \\
&\le M_T^2\, \tau \int_0^\tau \mathbb{E}\|\Delta F(s)\|^2\, ds \\
&\le M_T^2\, \tau \int_0^\tau L^2\, \mathbb{E}\|u(s) - v(s)\|^2\, ds \qquad \text{(Lipschitz of } \mathcal{F}_\theta\text{)} \\
&\le M_T^2\, L^2\, \tau^2\, \mathbb{E}\Big[\sup_{s\in[0,\tau]} \|u(s) - v(s)\|^2\Big] \qquad \text{(since } \textstyle\int_0^\tau f(s)\, ds \le \tau \sup_s f(s)\text{)} \\
&= M_T^2\, L^2\, \tau^2\, \|u - v\|_{\mathcal{X}_\tau}^2. \qquad \text{(definition of } \|\cdot\|_{\mathcal{X}_\tau}\text{)}
\end{aligned}$$

For the stochastic term, BDG and Assumption 3 yield

$$\begin{aligned}
&\mathbb{E}\Big[\sup_{t\in[0,\tau]} \Big\| \int_0^t S(t-s)\big(\Sigma(s, u(s)) - \Sigma(s, v(s))\big)\, dW_s\Big\|^2\Big] \\
&\quad \le C_{\mathrm{BDG}} M_T^2 \int_0^\tau \mathbb{E}\|\big(\Sigma(s, u(s)) - \Sigma(s, v(s))\big)Q^{1/2}\|_{\mathrm{HS}}^2\, ds \;\le\; C_{\mathrm{BDG}} M_T^2 L_\Sigma^2 \tau \|u - v\|_{\mathcal{X}_\tau}^2.
\end{aligned}$$

Combining the two bounds gives

$$\|\mathcal{T}u - \mathcal{T}v\|_{\mathcal{X}_\tau}^2 \le \Big(M_T^2 L^2 \tau^2 + C_{\mathrm{BDG}} M_T^2 L_\Sigma^2 \tau\Big)\|u - v\|_{\mathcal{X}_\tau}^2.$$

Hence for $\tau > 0$ sufficiently small, $\mathcal{T}$ is a contraction on $\mathcal{X}_\tau$. By Banach's fixed-point theorem, there exists a unique mild solution on $[0, \tau]$. Repeating the argument on successive intervals yields a unique mild solution on $[0, T]$. □

A.5 PROOF OF THEOREM 2 (STABILITY)

Proof. Let $\mu, \tilde{\mu}$ be mild solutions with initial conditions $\mu_0, \tilde{\mu}_0$. From the mild form and $S(t) = e^{t\mathcal{G}}$, for any $t \in [0, T]$ we have

$$\mu(t) - \tilde{\mu}(t) = S(t)(\mu_0 - \tilde{\mu}_0) + \int_0^t S(t-s)\big(\mathcal{F}_\theta(s, \mu(s)) - \mathcal{F}_\theta(s, \tilde{\mu}(s))\big)\, ds + \int_0^t S(t-s)\big(\Sigma(s, \mu(s)) - \Sigma(s, \tilde{\mu}(s))\big)\, dW_s.$$

Let $\Delta(t) := \mu(t) - \tilde{\mu}(t)$, $\Delta F(s) := \mathcal{F}_\theta(s, \mu(s)) - \mathcal{F}_\theta(s, \tilde{\mu}(s))$, and $\Delta\Sigma(s) := \Sigma(s, \mu(s)) - \Sigma(s, \tilde{\mu}(s))$.
For all $t \in [0, T]$,

