

# LEARNING SPATIOTEMPORAL DYNAMICAL SYSTEMS FROM POINT PROCESS OBSERVATIONS

Valerii Iakovlev Harri Lähdesmäki

Department of Computer Science, Aalto University, Finland

{valerii.iakovlev, harri.lahdesmaki}@aalto.fi

## ABSTRACT

Spatiotemporal dynamics models are fundamental for various domains, from heat propagation in materials to oceanic and atmospheric flows. However, currently available neural network-based spatiotemporal modeling approaches fall short when faced with data that is collected randomly over time and space, as is often the case with sensor networks in real-world applications like crowdsourced earthquake detection or pollution monitoring. In response, we developed a new method that can effectively learn spatiotemporal dynamics from such point process observations. Our model integrates techniques from neural differential equations, neural point processes, implicit neural representations and amortized variational inference to model both the dynamics of the system and the probabilistic locations and timings of observations. It outperforms existing methods on challenging spatiotemporal datasets by offering substantial improvements in predictive accuracy and computational efficiency, making it a useful tool for modeling and understanding complex dynamical systems observed under realistic, unconstrained conditions.

## 1 INTRODUCTION

In this work, consider the modeling of spatiotemporal dynamical systems whose dynamics are driven by partial differential equations. Such systems are ubiquitous and range from heat propagation in microstructures to the dynamics of oceanic currents. We use the data-driven modeling approach, where we observe a system at various time points and spatial locations and use the collected data to learn a model. In practice, the data is often collected by sensor networks that make measurements at random time points and random spatial locations. Such a measurement approach is used, for example, in crowdsourced earthquake monitoring where smartphones are used as measurement devices (Minson et al., 2015; Kong et al., 2016), in oceanographic monitoring where measurements are made by floating buoys (Albaladejo et al., 2010; Xu et al., 2014; Marin-Perianu et al., 2008), and for air pollution monitoring with vehicle-mounted sensors (Ma et al., 2008; Ghanem et al., 2004). This approach offers several advantages as it needs no sensor synchronization and allows the sensors to move freely. However, its random nature makes modeling more challenging as it requires capturing both the system dynamics and the random observation process.

We focus on the marked spatiotemporal point process setting, where the randomness in observation times and locations is defined in terms of the state of the underlying spatiotemporal dynamical system and marks are generated from the system state via an observation function. To the best of our knowledge, existing neural network-based methods cannot model such randomly observed spatiotemporal dynamical systems, mainly for two reasons. First, they do not model observation times and locations, hence they cannot predict where and when the next observation will be made.

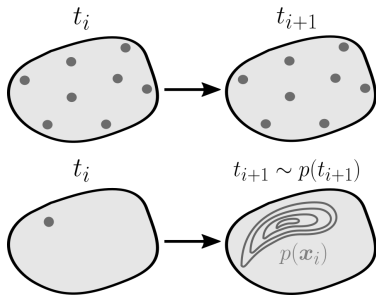


Figure 1: **Top:** Previous models assume dense observations at every time point and fixed spatiotemporal grids. **Bottom:** Our model works with extremely sparse observations and predicts where the next observations will happen.

Source code and datasets can be found in our [github repository](#).

Second, they assume the sensors form a fixed spatial grid and record data simultaneously, which is not the case in our setup where the data could come from as little as a single sensor at each time point (Fig. 1). For example, some earlier methods assume the observations are made on a fixed and regular spatiotemporal grid (Long et al., 2018; Geneva & Zabaras, 2020), other methods work with irregular, but still fixed, observation locations (Iakovlev et al., 2021; Lienen & Günnemann, 2022), and other works go further and allow the observation locations to change over time (Pfaff et al., 2021; Yin et al., 2023) but fix the observation times and assume dense observations. Whereas, another line of research has proposed methods that model only the spatiotemporal observation process without modeling the system dynamics (Chen et al., 2021; Zhu et al., 2021; Zhou et al., 2022; Zhou & Yu, 2023; Du et al., 2016).

Our work fills this gap and proposes a model for randomly observed spatiotemporal dynamical systems. Our model incorporates techniques from amortized variational inference (Kingma & Welling, 2013), neural differential equations (Chen et al., 2018; Rackauckas et al., 2020), neural point processes (Mei & Eisner, 2017; Chen et al., 2021), and implicit neural representations (Chen et al., 2023; Yin et al., 2023) to efficiently learn both the underlying system dynamics and the random observation process. Our model uses initial observations to obtain the variational estimate of the latent initial state via a transformer encoder (Vaswani et al., 2017), simulates the latent trajectory with neural ODEs (Chen et al., 2018), and uses implicit neural representations to parameterize the point process and observation distribution. Furthermore, we identify a computational bottleneck in the latent state evaluation and propose a technique to alleviate it, resulting in up to 4x faster training. Our model shows strong empirical results outperforming other models from the literature on challenging spatiotemporal datasets.

## 2 BACKGROUND

### 2.1 SPATIOTEMPORAL POINT PROCESSES

Spatiotemporal point processes (STPP) model sequences of events occurring in space and time. Each event has an associated event time  $t_i \in \mathbb{R}_{\geq 0}$  and event location  $\mathbf{x}_i \in \mathbb{R}^{d_x}$ . Given an event history  $\mathcal{H}_t \triangleq \{(t_i, \mathbf{x}_i) \mid t_i < t\}$  with all events up to time  $t$ , we can characterize an STPP by its conditional intensity function

$$\lambda^*(t, \mathbf{x}) \triangleq \lim_{\delta t \downarrow 0, \delta r \downarrow 0} \frac{\mathbb{P}(t_i \in [t, t + \delta t], \mathbf{x}_i \in B_{\delta r}(\mathbf{x}) \mid \mathcal{H}_t)}{\delta t |B_{\delta r}(\mathbf{x})|}, \quad (1)$$

where  $\delta t$  denotes an infinitesimal time interval, and  $B_{\delta r}(\mathbf{x})$  denotes a  $\delta r$ -ball centered at  $\mathbf{x}$ . Given a history  $\mathcal{H}_t$  with  $i - 1$  events,  $\lambda^*(t, \mathbf{x})$  describes the instantaneous probability of the next,  $i$ th, event occurring at time  $t$  and location  $\mathbf{x}$ . Given a sequence of  $N$  events  $\{(t_i, \mathbf{x}_i)\}_{i=1}^N$  on a bounded domain  $A \subset [0, T] \times \mathbb{R}^{d_x}$ , the log-likelihood for the STPP is evaluated as (Daley et al., 2003)

$$\log p(\{(t_i, \mathbf{x}_i)\}_{i=1}^N) = \sum_{i=1}^N \log \lambda^*(t_i, \mathbf{x}_i) - \int_A \lambda^*(t, \mathbf{x}) d\mathbf{x} dt. \quad (2)$$

Marked STPP extends the above simple STPP by a mark  $\mathbf{y}_i \in \mathbb{R}^{d_y}$  that is associated to each event  $(t_i, \mathbf{x}_i)$ .