$$\begin{aligned}
z(t) &:= \mathbb{E}\Big[\sup_{r\in[0,t]} \|\mu(r) - \tilde{\mu}(r)\|^2\Big] \\
&\le 3\,\mathbb{E}\Big[\sup_{r\in[0,t]} \|S(r)(\mu_0 - \tilde{\mu}_0)\|^2\Big] + 3\,\mathbb{E}\Big[\sup_{r\in[0,t]} \Big\|\int_0^r S(r-s)\Delta F(s)\, ds\Big\|^2\Big] \\
&\quad + 3\,\mathbb{E}\Big[\sup_{r\in[0,t]} \Big\|\int_0^r S(r-s)\Delta\Sigma(s)\, dW_s\Big\|^2\Big] \qquad \text{(mild form, } (a+b+c)^2 \le 3(a^2+b^2+c^2)\text{)} \\
&\le 3M_T^2\,\mathbb{E}\|\mu_0 - \tilde{\mu}_0\|^2 + 3M_T^2\, t\int_0^t \mathbb{E}\|\Delta F(s)\|^2\, ds + 3C_{\mathrm{BDG}} M_T^2 \int_0^t \mathbb{E}\|\Delta\Sigma(s)Q^{1/2}\|_{\mathrm{HS}}^2\, ds \\
&\qquad\qquad \text{(semigroup bound, Minkowski, BDG)} \\
&\le 3M_T^2\,\mathbb{E}\|\mu_0 - \tilde{\mu}_0\|^2 + 3M_T^2\, L^2 t\int_0^t \mathbb{E}\|\mu(s) - \tilde{\mu}(s)\|^2\, ds + 3C_{\mathrm{BDG}} M_T^2\, L_\Sigma^2 \int_0^t \mathbb{E}\|\mu(s) - \tilde{\mu}(s)\|^2\, ds \\
&\qquad\qquad \text{(Lipschitz)} \\
&\le 3M_T^2\,\mathbb{E}\|\mu_0 - \tilde{\mu}_0\|^2 + 3M_T^2\big(L^2 t + C_{\mathrm{BDG}} L_\Sigma^2\big)\int_0^t z(s)\, ds \qquad \text{(since } \|\Delta(s)\|^2 \le \sup_{\rho\le s}\|\Delta(\rho)\|^2\text{)} \\
&\le 3M_T^2\,\mathbb{E}\|\mu_0 - \tilde{\mu}_0\|^2 + 3M_T^2\big(L^2 T + C_{\mathrm{BDG}} L_\Sigma^2\big)\int_0^t z(s)\, ds.
\end{aligned}$$

Let $C_0 := 3M_T^2$ and $C_1 := 3M_T^2\big(L^2 T + C_{\mathrm{BDG}} L_\Sigma^2\big)$.
Then $z(t) \le C_0\,\mathbb{E}\|\mu_0 - \tilde{\mu}_0\|^2 + C_1 \int_0^t z(s)\, ds$, so Grönwall's inequality yields

$$\mathbb{E}\Big[\sup_{t\in[0,T]} \|\mu(t) - \tilde{\mu}(t)\|^2\Big] = z(T) \le C_0 e^{C_1 T}\mathbb{E}\|\mu_0 - \tilde{\mu}_0\|^2.$$

This implies the claimed stability bound ($C_T = C_0 e^{C_1 T}$). □

A.6 PROOF OF THEOREM 3 (GALERKIN CONVERGENCE)

Proof. Let $P_K : \mathcal{H} \to V_K$ be the orthogonal projection. Since the Galerkin dynamics evolve in the finite-dimensional space V_K and the projected drift/diffusion inherit the same Lipschitz and growth bounds, Eq.(31) admits a unique solution μ_K (by the same argument as Theorem 1).

Define the error $e_K(t) := \mu(t) - \mu_K(t)$. Because $\mathcal{G}$ is bounded, the mild formulation is equivalent to the integral form of Eq.(30):

$$\mu(t) = \mu_0 + \int_0^t \big(\mathcal{G}\mu(s) + \mathcal{F}_\theta(s, \mu(s))\big)\, ds + \int_0^t \Sigma(s, \mu(s))\, dW_s,$$

and similarly from Eq.(31),

$$\mu_K(t) = P_K\mu_0 + \int_0^t P_K\big(\mathcal{G}\mu_K(s) + \mathcal{F}_\theta(s, \mu_K(s))\big)\, ds + \int_0^t P_K\Sigma(s, \mu_K(s))\, dW_s.$$

Subtracting gives

$$e_K(t) = (I - P_K)\mu_0 + \int_0^t (I - P_K)\big(\mathcal{G}\mu(s) + \mathcal{F}_\theta(s, \mu(s))\big)\, ds + \int_0^t (I - P_K)\Sigma(s, \mu(s))\, dW_s$$
$$+ \int_0^t P_K\Big(\mathcal{G}e_K(s) + \big(\mathcal{F}_\theta(s, \mu(s)) - \mathcal{F}_\theta(s, \mu_K(s))\big)\Big)\, ds + \int_0^t P_K\big(\Sigma(s, \mu(s)) - \Sigma(s, \mu_K(s))\big)\, dW_s.$$

Let $Z_K(t) := \mathbb{E}\big[\sup_{r\in[0,t]} \|e_K(r)\|^2\big]$. Using Cauchy–Schwarz for the deterministic integrals, the BDG inequality for the stochastic integrals, the boundedness of $\mathcal{G}$, and the Lipschitz bounds on $\mathcal{F}_\theta$ and $\Sigma(\cdot)Q^{1/2}$, we obtain for all $t \in [0, T]$,

$$Z_K(t) \;\le\; \alpha_K(t) \;+\; \beta \int_0^t Z_K(s)\, ds, \tag{36}$$

where

$$\alpha_K(t) := C_0\Bigg(\mathbb{E}\|(I - P_K)\mu_0\|^2 + t\int_0^t \mathbb{E}\|(I - P_K)\mathcal{G}\mu(s)\|^2\, ds + t\int_0^t \mathbb{E}\|(I - P_K)\mathcal{F}_\theta(s, \mu(s))\|^2\, ds$$
$$+ C_{\mathrm{BDG}} \int_0^t \mathbb{E}\|(I - P_K)\Sigma(s, \mu(s))Q^{1/2}\|_{\mathrm{HS}}^2\, ds\Bigg),$$
$$\beta := C_1\Big(t\|\mathcal{G}\|^2 + tL^2 + C_{\mathrm{BDG}}L_\Sigma^2\Big) \;\le\; C_1\Big(T\|\mathcal{G}\|^2 + TL^2 + C_{\mathrm{BDG}}L_\Sigma^2\Big),$$

for constants $C_0, C_1 > 0$ independent of K (depending only on norm inequalities). Applying Grönwall's inequality to Eq.(36) yields

$$Z_K(T) \le \alpha_K(T) \exp(\beta T).$$

Since $P_K \to I$ strongly on $\mathcal{H}$ (Assumption 5) and μ has finite second moments on $[0, T]$ (Theorem 1 and linear growth), we have $\alpha_K(T) \to 0$ as $K \to \infty$ by dominated convergence theorem. Therefore $Z_K(T) \to 0$, i.e.,

$$\mathbb{E}\Big[\sup_{t\in[0,T]} \|\mu(t) - \mu_K(t)\|^2\Big] \to 0 \quad \text{as } K \to \infty.$$

□

B ELBO DETAILS

The nonlocal kernel matrix A_{NL}, Neural ODE parameters θ, and basis parameters are optimized by maximizing an evidence lower bound (ELBO) via stochastic variational inference (SVI). Let the variational family factorize as $q(\sigma_{\mathrm{obs}}, \sigma_{\mathrm{proc}}, \mathbf{z}_{1:T}) = q(\sigma_{\mathrm{obs}})\, q(\sigma_{\mathrm{proc}}) \prod_{k=1}^T q(\mathbf{z}_k \mid \mathcal{D}_{1:k}, \sigma_{\mathrm{obs}}, \sigma_{\mathrm{proc}})$, where each $q(\mathbf{z}_k \mid \cdot)$ is Gaussian with moments given by the predict–update recursion in Sec. 3.1. Specifically, we maximize

$$\mathcal{L}_{\mathrm{ELBO}} = \mathbb{E}_{q(\sigma_{\mathrm{obs}}, \sigma_{\mathrm{proc}}, \mathbf{z}_{1:T})}\Bigg[\sum_{k=1}^T \log p(\mathbf{y}_k \mid \mathbf{z}_k, \sigma_{\mathrm{obs}}) + \log p(\mathbf{z}_1) + \sum_{k=2}^T \log p(\mathbf{z}_k \mid \mathbf{z}_{k-1}, \sigma_{\mathrm{proc}}) - \log q(\mathbf{z}_{1:T} \mid \sigma_{\mathrm{obs}}, \sigma_{\mathrm{proc}})\Bigg]$$
$$- \mathrm{KL}(q(\sigma_{\mathrm{obs}}) \,\|\, p(\sigma_{\mathrm{obs}})) - \mathrm{KL}(q(\sigma_{\mathrm{proc}}) \,\|\, p(\sigma_{\mathrm{proc}}))\,, \tag{37}$$