### 2.2 ORDINARY AND PARTIAL DIFFERENTIAL EQUATIONS

Given a deterministic continuous-time dynamic system with state  $\mathbf{z}(t) \in \mathbb{R}^{d_z}$ , we can describe the evolution of its state in terms of an ordinary differential equation (ODE)

$$\frac{d\mathbf{z}(t)}{dt} = f(t, \mathbf{z}(t)). \quad (3)$$

For an initial state  $\mathbf{z}_1$  at time  $t_1$  we can solve the ODE to obtain the system state  $\mathbf{z}(t)$  at later times  $t > t_1$ . The solution exists and is unique if  $f$  is continuous in time and Lipschitz continuous in state (Coddington et al., 1956), and can be obtained either analytically or using numerical ODE solvers (Heirer et al., 1987). In this work we solve ODEs numerically using ODE solvers from

`torchdiffeq` (Chen, 2018) package. Similarly, the dynamics of spatiotemporal systems with state  $\mathbf{z}(t, \mathbf{x})$  defined over both space and time is described in terms of a partial differential equation (PDE)

$$\frac{\partial \mathbf{z}(t, \mathbf{x})}{\partial t} = F(t, \mathbf{x}, \mathbf{z}(t, \mathbf{x}), \nabla \mathbf{z}(t, \mathbf{x})) \quad (4)$$

which incorporates both temporal and spatial derivatives, indicated by  $\frac{\partial}{\partial t}$  and  $\nabla$ , respectively.

### 3 PROBLEM SETUP

In this work, we model spatiotemporal dynamical systems from data. The data consists of multiple trajectories collected by observing a system over a period of time. A trajectory consists of  $N$  triplets  $\{(t_i, \mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^N$ , where  $\mathbf{y}_i \in \mathbb{R}^{d_y}$  is a system observation, with  $t_i \in \mathbb{R}_{\geq 0}$  and  $\mathbf{x}_i \in \mathbb{R}^{d_x}$  being the corresponding observation time and location. The number of observations  $N$  can change across different trajectories in the dataset. Due to randomness of the observation process we assume the observation times and locations do not overlap (neither within a trajectory nor across trajectories), resulting in a single observation  $\mathbf{y}_i$  per time point and location. For brevity, we describe our method for a single observed trajectory, but extension to multiple trajectories is straightforward.

We assume the data generating process (DGP) consists of a latent spatiotemporal state  $\mathbf{u}(t, \mathbf{x}) \in \mathbb{R}^{d_u}$  with  $t \in \mathbb{R}_{\geq 0}$  and  $\mathbf{x} \in \mathbb{R}^{d_x}$ , whose dynamics are governed by a PDE. The latent state directly affects the times and locations of the observations, and is only partially observed:

$$\mathbf{u}(t_1, \cdot) \sim p(\mathbf{u}), \quad \frac{\partial \mathbf{u}(t, \mathbf{x})}{\partial t} = F(\mathbf{u}(t, \mathbf{x}), \nabla \mathbf{u}(t, \mathbf{x})), \quad (5)$$

$$t_i, \mathbf{x}_i \sim \text{nhpp}(\lambda(\mathbf{u}(t, \mathbf{x}))), \quad i = 1, \dots, N, \quad (6)$$

$$\mathbf{y}_i \sim p(\mathbf{y}_i | g(\mathbf{u}(t_i, \mathbf{x}_i))), \quad i = 1, \dots, N. \quad (7)$$

According to this process, each trajectory is generated as follows. We first sample the latent field  $\mathbf{u}(t_1, \cdot)$  at initial time point  $t_1$  and specify its dynamics by a PDE (Eq. 5). Next, we sample the observation times  $t_i$  and locations  $\mathbf{x}_i$  from a non-homogeneous Poisson process (nhpp) with intensity  $\lambda$  that is a function of the latent state (Eq. 6). Finally, we sample  $\mathbf{y}_i$  from the observation distribution parameterized by a function  $g$  (Eq. 7). All datasets in this work were generated according to this process (see Appendix A for details). We assume the data generating process is fully unknown, and our goal is to construct and learn its model from the data.

## 4 METHODS

In this section we describe our proposed model, its components, and the parameter inference method.

### 4.1 MODEL

Our goal is to model the true DGP (Eqs. 5-7). To this end, we define our generative model as:

$$\mathbf{z}(t_1) \sim p(\mathbf{z}(t_1)), \quad \frac{d\mathbf{z}(\tau)}{d\tau} = f(\mathbf{z}(\tau)), \quad (8)$$

$$\mathbf{u}(t, \mathbf{x}) = \phi(\mathbf{z}(t), \mathbf{x}), \quad (9)$$

$$t_i, \mathbf{x}_i \sim \text{nhpp}(\lambda(\mathbf{u}(t, \mathbf{x}))), \quad i = 1, \dots, N, \quad (10)$$

$$\mathbf{y}_i \sim p(\mathbf{y}_i | g(\mathbf{u}(t_i, \mathbf{x}_i))), \quad i = 1, \dots, N, \quad (11)$$

and the corresponding joint distribution is (details in App. B):

$$p(\mathbf{z}_1, \{t_i, \mathbf{x}_i, \mathbf{y}_i\}_{i=1}^N) = p(\mathbf{z}_1) p(\{t_i, \mathbf{x}_i\}_{i=1}^N | \mathbf{z}_1) \prod_{i=1}^N p(\mathbf{y}_i | \mathbf{z}_1, t_i, \mathbf{x}_i). \quad (12)$$

Below, we describe each component in detail, and a diagram of our model is shown in Figure 2.

**Latent dynamics (Eq. 8).** Our goal is to model the latent PDE dynamics of the DGP (Eq. 5). To do that, we introduce a low-dimensional *temporal* latent state  $\mathbf{z}(\tau) \in \mathbb{R}^{d_z}$  with  $\tau \in \mathbb{R}_{\geq 0}$  that encodes the *spatiotemporal* field  $\mathbf{u}(\tau, \cdot)$ , and introduce a neural network-parameterized ODE governing its dynamics (Eq. 8). We use a low-dimensional latent state because it allows to simulate the dynamics considerably faster than a full-grid spatiotemporal discretization (Wu et al., 2022; Yin et al., 2023). During our experiments, we observed that it takes ODE solvers a long time to solve the ODE and that this time increases with the number of observations. We found this happens because of a bottleneck in existing ODE solvers caused by extremely dense time grids (such as in Fig. 3). Such time grids force the solver to choose small steps which prevents it from adaptively selecting the optimal step size thus reducing its efficiency. To alleviate this bottleneck, we move away from the original dense time grid  $t_1, \dots, t_N$  to an auxiliary sparse grid  $\tau_1, \dots, \tau_n$ , with  $n \ll N$ , where the first and last time points coincide with the original grid:  $\tau_1 = t_1, \tau_n = t_N$  (see Fig. 2). Then, we solve for the latent state  $\mathbf{z}(t)$  only at the sparse grid, which allows the ODE solver to choose the optimal step size and results in approximately an order of magnitude (up to 9x) faster simulations (see Sec. 5). Finally, we use the computed latent states  $\mathbf{z}(\tau_1), \dots, \mathbf{z}(\tau_n)$  to approximate the original latent state  $\mathbf{z}(t)$  at intermediate time points via interpolation:  $\tilde{\mathbf{z}}(t) = \text{interpolate}(t; \mathbf{z}(\tau_1), \dots, \mathbf{z}(\tau_n))$ .

**Latent state decoding (Eq. 9).** Next, we need to recover the latent spatiotemporal state  $\mathbf{u}(t, \mathbf{x})$  from the low-dimensional representation  $\tilde{\mathbf{z}}(t)$ . We define the latent spatiotemporal state  $\mathbf{u}(t, \mathbf{x})$  as a function  $\phi(\tilde{\mathbf{z}}(t), \mathbf{x})$  of the latent state  $\tilde{\mathbf{z}}(t)$  and evaluation location  $\mathbf{x}$  (Eq. 9). In particular, we parameterize  $\phi(\tilde{\mathbf{z}}(t), \mathbf{x})$  by an MLP which takes  $\tilde{\mathbf{z}}(t) + \text{proj}(\mathbf{x})$  as the input and maps it directly to  $\mathbf{u}(t, \mathbf{x})$ , where  $\text{proj}(\mathbf{x}) \in \mathbb{R}^{d_z}$  is a trainable linear mapping of  $\mathbf{x}$  to  $\mathbb{R}^{d_z}$ . This formulation results in space-time continuous  $\mathbf{u}(t, \mathbf{x})$  which can be evaluated at any time point and spatial location, which is required to define the intensity function  $\lambda(\mathbf{u}(t, \mathbf{x}))$  and observation function  $g(\mathbf{u}(t, \mathbf{x}))$ .