where $p(\sigma_{\mathrm{obs}})$ and $p(\sigma_{\mathrm{proc}})$ are specified by the priors in Level 3, $p(\mathbf{z}_1) = \mathcal{N}(\mathbf{0}, \sigma_0^2 I_K)$, and $p(\mathbf{z}_k \mid \mathbf{z}_{k-1}, \sigma_{\mathrm{proc}})$ is the Gaussian transition approximation in Eq.(12). The likelihood term is evaluated on non-missing entries under the Gaussian observation model Eq.(15).

Concrete forms of each term. We now expand the individual terms in Eq.(37). All Gaussian log-densities below use the convention $\log \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = -\frac{d}{2}\log(2\pi) - \frac{1}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})$.

Observation log-likelihood. At time t_k, let $\mathcal{S}_k$ denote the set of n_k observed locations. Under Eq.(15):

$$\log p(\mathbf{y}_k \mid \mathbf{z}_k, \sigma_{\mathrm{obs}}) = -\frac{n_k}{2}\log(2\pi\sigma_{\mathrm{obs}}^2) - \frac{1}{2\sigma_{\mathrm{obs}}^2}\|\mathbf{y}_k - \boldsymbol{\Phi}_k \mathbf{z}_k\|^2, \tag{38}$$

where $\mathbf{\Phi}_k = \mathbf{\Phi}(\mathcal{S}_k) \in \mathbb{R}^{n_k \times K}$ contains rows corresponding to observed locations only.

Transition log-density. Under the discretized transition Eq.(12), for $k \geq 2$:

$$\log p(\mathbf{z}_k \mid \mathbf{z}_{k-1}, \sigma_{\text{proc}}) = -\frac{K}{2} \log(2\pi\sigma_{\text{proc}}^2 \Delta t_k) - \frac{\|\mathbf{z}_k - \hat{\mathbf{z}}_k\|^2}{2\sigma_{\text{proc}}^2 \Delta t_k}, \tag{39}$$

where $\hat{\mathbf{z}}_k = \text{ODESolve}(\mathbf{z} \mapsto A_{\text{NL}}\mathbf{z} + g_\theta(\mathbf{z}, t),\ \mathbf{z}_{k-1},\ t_{k-1},\ t_k)$.

Prior log-density.

$$\log p(\mathbf{z}_1) = -\frac{K}{2} \log(2\pi\sigma_0^2) - \frac{\|\mathbf{z}_1\|^2}{2\sigma_0^2}. \tag{40}$$

Variational entropy. Since $q(\mathbf{z}_{1:T} \mid \sigma_{\text{obs}}, \sigma_{\text{proc}}) = \prod_{k=1}^{T} q(\mathbf{z}_k \mid \mathcal{D}_{1:k}, \sigma_{\text{obs}}, \sigma_{\text{proc}})$ with $q(\mathbf{z}_k \mid \cdot) = \mathcal{N}(\mathbf{m}_k^+, \mathbf{P}_k^+)$, where $(\mathbf{m}_k^+, \mathbf{P}_k^+)$ depend on $(\sigma_{\text{obs}}, \sigma_{\text{proc}})$ through the filtering recursion, the conditional entropy decomposes as:

$$-\mathbb{E}_q[\log q(\mathbf{z}_{1:T} \mid \sigma_{\text{obs}}, \sigma_{\text{proc}})] = \sum_{k=1}^{T} \frac{1}{2} \log |2\pi e\, \mathbf{P}_k^+| = \sum_{k=1}^{T} \left(\frac{K}{2} \log(2\pi e) + \frac{1}{2} \log |\mathbf{P}_k^+| \right). \tag{41}$$