**Intensity function (Eq. 10).** We parameterize the intensity function  $\lambda(\mathbf{u}(t, \mathbf{x}))$  by an MLP which takes  $\mathbf{u}(t, \mathbf{x})$  as the input and maps it to intensity of the Poisson process (Eq. 10). To ensure non-negative values of the intensity function and improve its numerical stability we further exponentiate the output of the MLP and add a small constant to it.

**Observation function (Eq. 11).** The observation function  $g(\mathbf{u}(t, \mathbf{x}))$  maps the latent spatiotemporal state  $\mathbf{u}(t, \mathbf{x})$  to parameters of the observation model (Eq. 11). Therefore, its exact structure depends on the observation model. In this work the observation model is a normal distribution, where the variance is fixed and the observation function  $g(\mathbf{u}(t, \mathbf{x}))$  returns the mean.

Details of the model specification and parameterization can be found in Appendix B and C.

## 4.2 PARAMETER AND LATENT STATE INFERENCE

To infer the model parameters and posterior of the latent initial state  $\mathbf{z}_1$  we use variational inference (Blei et al., 2017). We define an approximate posterior  $q(\mathbf{z}_1; \psi)$  with variational parameters  $\psi$  to approximate the true posterior  $p(\mathbf{z}_1 | \{t_i, \mathbf{x}_i, \mathbf{y}_i\}_{i=1}^N)$  and then minimize the Kullback-Leibler divergence

$$\text{KL} [q(\mathbf{z}_1; \psi) || p(\mathbf{z}_1 | \{t_i, \mathbf{x}_i, \mathbf{y}_i\}_{i=1}^N)] \quad (13)$$

over the model and variational parameters to obtain an estimate of the model parameters and approximation of the posterior.

To avoid optimizing the local variational parameters  $\psi$  for each trajectory in the dataset, we use amortization (Kingma & Welling, 2013) and define  $\psi$  as the output of an encoder:

$$\psi = \text{Encoder}(\{t_i, \mathbf{x}_i, \mathbf{y}_i\}_{i=1}^N), \quad (14)$$

which maps observations  $\{t_i, \mathbf{x}_i, \mathbf{y}_i\}_{i=1}^N$  to the local variational parameters  $\psi$ . This allows to optimize a fixed set of encoder parameters instead of a dataset size-dependent set of local variational parameters. We discuss the structure of the encoder in the next section.

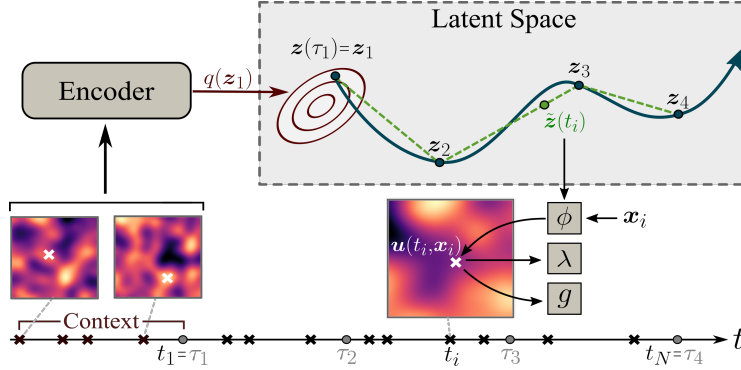


Figure 2: **Model diagram.** Black crosses on the time axis are observation times, while white crosses on field images are the corresponding observation locations. The initial observations (context) are mapped by the encoder to the latent initial state distribution  $q(\mathbf{z}_1)$ . The initial state  $\mathbf{z}_1$  is then sampled from that distribution and evolved through time using the dynamics model. The latent trajectory is evaluated only at the sparse time grid  $\tau_1, \dots, \tau_n$  and the latent state  $\tilde{\mathbf{z}}(t)$  at other time points is evaluated via interpolation. The latent state  $\tilde{\mathbf{z}}(t)$  is then mapped by  $\phi$  to the spatiotemporal state  $\mathbf{u}(t, \mathbf{x})$ , which is used to parameterize the point process and observation distribution (via mappings  $\lambda$  and  $g$ , respectively) for predicting subsequent observation times and locations.

In practice, instead of minimizing the KL divergence we, equivalently, maximize the evidence lower bound (ELBO) which for our model is defined as:

$$\mathcal{L} = \underbrace{\sum_{i=1}^N \mathbb{E}_{q(\mathbf{z}_1; \psi)} [\ln p(\mathbf{y}_i | t_i, \mathbf{x}_i, \mathbf{z}_1)]}_{(i) \text{ Expected observation log-lik.}} + \underbrace{\mathbb{E}_{q(\mathbf{z}_1; \psi)} \left[ \sum_{i=1}^N \ln \lambda(\mathbf{u}(t_i, \mathbf{x}_i)) - \int \lambda(\mathbf{u}(t, \mathbf{x})) d\mathbf{x} dt \right]}_{(ii) \text{ Expected STPP log-lik.}} \quad (15)$$

$$- \underbrace{\text{KL}[q(\mathbf{z}_1; \psi) \| p(\mathbf{z}_1)]}_{(iii) \text{ KL between prior and posterior}}. \quad (16)$$

The ELBO is maximized wrt. the model and encoder parameters. Appendix B contains detailed derivation of the ELBO, and fully specifies the model and the approximate posterior. While the term (iii) can be computed analytically, computation of terms (i) and (ii) involves approximations: Monte Carlo integration for the expectations and intensity integral, and numerical ODE solvers for the solution of the initial value problems. Appendix B details the computation of ELBO.

### 4.3 ENCODER

Our encoder maps the observations  $\{t_i, \mathbf{x}_i, \mathbf{y}_i\}_{i=1}^N$  to the local variational parameters  $\psi$  of  $q(\mathbf{z}_1; \psi)$  using Eq. 14. This process begins by embedding each observation into a high-dimensional feature space. The embedded sequence is then processed by a stack of transformer encoder layers (Vaswani et al., 2017), after which the output is mapped to the variational parameters  $\psi$ . Below, we describe this process in more detail.

First, we convert the observation sequence  $\{t_i, \mathbf{x}_i, \mathbf{y}_i\}_{i=1}^N$  into a sequence of vectors  $\{\mathbf{b}_i\}_{i=1}^N$ , where

$$\mathbf{b}_i = \text{proj}(t_i) + \text{proj}(\mathbf{x}_i) + \text{proj}(\mathbf{y}_i), \quad (17)$$

and proj are separate trainable linear projections. This additive decomposition is simple yet shows strong empirically performance in our experiments. Since  $t_i$  and  $\mathbf{x}_i$  already provide positional information, we omit positional encodings. Additionally, we introduce an aggregation token, [AGG], with a learnable representation to indicate where the encoder should aggregate information about the observations. This gives us the following encoder input sequence:  $\{\{\mathbf{b}_i\}_{i=1}^N, [\text{AGG}]\}$ . Next, a stack of transformer encoder layers processes the input sequence, producing the output sequence  $\{\{\bar{\mathbf{b}}_i\}_{i=1}^N, [\overline{\text{AGG}}]\}$ . Finally, we map the output aggregation token  $[\overline{\text{AGG}}]$  to the variational parameters  $\psi$ . This mapping depends on the form of the approximate posterior (see Appendix B for specification of the mapping).

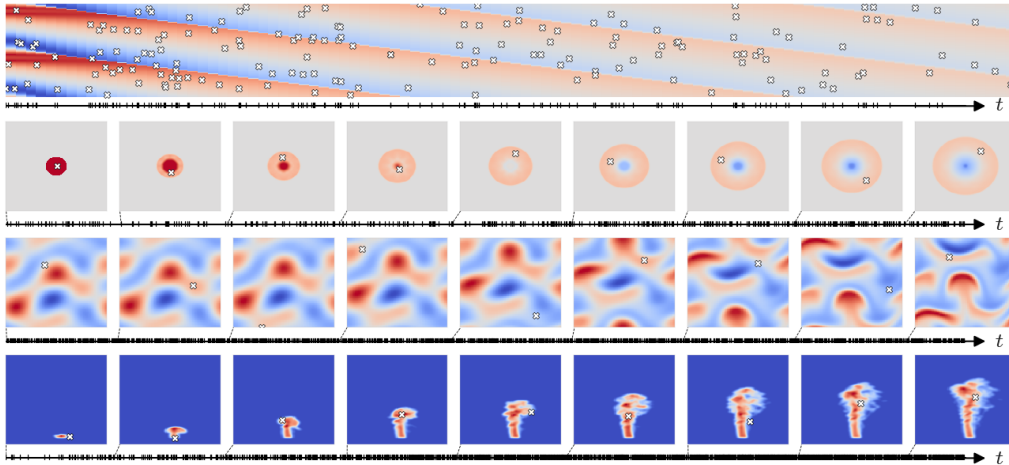


Figure 3: **Examples of trajectories from our datasets.** White crosses indicate the observation locations  $\mathbf{x}$ , marks on the horizontal time grid indicate the corresponding observation times  $t$ , and color-coded fields denote the system state with observations  $\mathbf{y}$  at white crosses  $(t, \mathbf{x})$ . **Row 1:** Burger’s (1D) dataset. The full trajectory is shown, with 1-D measurement location  $\mathbf{x}$  along the vertical and time  $t$  along the horizontal directions. **Rows 2, 3 and 4:** Shallow Water, Navier-Stokes, and Scalar Flow datasets, respectively. Due to the 2D nature of the systems, only snapshots of the system state are plotted.

#### 4.4 FORECASTING

Given a set of initial observations  $\{t_i^*, \mathbf{x}_i^*, \mathbf{y}_i^*\}_{i=1}^{N_{\text{ctx}}}$  of a test trajectory (we call it ”context”), we want to predict how the system behaves at some time point  $t > t_{N_{\text{ctx}}}$  and spatial location  $\mathbf{x}$ . Given the context, we predict the observation  $\mathbf{y}$  at  $t$  and  $\mathbf{x}$  as the expectation of the following approximate posterior predictive distribution:

$$p(\mathbf{y}|t, \mathbf{x}, \{t_i^*, \mathbf{x}_i^*, \mathbf{y}_i^*\}_{i=1}^{N_{\text{ctx}}}) = \int p(\mathbf{y}|t, \mathbf{x}, \mathbf{z}_1) q_{\psi^*}(\mathbf{z}_1) d\mathbf{z}_1, \quad (18)$$

where  $\psi^* = \text{Encoder}(\{t_i^*, \mathbf{x}_i^*, \mathbf{y}_i^*\}_{i=1}^{N_{\text{ctx}}})$ , and the expectation is approximated via Monte Carlo integration. In Eq. 18, we omit explicitly conditioning on training data for brevity. Similarly, the distribution over the observation times and locations is evaluated as:

$$p(t, \mathbf{x}|\{t_i^*, \mathbf{x}_i^*, \mathbf{y}_i^*\}_{i=1}^{N_{\text{ctx}}}) = \int p(t, \mathbf{x}|\mathbf{z}_1) q_{\psi^*}(\mathbf{z}_1) d\mathbf{z}_1. \quad (19)$$

In our experiments, the context for each trajectory is defined as all observations withing a fixed initial time period (see Appendix A for details).

## 5 EXPERIMENTS

In this section we demonstrate properties of our method and compare it against other methods from the literature. Our datasets are generated by three commonly-used PDE systems: Burgers’ (models nonlinear 1D wave propagation), Shallow Water (models 2D wave propagation under the gravity), and Navier-Stokes with transport (models the spread of a pollutant in a liquid over a 2D domain). In addition to the synthetic data, we include a real-world dataset Scalar Flow (Eckert et al., 2019), which contains observations of smoke plumes raising in warm air. See Appendix A for details about the dataset generation. Figure 3 shows examples of trajectories from the datasets.

In all cases our model has at most 3 million parameters, and training takes at most 1.5 hours on a single GeForce RTX 3080 GPU. The training is done for 25k iterations with learning rate  $3e-4$  and batch size 32. We use the adaptive ODE solver (dopri5) from `torchdiffeq` package with relative

and absolute tolerance set to  $1e-5$ . As performance metrics we use mean absolute error (MAE) for the observations  $\mathbf{y}_i$ , and event-averaged log-likelihood of the point process for  $t_i$  and  $\mathbf{x}_i$ , both evaluated on the test set. See Appendix C for details about our training setup.

**Context Size.** Our encoder maps the initial observations (context)  $\{t_i^*, \mathbf{x}_i^*, \mathbf{y}_i^*\}_{i=1}^{N_{\text{ctx}}}$  to parameters  $\psi$  of the approximate posterior  $q(\mathbf{z}_1; \psi)$ . Here we look at how accuracy of the state predictions and process likelihood is affected by the context size. We train and test our model with different context sizes: full (using all  $N_{\text{ctx}}$  points), half of the context (using points from  $N_{\text{ctx}}/2$  to  $N_{\text{ctx}}$ ), etc., and show the results in Figure 4. We see that both MAE and process likelihood improve as we increase the context size, but the improvements tend to saturate for larger contexts. This indicates that the encoder mostly uses observations that are close to the time point at which the latent state is inferred, and does not utilize observations that are too far away from it. This effect is especially visible with synthetic data, but the plots suggest the real-world dataset Scalar Flow could still benefit from larger context.

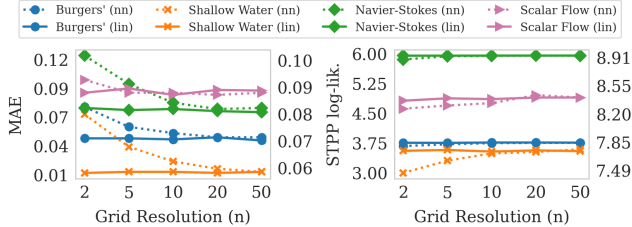


Figure 5: Test MAE ( $\downarrow$ ) and process log-likelihood ( $\uparrow$ ) for different temporal grid resolutions ( $n$ ) and interpolation methods (nn = nearest neighbor, lin = linear). For Scalar Flow we use the right axis for better visibility.

**Latent State Interpolation.** As discussed in Section 4.1, we do not evaluate the latent state at the full time grid  $t_1, \dots, t_N$  using the ODE solver (we call it the sequential method). Instead, we use an adaptive ODE solver to evaluate the latent state only at a sparser time grid  $\tau_1, \dots, \tau_n$ , and then interpolate the evaluations to the full grid. Here we look at how this approach affects training times, and investigate the effects of the grid resolution  $n$  and different interpolation methods. We vary the resolution from coarse  $n = 2$  to fine  $n = 50$ , and test nearest neighbor and linear interpolation methods. Figure 5 shows that both MAE and process log-likelihood improve with the grid resolution, but only when we use nearest neighbor interpolation. With linear interpolation, the model achieves its best performance with just  $n = 2$ , meaning that interpolation is done between only two points in the latent space, and further increasing the resolution does not improve the results. To study this observation in more detail, we provide additional experiments in Appendix D. Both interpolation methods achieve the same optimal performance when  $n \geq 20$ . This effect is visible for both synthetic and, to a smaller extent, real data. Furthermore, Table 1 shows that our method results in up to 9x faster latent state evaluation than the sequential method, which in our case translates in up to 4x faster training.