C ADDITIONAL EXPERIMENTAL DETAILS

C.1 DATASET DETAILS

D1: Advection–diffusion. We use constant velocity $\mathbf{v} = (0.5,\ 0.3)$, diffusion coefficient $\kappa = 0.1$, and forcing $F(x, t)$ given by two spatially localized Gaussian sources with oscillating amplitudes. We discretize on a 64×64 grid with $\Delta t_{\text{sim}} = 0.01$ and save snapshots every $\Delta t_{\text{save}} = 1.0$, yielding $T = 151$ frames over $[0, 150]$. Sensor observations are obtained by evaluating u at $S = 50$ irregularly placed coordinates and adding Gaussian noise with $\sigma_{\text{noise}} = 0.1$. The first 75% of frames are used for training and the remainder for evaluation, with rolling-origin windows of context length $L{=}50$, forecast horizon $H{=}10$, and stride $r{=}5$.

D2: Nonlocal IDE. We simulate on $D = [0, 1]^2$ with periodic boundary conditions on a 64×64 grid, $T_{\text{final}} = 20$, and step size $\Delta t = 0.1$, integrated with a fourth-order Runge–Kutta (RK4) method. We observe $S = 36$ sensors on a regular spatial grid and add Gaussian noise with $\sigma_{\text{noise}} = 0.05$. The local diffusion coefficient is $\kappa = 0.01$, and $f(x, t)$ consists of two spatially localized Gaussian sources with oscillating amplitudes.

The ground-truth kernel $G_\star$ has rank $R_\star = 4$ with eigenvalues $\lambda_p = 1/(p{+}1)$. The four modes ψ_p are defined on normalized coordinates $(\tilde{x}, \tilde{y}) \in [0, 1]^2$:

$$\begin{aligned}
\psi_1 &= \sin(2\pi\tilde{y}) && \text{(north–south coupling)},\\
\psi_2 &= \sin(2\pi\tilde{x}) && \text{(east–west coupling)},\\
\psi_3 &= \sin(2\pi\tilde{x})\sin(2\pi\tilde{y}) && \text{(four-corner interaction)},\\
\psi_4 &= \sin(4\pi\tilde{x})\cos(4\pi\tilde{y}) && \text{(multi-scale coupling)},
\end{aligned}$$

each L^2-normalized over D. These modes create nonlocal interactions that couple spatially distant regions, mimicking teleconnection-like patterns where the evolution at one location depends on remote areas rather than only the immediate neighborhood. The first 75% of time steps are used for training and the remainder for evaluation, with context length $L{=}50$, forecast horizon $H{=}10$, and stride $r{=}5$.

D3: EPA $PM_{2.5}$. Daily mean $PM_{2.5}$ concentrations are obtained from the U.S. EPA Air Quality System (AQS) for the year 2024. We select stations in the northeastern United States (latitude $35°$- $45°$N, longitude $70°$–$85°$W) with at least 80% temporal coverage, yielding $S{=}50$ stations over $T{=}366$ days. Values are capped at $150\,\mu\text{g/m}^3$ prior to the $\log(1{+}y)$ transform. The natural missing rate is approximately 1–2%; we additionally introduce artificial missing-at-random during training. We use $L{=}5$ days and $H{=}1$ day with stride $r{=}1$.

C.2 METRICS

RMSE. Root mean squared error between the predictive mean and ground truth, averaged over all forecast windows, horizons, and sensors:

$$\mathrm{RMSE} = \sqrt{\frac{1}{|\mathcal{W}|}\sum_{w\in\mathcal{W}}\frac{1}{H\cdot S_w}\sum_{h=1}^{H}\sum_{s=1}^{S_w}\big(\hat{y}_{s,h}^{(w)} - y_{s,h}^{(w)}\big)^2},$$

where $\mathcal{W}$ indexes rolling windows, H is the forecast horizon, S_w is the number of evaluated sensors in window w, and $\hat{y}$ denotes the predictive mean.