**Latent Space Dimension.** Our model assumes the system dynamics can be accurately modeled in a low-dimensional latent space. In this experiment we show that this assumption is valid at least for the four systems that we consider and best predictive accuracy is achieved for relatively low-dimensional latent space dimension  $d_z$ . In Figure 6 we show how MAE and process log-likelihood improve as we increase the latent state dimension. For most datasets the predictive accuracy converges for as low as 16-dimensional latent space, except for the Navier-Stokes dataset, where  $d_z = 64$  is required, likely

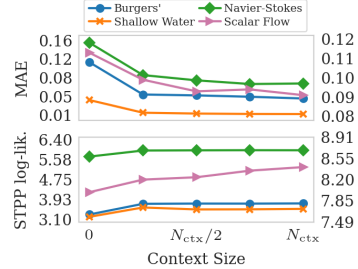


Figure 4: Test MAE ( $\downarrow$ ) and log-likelihood ( $\uparrow$ ) vs. context size. For Scalar Flow we use the right axis for better visibility.

Dataset	Interp.	Seq.
Burgers'	0.007	0.018
Shallow Water	0.011	0.045
Navier-Stokes	0.012	0.108
Scalar Flow	0.011	0.089

Table 1: Latent state evaluation time (sec), interpolation vs. sequential method. Averaged time over all trajectories.

due to its more intricate state (as can be qualitatively observed in Fig. 3). We note that  $d_z$  affects the number of model parameters, but the difference between 1- and 64-dimensional latent space is less than 3%.

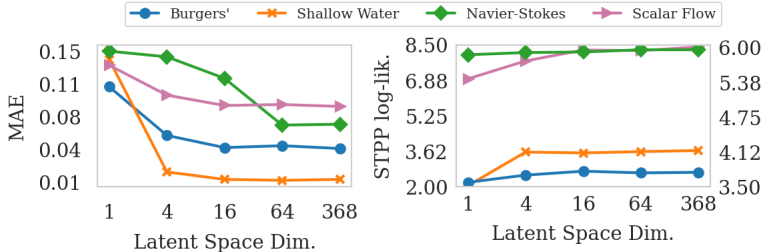


Figure 6: Test MAE ( $\downarrow$ ) and process log-likelihood ( $\uparrow$ ) for different latent space dimensions ( $d_z$ ). For Burgers' and Navier-Stokes we use the right axis for better visibility.

### Interaction Between Observation and Process Models.

In our model the observation times and locations are modeled by a point process model (Eq. 10), while observations are modeled by an observation model (Eq. 11). Intuitively, one could expect some interaction between them and ask the question of how important is one model for the performance of the other. In this experiment we look at the effect of removing the point process model on the performance of the observation model and vice versa. We remove the point process model by removing Eq. 10 from our model, so that we do not model the observation times and locations. Similarly, we remove the observation model by removing Eq. 11 from our model and also removing observations  $y_i$  from our data, so that there is no information about the observation values. In Table 2 we show that removing the point process model seems to have practically no effect on MAE, while removing the observation model decreases the process likelihood. This shows that if we know the system states we can model the observation times and locations more accurately.

Dataset	MAE	MAE (-STPP)	Log-lik.	Log-lik (-Obs.)
Burgers'	0.042	0.043	3.77	3.70
Shallow Water	0.012	0.012	3.67	3.56
Navier-Stokes	0.071	0.071	5.96	5.92
Scalar Flow	0.090	0.089	8.43	6.87

Table 2: Test MAE ( $\downarrow$ ) and log-likelihood ( $\uparrow$ ) of the full model, and after removing the point process (-STPP) or the observation model (-Obs.).

**Comparison to Other Methods.** We compare our method against two groups of methods from the literature. The first group consists of neural spatiotemporal dynamic models that can model only the observations  $y_i$  but not the time points and spatial locations. This group includes FEN (Lienen & Günnemann, 2022) (graph neural network-based model, simulates the system dynamics at every point on the spatial grid), NSPDE (Salvi et al., 2022) (space- and time-continuous model simulating the dynamics in the spectral domain), and CNN-ODE (our simple baseline that uses a CNN encoder/decoder to map observations to/from the latent space, and uses neural ODEs (Chen et al., 2018) to model the latent space dynamics). Note that due to their assumption about multiple observations per time point, dynamic models require time binning (see Appendix C.4). The second group consists of neural spatiotemporal point processes that model only the observation times and locations. This group consists of DSTPP (Yuan et al., 2023), NSTPP (Chen et al., 2021), and AutoSTPP (Zhou & Yu, 2023). We also provide simple baselines for both groups: median predictor for dynamic models, and constant (learnable) intensity for the point process group. Hyperparameters of each method were tuned for the best performance. See Appendix C.4 for details about the models and hyperparameters used.

Table 3 shows the comparison results. We see that most methods from the first group perform rather poorly and fail to beat even the MAE of the median predictor, with the CNN-ODE model showing the best results. However, CNN-ODE still performs considerably worse than our model on synthetic data, likely because of the loss of information caused by time binning. On the real-world Scalar Flow dataset CNN-ODE and our model have similar performance. In the second group most methods



Model	Burgers'		Shallow Water		Navier-Stokes		Scalar Flow	
	MAE ( $\downarrow$ )	Log-lik. ( $\uparrow$ )	MAE ( $\downarrow$ )	Log-lik. ( $\uparrow$ )	MAE ( $\downarrow$ )	Log-lik. ( $\uparrow$ )	MAE ( $\downarrow$ )	Log-lik. ( $\uparrow$ )
Median Pred.	0.161	-	0.148	-	0.155	-	0.140	-
FEN	N/A	-	$0.336 \pm 0.008$	-	$0.158 \pm 0.006$	-	$0.197 \pm 0.003$	-
NSPDE	$0.166 \pm 0.002$	-	$0.339 \pm 0.002$	-	$0.098 \pm 0.001$	-	$0.142 \pm 0.003$	-
CNN-ODE	$0.076 \pm 0.000$	-	$0.032 \pm 0.001$	-	$0.077 \pm 0.002$	-	<b><math>0.091 \pm 0.001</math></b>	-
Const. Intensity	-	3.59	-	2.02	-	5.86	-	6.87
DSTPP	-	$0.67 \pm 0.13$	-	$0.08 \pm 0.04$	-	$0.41 \pm 0.04$	-	$3.02 \pm 0.01$
NSTPP	-	$2.35 \pm 0.06$	-	$1.53 \pm 0.13$	-	$3.83 \pm 0.14$	-	$6.69 \pm 0.03$
AutoSTPP	-	N/A	-	$3.37 \pm 0.01$	-	$5.91 \pm 0.01$	-	<b><math>8.49 \pm 0.05</math></b>
Ours	<b><math>0.043 \pm 0.002</math></b>	<b><math>3.77 \pm 0.02</math></b>	<b><math>0.012 \pm 0.001</math></b>	<b><math>3.62 \pm 0.05</math></b>	<b><math>0.071 \pm 0.001</math></b>	<b><math>5.96 \pm 0.01</math></b>	<b><math>0.092 \pm 0.002</math></b>	$8.41 \pm 0.04$

Table 3: Model comparisons. MAE ( $\downarrow$ ) and Log-lik (per event) ( $\uparrow$ ) on test data. The first group of methods (FEN, NSPDE, CNN-ODE) contains neural spatiotemporal dynamical models and are evaluated using MAE. The second group (DSTPP, NSTPP, AutoSTPP) contains neural spatiotemporal point process models and are evaluated using Log-lik. Error bars represent one standard error calculated over 5 realizations of a random seed controlling the model initialization.

showed poor results and failed to beat the simple constant intensity baseline, however AutoSTPP shows very strong performance, although still not outperforming our method on the synthetic data, and being close to it on the Scalar Flow dataset.