CRPS. The continuous ranked probability score (CRPS) measures the compatibility of a predictive CDF F with the observed value y:

$$\mathrm{CRPS}(F, y) = \int_{-\infty}^{\infty}\big(F(z) - \mathbf{1}[z \geq y]\big)^2 dz.$$

For NLBST, the predictive distribution is obtained by propagating $N{=}100$ posterior samples through the nonlinear ODE dynamics, yielding a distribution-free empirical CRPS that captures non-Gaussian effects from hyperparameter uncertainty and nonlinear propagation. For baselines whose uncertainty estimates reduce to a predictive mean and variance (MC Dropout, VAE sampling), we compute CRPS under a Gaussian approximation in closed form via $\mathrm{CRPS}(\mathcal{N}(\mu,\sigma^2), y) = \sigma\big[\bar{z}\,(2\Phi(\bar{z}) - 1) + 2\varphi(\bar{z}) - \pi^{-1/2}\big]$, where $\bar{z} = (y-\mu)/\sigma$.

C.3 IMPLEMENTATION DETAILS

Model configuration. We use $K{=}16$ RBF basis functions for **D1**, and $K{=}24$ Fourier basis functions for **D2** and **D3**. For **D2** and **D3**, we use a truncated Fourier basis on $[0,1]^d$ which is L^2-orthonormal. For the RBF basis (**D1**), basis centers are initialized from sensor coordinates, and the lengthscale is set based on inter-sensor distances; basis parameters are fixed during training for stability. The Neural ODE residual g_θ is a 2-layer multi-layer perceptron (MLP) with hidden dimension 64 and tanh activation. ODE integration uses a fourth-order Runge–Kutta (RK4) method. All spatial coordinates are normalized to $[0,1]^d$ prior to basis construction and GP kriging. Min-max normalization is applied per sensor for all methods including NLBST.

Training and prediction. We optimize the ELBO via stochastic variational inference (SVI) with Adam (learning rate 10^{-3}) and gradient clipping (max norm 1). NLBST is trained for 200 epochs with early stopping of patience 10. The variational guide performs full-covariance Kalman filtering with inline predict-update steps (Sec. 3). For probabilistic forecasts, we draw $N{=}100$ Monte Carlo samples from the variational posterior and decode via the spatial basis. All results are averaged over 10 runs with different random seeds.

C.4 BASELINE IMPLEMENTATION DETAILS

All baselines are trained on the same data splits and evaluated under the same protocol as NLBST. Since most baselines were not originally designed for simultaneous irregular time series forecasting, spatial generalization, and uncertainty quantification, we adapt each method to provide predictions at unmeasured locations and calibrated predictive intervals. Table 7 summarizes the key modeling capabilities of each method.

Linear-DSTM [Wikle and Hooten, 2010]. A discrete-time linear dynamic spatio-temporal model $\mathbf{z}_t = A\mathbf{z}_{t-1} + \boldsymbol{\epsilon}_t$, $\mathbf{y}_t = \boldsymbol{\Phi}\mathbf{z}_t + \nu_t$, using the same spatial basis as NLBST. Inference includes the standard Kalman filter. This baseline isolates the contribution of nonlinear dynamics and continuous-time propagation.

GRU-D [Che et al., 2018]. GRU with trainable exponential decay on both input and hidden states to handle irregular time gaps. Observations are imputed via a learned convex combination of the last observed value and the empirical mean, weighted by a time-decay gate. Since GRU-D operates on a fixed sensor set, spatial prediction at unmeasured locations uses GP kriging with an RBF kernel. We used MC dropout for uncertainty estimation.

Table 7: Modeling properties of baselines. ✓: native support; △: partial or approximate support; ×: not supported (adaptation required).

Model	Spatially inductive	Irregular-time	Principled UQ	Nonlocal
NLBST (Ours)	✓	✓	✓	✓
Linear-DSTM	✓	×	✓	×
GRU-D	×	✓	×	×
Latent ODE	×	✓	△	×
FNO	×	×	×	△
GraFITi	×	✓	×	△
APN	×	✓	×	×

Latent ODE [Rubanova et al., 2019]. An ODE-RNN encoder processes observations in forward time to produce a variational posterior $q(\mathbf{z}_0)$; latent states are then decoded via a Neural ODE. Uncertainty is obtained by sampling from the VAE posterior. Spatial prediction at unmeasured locations uses GP kriging of the decoded measured-location outputs.