These experiments show the challenging nature of modeling randomly observed dynamical systems, and demonstrate our method’s properties and strong performance.

## 6 RELATED WORK

**Neural Point Processes.** Traditionally, intensity functions are simple parametric functions incorporating domain knowledge about the system being modeled. While simple and interpretable, this approach might require strong domain knowledge and might have limited expressivity due to overly simplistic form of the intensity function. These limitations lead to the development of neural point processes which parameterize the intensity function by a neural network, leading to improved flexibility and expressivity. For example [Mei & Eisner \(2017\)](#); [Omi et al. \(2019\)](#); [Jia & Benson \(2019\)](#); [Zuo et al. \(2020\)](#) use neural networks to model time-dependent intensity functions in a flexible and efficient manner, with neural architectures ranging from recurrent neural networks to transformers and neural ODEs. These techniques consistently outperform classical intensity parameterizations and do not require strong domain expertise to choose the right form of the intensity function. Other works extend this idea to marked point processes where intensity is a function of time and also of a discrete or continuous mark. For example, [Chen et al. \(2021\)](#); [Zhu et al. \(2021\)](#); [Zhou et al. \(2022\)](#); [Zhou & Yu \(2023\)](#); [Yuan et al. \(2023\)](#) use spatial coordinates as a mark and model the intensity using a wide range of methods aimed at improving flexibility and efficiency. Works such as [Du et al. \(2016\)](#); [Xiao et al. \(2019\)](#); [Boyd et al. \(2020\)](#) use discrete observations as marks, and [Sharma et al. \(2018\)](#) use continuous marks with underlying dynamic model, which makes their work most similar to ours, with major differences related to the model, training process, and them working only with temporal processes, whereas we are dealing with more general spatiotemporal processes.

**Neural Spatiotemporal Dynamic Models.** To model the spatiotemporal dynamics our method uses the “encode-process-decode” approach employed in many other works. The main idea of the approach consists of taking a set of initial observations and using it to estimate the latent initial state. A dynamics function is then used to map the latent state forward in time to other time points either discretely ([Long et al., 2018](#); [HAN et al., 2022](#); [Wu et al., 2022](#)) or continuously ([Yildiz et al., 2019](#); [Salvi et al., 2022](#)). Finally, the latent state is mapped to the observations using discrete ([Yildiz et al., 2019](#); [Wu et al., 2022](#)) or continuous ([Yin et al., 2023](#); [Chen et al., 2023](#)) decoder. Such methods assume multiple observations per time point and do not model the observation times and locations. However, in contrary to many previous works that use the encode-process-decode approach to learn a discriminative model, we use the auto-encoding variational Bayes to learn a generative model of the underlying dynamical system.

**Implicit Neural Representations.** Our model represents the continuous latent spatiotemporal state  $\mathbf{u}(t, \mathbf{x})$  in terms of a function  $\phi(\tilde{\mathbf{z}}(t), \mathbf{x})$ . This approach to representing continuous field is called

implicit neural representations (Park et al., 2019; Chen & Zhang, 2019; Mescheder et al., 2019; Chibane et al., 2020), and it has found applications in many other works related to modeling of spatiotemporal systems (Chen et al., 2023; Yin et al., 2023).

## 7 CONCLUSION

In this work, we developed a method for modeling spatiotemporal dynamical systems from marked point process observations. Along with our method, we proposed an interpolation-based technique to greatly speed up the training time. In the experiments we showed that our method effectively utilizes the context of various lengths to improve accuracy of the initial state inference, uses the system state observations to better model the observation times and locations, and that low-resolution linear interpolation is sufficient to accurately represent the latent trajectories. We further demonstrated that our method achieves strong performance on challenging spatiotemporal datasets and outperforms other methods from the literature.

**Limitations.** Using a Poisson process for sampling event times and locations can be limiting in certain applications. First, the assumption of non-overlapping time points may not hold in scenarios where effectively simultaneous observations occur, such as in high-frequency sensor networks. Second, the lack of interaction between the event occurrences and system dynamics may not hold if the act of observation influences the system (e.g., triggering human actions or affecting sensor availability). Finally, during data collection faithfully capturing the event intensity requires sufficiently dense sensor placement, especially in high-intensity regions. Future work could relax these assumptions to better model such scenarios, for example, by using different point process models and adapting the architecture to efficiently account for potential interactions between the observation events and system dynamics.

## REFERENCES

- Cristina Albaladejo, Pedro Sánchez, Andrés Iborra, Fulgencio Soto, Juan A López, and Roque Torres. Wireless sensor networks for oceanographic monitoring: A systematic review. *Sensors*, 10(7): 6948–6968, 2010.
- David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017.
- Alex Boyd, Robert Bamler, Stephan Mandt, and Padhraic Smyth. User-dependent neural sequence models for continuous-time event data. *Advances in Neural Information Processing Systems*, 33: 21488–21499, 2020.
- Peter Yichen Chen, Jinxu Xiang, Dong Heon Cho, Yue Chang, G A Pershing, Henrique Teles Maia, Maurizio M Chiaramonte, Kevin Thomas Carlberg, and Eitan Grinspun. CROM: Continuous reduced-order modeling of PDEs using implicit neural representations. In *The Eleventh International Conference on Learning Representations*, 2023. URL <https://openreview.net/forum?id=FUORz1tG8Og>.
- Ricky T. Q. Chen. torchdiffeq, 2018. URL <https://github.com/rtqichen/torchdiffeq>.
- Ricky T. Q. Chen, Brandon Amos, and Maximilian Nickel. Neural spatio-temporal point processes. In *International Conference on Learning Representations*, 2021.
- Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. *Advances in neural information processing systems*, 31, 2018.
- Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5939–5948, 2019.
- Julian Chibane, Thiemo Alldieck, and Gerard Pons-Moll. Implicit functions in feature space for 3d shape reconstruction and completion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6970–6981, 2020.