FNO [Li et al., 2020b]. A Fourier Neural Operator with 1-D spectral convolution layers operating along the temporal axis over all measured sensors jointly. At each layer, the input passes through a spectral convolution (FFT $\rightarrow$ linear mixing in frequency $\rightarrow$ iFFT) and a pointwise linear transform, combined via a residual connection. Forecasting is autoregressive: at each horizon step, the model appends its own prediction and re-applies the Fourier blocks. FNO achieves global mixing via spectral convolution but does not learn an explicit nonlocal kernel. Uncertainty is provided via MC dropout, and spatial prediction uses GP kriging.

GraFITi [Yalavarthi et al., 2024]. A bipartite graph attention model that represents observations as edges connecting time nodes and channel (sensor) nodes. Missing observations correspond to absent edges, so irregular and incomplete data are handled natively. Message passing alternates between channel$\rightarrow$time and time$\rightarrow$channel attention, with edge embeddings updated at each layer. We adapt GraFITi for forecasting by marking future time–channel positions as target edges with unknown values; the model predicts these entries in a single forward pass. GraFITi achieves implicit cross-channel coupling via bipartite attention but lacks an explicit spatial nonlocal mechanism. Uncertainty is obtained via MC dropout. Spatial prediction at unmeasured locations uses GP kriging with an RBF kernel whose length scale is set to 0.3.

APN [Liu et al., 2026]. A patch-based architecture using Time-Aware Patch Aggregation (TAPA): per-variable soft patches with learnable sigmoid boundaries aggregate context observations, followed by cross-patch attention and a time-query decoder. The decoder conditions on learned time encodings to produce predictions at arbitrary future times. Uncertainty is obtained via MC dropout. Spatial prediction at unmeasured locations uses GP kriging with the same RBF kernel configuration as GraFITi.

Spatial interpolation for transductive baselines. All baselines except Linear-DSTM are spatially transductive: they produce outputs only at the training sensor locations. For these methods, predictions at unmeasured locations are obtained via GP kriging weights $\mathbf{W} = K(\mathcal{S}_u, \mathcal{S}_m)\, K(\mathcal{S}_m, \mathcal{S}_m)^{-1}$ with an RBF kernel (length scale 0.3) and a nugget of 10^{-2} for numerical stability. In contrast, NLBST and Linear-DSTM predict at unmeasured locations by evaluating the spatial basis at new coordinates, requiring no post-hoc interpolation.

Uncertainty quantification. Linear-DSTM and NLBST provide principled Bayesian uncertainty via the Kalman filtering posterior. Latent ODE obtains uncertainty through VAE posterior sampling. GRU-D, FNO, GraFITi, and APN use MC dropout (50 forward passes) to approximate predictive distributions.

Training details. All baselines are trained with Adam (lr 10^{-3}) and gradient clipping (max norm 1.0). For fair comparison, all methods use the same context length L and forecast horizon H as NLBST within each experimental setting. All models are trained for 200 epochs with early stopping of patience 10, and use the same preprocessing as NLBST. All results are averaged over 10 runs with different random seeds.

D ADDITIONAL EXPERIMENTAL RESULTS

D.1 $PM_{2.5}$ SUPPLEMENTARY RESULTS

Table 8 reports the robustness of NLBST to increasing missing rates on **D3** with 40% spatial holdout. RMSE remains stable across $\rho_m \in \{0, 10, 20, 30\}\%$ for both measured and unmeasured locations, confirming that the NLBST gracefully handles observation sparsity by widening the prior uncertainty when measurements are absent.

Table 9 examines spatial generalization on **D3** as the fraction of held-out stations increases. Performance at unmeasured locations remains comparable to measured locations even at 40% holdout, demonstrating that the spatial basis provides effective inductive prediction to unseen coordinates on real-world data.

Table 8: (**RQ1**) Robustness to missing data on **D3** ($PM_{2.5}$) (NLBST, 40% spatial holdout). RMSE and CRPS in $\log(1+y)$ space (mean $\pm$ std over 10 seeds).

ρ_m	RMSE (meas.)	RMSE (unmeas.)
0%	0.554 ± 0.045	0.558 ± 0.053
10%	0.554 ± 0.047	0.558 ± 0.054
20%	0.551 ± 0.044	0.555 ± 0.052
30%	0.546 ± 0.045	0.552 ± 0.052

Table 9: (**RQ2**) Spatial generalization on **D3** ($PM_{2.5}$) (NLBST). RMSE in $\log(1+y)$ space (mean $\pm$ std over 10 seeds).

Holdout %	RMSE (measured)	RMSE (unmeasured)
0%	0.567 ± 0.003	—
20%	0.542 ± 0.026	0.523 ± 0.025
40%	0.552 ± 0.047	0.557 ± 0.053

D.2 DYNAMICS ABLATION (RQ4)

Table 10 isolates the contribution of each component in the NLBST dynamics on **D2** and **D3**. On **D2**, removing the nonlocal coupling (neural only) degrades CRPS most severely, while removing the neural residual (linear only) primarily affects RMSE, indicating that the two components play complementary roles in prediction accuracy and uncertainty calibration. The same pattern holds on the real-world **D3** dataset: both ablations increase measured-location RMSE relative to the full model ($0.480 \rightarrow 0.495/0.496$), confirming that both A_{NL} and g_θ provide complementary performance gains rather than introducing redundant complexity.

Table 10: (**RQ3**, **RQ4**) Dynamics ablation on **D2** (Nonlocal IDE) and **D3** ($PM_{2.5}$). Mean (std) over 10 seeds; **D3** reports RMSE at measured locations. Best per column in **bold**.

	D2		**D3**
Dynamics	RMSE↓	CRPS↓	RMSE↓
$A_{\mathrm{NL}}\mathbf{z} + g_\theta$ (full)	**9.54 (3.36)**	**6.84 (2.26)**	**0.480 (0.010)**
$A_{\mathrm{NL}}\mathbf{z}$ (linear only)	10.85 (2.70)	7.20 (1.40)	0.495 (0.006)
$g_\theta(\mathbf{z}, t)$ (neural only)	10.84 (2.73)	8.03 (1.42)	0.496 (0.006)

D.3 LONG-HORIZON FORECASTING

To assess robustness beyond the short-horizon setting in the main experiments (Table 2, $H{=}1$), we evaluate long-horizon forecasting on **D3** with $H \in \{10, 30, 50\}$. The setup matches the main D3 experiments, except that we use $T_{\mathrm{train}}{=}200$ and $T_{\mathrm{test}}{=}166$ to support evaluation up to $H{=}50$, with the context window scaling with H $(14, 30, 50)$. Means over 3 seeds are reported in Table 11. NLBST consistently achieves the lowest RMSE and CRPS across all horizons. Notably,

Linear-DSTM—which shares the same spatial basis and Kalman updates but uses discrete-time linear dynamics—degrades catastrophically as H grows (RMSE $1.43 \rightarrow 19.17 \rightarrow 63.23$), whereas NLBST remains stable (RMSE ≤ 0.59 throughout). This highlights that continuous-time nonlinear dynamics are essential for stable extrapolation over long horizons.

Table 11: (**RQ1**, **RQ3**) Long-horizon forecasting on **D3** ($PM_{2.5}$) with $H \in \{10, 30, 50\}$. Mean over 3 seeds; RMSE and CRPS in $\log(1+y)$ space. Best per column in **bold**.

	H=10		H=30		H=50	
Model	RMSE	CRPS	RMSE	CRPS	RMSE	CRPS
NLBST (Ours)	**0.552**	**0.322**	**0.590**	**0.347**	**0.490**	**0.286**
Linear-DSTM	1.432	0.778	19.165	9.106	63.225	35.082
GRU-D	1.260	1.100	1.314	1.151	1.238	1.080
Latent ODE	1.287	0.906	1.232	0.848	1.304	0.925
FNO	1.264	0.748	1.288	0.781	1.407	0.830
GraFITi	0.977	0.779	0.863	0.684	0.888	0.708
APN	0.781	0.579	0.953	0.765	0.787	0.595