- Earl A Coddington, Norman Levinson, and T Teichmann. Theory of ordinary differential equations, 1956.
- Daryl J Daley, David Vere-Jones, et al. *An introduction to the theory of point processes: volume I: elementary theory and methods*. Springer, 2003.
- Nan Du, Hanjun Dai, Rakshit Trivedi, Utkarsh Upadhyay, Manuel Gomez-Rodriguez, and Le Song. Recurrent marked temporal point processes: Embedding event history to vector. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1555–1564, 2016.
- Marie-Lenat Eckert, Kiwon Um, and Nils Thuerey. Scalarflow: A large-scale volumetric data set of real-world scalar transport flows for computer animation and machine learning. *ACM Transactions on Graphics*, 38(6):239, 2019.
- Nicholas Geneva and Nicholas Zabaras. Modeling the dynamics of pde systems with physics-constrained deep auto-regressive networks. *Journal of Computational Physics*, 403:109056, 2020.
- Moustafa Ghanem, Yike Guo, John Hassard, Michelle Osmond, and Mark Richards. Sensor grids for air pollution monitoring. In *Proc. 3rd UK e-Science All Hands Meeting, Nottingham, UK*, 2004.
- Pablo Gómez, Gabriele Meoni, and Håvard Hem Toftevaag. torchquad, 2024. URL <https://github.com/esa/torchquad>.
- XU HAN, Han Gao, Tobias Pfaff, Jian-Xun Wang, and Liping Liu. Predicting physics in mesh-reduced space with temporal attention. In *International Conference on Learning Representations*, 2022. URL <https://openreview.net/forum?id=XctLdNfCmP>.
- E Heirer, SP Nørsett, and G Wanner. Solving ordinary differential equations i: Nonstiff problems, 1987.
- Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus), 2016.
- Philipp Holl, Vladlen Koltun, Kiwon Um, and Nils Thuerey. phiflow: A differentiable pde solving framework for deep learning via physical simulations. In *NeurIPS Workshop on Differentiable vision, graphics, and physics applied to machine learning*, 2020. URL <https://montrealrobotics.ca/diffcvgp/assets/papers/3.pdf>.
- Valerii Iakovlev, Markus Heinonen, and Harri Lähdesmäki. Learning continuous-time {pde}s from sparse data with graph neural networks. In *International Conference on Learning Representations*, 2021. URL <https://openreview.net/forum?id=aUX5PlaQ7Oy>.
- Junteng Jia and Austin R Benson. Neural jump stochastic differential equations. *Advances in Neural Information Processing Systems*, 32, 2019.
- Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.
- Qingkai Kong, Richard M Allen, Louis Schreier, and Young-Woo Kwon. Myshake: A smartphone seismic network for earthquake early warning and beyond. *Science advances*, 2(2):e1501055, 2016.
- PA W Lewis and Gerald S Shedler. Simulation of nonhomogeneous poisson processes by thinning. *Naval research logistics quarterly*, 26(3):403–413, 1979.
- Marten Lienen and Stephan Günnemann. Learning the dynamics of physical systems from sparse observations with finite element networks. In *International Conference on Learning Representations*, 2022. URL <https://openreview.net/forum?id=HFmAukZ-k-2>.
- Zichao Long, Yiping Lu, Xianzhong Ma, and Bin Dong. Pde-net: Learning pdes from data. In *International conference on machine learning*, pp. 3208–3216. PMLR, 2018.
- Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL <https://openreview.net/forum?id=Bkg6RiCqY7>.

- Yajie Ma, Mark Richards, Moustafa Ghanem, Yike Guo, and John Hassard. Air pollution monitoring and mining based on sensor grid in london. *Sensors*, 8(6):3601–3623, 2008.
- Mihai Marin-Perianu, Supriyo Chatterjea, Raluca Marin-Perianu, Stephan Bosch, Stefan Dulman, Stuart Kininmonth, and Paul Havinga. Wave monitoring with wireless sensor networks. In *2008 International Conference on Intelligent Sensors, Sensor Networks and Information Processing*, pp. 611–616. IEEE, 2008.
- Hongyuan Mei and Jason M Eisner. The neural hawkes process: A neurally self-modulating multivariate point process. *Advances in neural information processing systems*, 30, 2017.
- Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4460–4470, 2019.
- Sarah E Minson, Benjamin A Brooks, Craig L Glennie, Jessica R Murray, John O Langbein, Susan E Owen, Thomas H Heaton, Robert A Iannucci, and Darren L Hauser. Crowdsourced earthquake early warning. *Science advances*, 1(3):e1500036, 2015.
- Takahiro Omi, Kazuyuki Aihara, et al. Fully neural network based model for general temporal point processes. *Advances in neural information processing systems*, 32, 2019.
- Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. DeepSDF: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 165–174, 2019.
- Tobias Pfaff, Meire Fortunato, Alvaro Sanchez-Gonzalez, and Peter Battaglia. Learning mesh-based simulation with graph networks. In *International Conference on Learning Representations*, 2021. URL [https://openreview.net/forum?id=roNqYL0\\_XP](https://openreview.net/forum?id=roNqYL0_XP).
- Christopher Rackauckas, Yingbo Ma, Julius Martensen, Collin Warner, Kirill Zubov, Rohit Supekar, Dominic Skinner, Ali Ramadhan, and Alan Edelman. Universal differential equations for scientific machine learning. August 2020. doi: 10.21203/rs.3.rs-55125/v1. URL <http://dx.doi.org/10.21203/RS.3.RS-55125/V1>.
- Cristopher Salvi, Maud Lemercier, and Andris Gerasimovics. Neural stochastic pdes: Resolution-invariant learning of continuous spatiotemporal dynamics. *Advances in Neural Information Processing Systems*, 35:1333–1344, 2022.
- Anuj Sharma, Robert Johnson, Florian Engert, and Scott Linderman. Point process latent variable models of larval zebrafish behavior. *Advances in Neural Information Processing Systems*, 31, 2018.
- Makoto Takamoto, Timothy Praditia, Raphael Leiteritz, Dan MacKinlay, Francesco Alesiani, Dirk Pflüger, and Mathias Niepert. PDEBENCH: AN EXTENSIVE BENCHMARK FOR SCIENTIFIC MACHINE LEARNING. In *ICLR 2023 Workshop on Physics for Machine Learning*, 2023. URL <https://openreview.net/forum?id=b8SwOxZQ2kj>.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- Tailin Wu, Takashi Maruyama, and Jure Leskovec. Learning to accelerate partial differential equations via latent global evolution. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL <https://openreview.net/forum?id=xvZtgp5wyYT>.
- Shuai Xiao, Junchi Yan, Mehrdad Farajtabar, Le Song, Xiaokang Yang, and Hongyuan Zha. Learning time series associated event sequences with recurrent point process networks. *IEEE transactions on neural networks and learning systems*, 30(10):3124–3136, 2019.
- Guobao Xu, Weiming Shen, and Xianbin Wang. Applications of wireless sensor networks in marine environment monitoring: A survey. *Sensors*, 14(9):16932–16954, 2014.

- Cagatay Yildiz, Markus Heinonen, and Harri Lahdesmaki. Ode2vae: Deep generative second order odes with bayesian neural networks. *Advances in Neural Information Processing Systems*, 32, 2019.
- Yuan Yin, Matthieu Kirchmeyer, Jean-Yves Franceschi, Alain Rakotomamonjy, and patrick gallinari. Continuous PDE dynamics forecasting with implicit neural representations. In *The Eleventh International Conference on Learning Representations*, 2023. URL <https://openreview.net/forum?id=B73niNjbPs>.
- Yuan Yuan, Jingtao Ding, Chenyang Shao, Depeng Jin, and Yong Li. Spatio-temporal diffusion point processes. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '23. ACM, August 2023. doi: 10.1145/3580305.3599511. URL <http://dx.doi.org/10.1145/3580305.3599511>.
- Zihao Zhou and Rose Yu. Automatic integration for spatiotemporal neural point processes. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL <https://openreview.net/forum?id=DeblyPlzMN>.
- Zihao Zhou, Xingyi Yang, Ryan Rossi, Handong Zhao, and Rose Yu. Neural point process for learning spatiotemporal event dynamics. In *Learning for Dynamics and Control Conference*, pp. 777–789. PMLR, 2022.
- Shixiang Zhu, Shuang Li, Zhigang Peng, and Yao Xie. Imitation learning of neural spatio-temporal point processes. *IEEE Transactions on Knowledge and Data Engineering*, 34(11):5391–5402, 2021.
- Simiao Zuo, Haoming Jiang, Zichong Li, Tuo Zhao, and Hongyuan Zha. Transformer hawkes process. In *International conference on machine learning*, pp. 11692–11702. PMLR, 2020.

## A DATA

As described in Section 3, we assume the following data generating process

$$\mathbf{u}(t_1, \cdot) \sim p(\mathbf{u}), \quad \frac{\partial \mathbf{u}(t, \mathbf{x})}{\partial t} = F(\mathbf{u}(t, \mathbf{x}), \nabla \mathbf{u}(t, \mathbf{x})), \quad (20)$$

$$t_i, \mathbf{x}_i \sim \text{nhpp}(\lambda(\mathbf{u}(t, \mathbf{x}))), \quad i = 1, \dots, N, \quad (21)$$

$$\mathbf{y}_i \sim p(\mathbf{y}_i | g(\mathbf{u}(t_i, \mathbf{x}_i))), \quad i = 1, \dots, N, \quad (22)$$

We selected three commonly used PDE systems: Burgers', Shallow Water, and Navier-Stokes with transport. Next we discuss the data generating process for each dataset.

### A.1 BURGERS'

The Burgers' system in 1D is characterized by a 1-dimensional state  $u(t, x)$  and the following PDE dynamics:

$$\frac{\partial u}{\partial t} + u \frac{\partial u}{\partial x} = \nu \frac{\partial^2 u}{\partial x^2}, \quad (23)$$

where  $\nu$  is the diffusion coefficient. We obtain data for this system from the PDEBench dataset (Takamoto et al., 2023). The system is simulated on a time interval  $[0, 2]$  and on a spatial domain  $[0, 1]$  with periodic boundary conditions. The dataset contains 10000 trajectories, each trajectory is evaluated on a uniform temporal grid with 101 time points, and uniform spatial grid with 256 nodes.

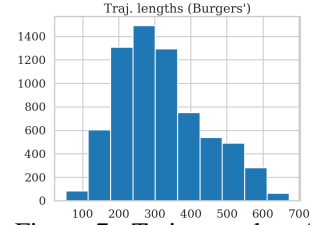


Figure 7: Trajectory lengths for Burgers' dataset.

Since the data is discrete, we define continuous field  $u(t, x)$  by linearly interpolating the data across space and time. Next, we use the interpolant to define the intensity function

$$\lambda(u(t, x)) = 1600|u(t, x)|, \quad (24)$$

and use it to simulate the non-homogeneous Poisson process using the thinning algorithm (Lewis & Shedler, 1979). As the result, for each the 10000 trajectories, we obtain a sequence of time points and spatial locations  $\{t_i, x_i\}_{i=1}^N$ , where  $N$  can be different for each trajectory. Finally, we compute the observations as

$$\mathbf{y}_i = u(t_i, x_i). \quad (25)$$

We use the first 0.5 seconds as the context for the initial state inference, and further filter the dataset to ensure that each trajectory has at least 10 points in the context, resulting in approximately 7000 trajectories. We use 80%/10%/10% split for training/validation/testing.

### A.2 SHALLOW WATER

The Shallow Water system in 2D is characterized by a 3-dimensional state

$$\mathbf{u}(t, \mathbf{x}) = [h(t, \mathbf{x}), u(t, \mathbf{x}), v(t, \mathbf{x})] \quad (26)$$

and the following PDE dynamics:

$$\frac{\partial h}{\partial t} + u \frac{\partial h}{\partial x} + v \frac{\partial h}{\partial y} = 0, \quad (27)$$

$$\frac{\partial u}{\partial t} + u \frac{\partial u}{\partial x} + v \frac{\partial u}{\partial y} + g \frac{\partial h}{\partial x} = 0, \quad (28)$$

$$\frac{\partial v}{\partial t} + u \frac{\partial v}{\partial x} + v \frac{\partial v}{\partial y} + g \frac{\partial h}{\partial y} = 0, \quad (29)$$

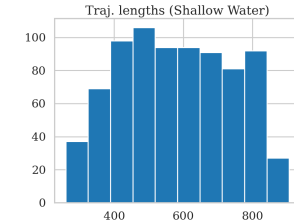


Figure 8: Trajectory lengths for Shallow Water dataset.

where  $g$  is the gravitational acceleration,  $h(t, \mathbf{x})$  is wave height, and  $u(t, \mathbf{x})$  and  $v(t, \mathbf{x})$  are horizontal and vertical velocities, respectively. We obtain data for this system from the PDEBench dataset

(Takamoto et al., 2023). The system is simulated on a time interval  $[0, 1]$  and on a spatial domain  $[-2.5, 2.5]^2$ . The dataset contains 1000 trajectories, each trajectory is evaluated on a uniform temporal grid with 101 time points, and uniform 128-by-128 spatial grid.

Since the data is discrete, we define continuous field  $\mathbf{u}(t, \mathbf{x})$  by linearly interpolating the data across space and time. Next, we use the interpolant to define the intensity function

$$\lambda(\mathbf{u}(t, \mathbf{x})) = 500|h(t, \mathbf{x}) - 1|, \quad (30)$$

indicating that measurements are made only when the wave height is below or above the baseline of 1. We use the intensity function to simulate the non-homogeneous Poisson process using the thinning algorithm (Lewis & Shedler, 1979). As the result, for each of the 1000 trajectories, we obtain a sequence of time points and spatial locations  $\{t_i, \mathbf{x}_i\}_{i=1}^N$ , where  $N$  can be different for each trajectory. Finally, we compute the observations as

$$\mathbf{y}_i = h(t_i, \mathbf{x}_i), \quad (31)$$

observing only the wave height.

We use the first 0.1 seconds as the context for the initial state inference, and further filter the dataset to ensure that each trajectory has at least 10 points in the context, resulting in approximately 800 trajectories. We use 80%/10%/10% split for training/validation/testing.

### A.3 NAVIER-STOKES

The Navier-Stokes system with a transport equation is characterized by a 4-dimensional state

$$\mathbf{u}(t, \mathbf{x}) = [c(t, \mathbf{x}), u(t, \mathbf{x}), v(t, \mathbf{x}), p(t, \mathbf{x})] \quad (32)$$

and the following PDE dynamics:

$$\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} = 0, \quad (33)$$

$$\frac{\partial u}{\partial t} + u \frac{\partial u}{\partial x} + v \frac{\partial u}{\partial y} = -\frac{\partial p}{\partial x} + \nu \left( \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} \right), \quad (34)$$

$$\frac{\partial v}{\partial t} + u \frac{\partial v}{\partial x} + v \frac{\partial v}{\partial y} = -\frac{\partial p}{\partial y} + \nu \left( \frac{\partial^2 v}{\partial x^2} + \frac{\partial^2 v}{\partial y^2} \right) + c, \quad (35)$$

$$\frac{\partial c}{\partial t} = -u \frac{\partial c}{\partial x} - v \frac{\partial c}{\partial y} + \nu \left( \frac{\partial^2 c}{\partial x^2} + \frac{\partial^2 c}{\partial y^2} \right), \quad (36)$$

where the diffusion constant  $\nu = 0.002$ , and  $c(t, \mathbf{x})$  is concentration of the transported species,  $p(t, \mathbf{x})$  is pressure, and  $u(t, \mathbf{x})$  and  $v(t, \mathbf{x})$  are horizontal and vertical velocities, respectively. For each trajectory, we start with zero initial velocities and pressure, and the initial scalar field  $c_0(x, y)$  is generated as:

$$\tilde{c}_0(x, y) = \sum_{k, l=-N}^N \lambda_{kl} \cos(2\pi(kx + ly)) + \gamma_{kl} \sin(2\pi(kx + ly)), \quad (37)$$

$$c_0(x, y) = \frac{\tilde{c}_0(x, y) - \min(\tilde{c}_0)}{\max(\tilde{c}_0) - \min(\tilde{c}_0)}, \quad (38)$$

where  $N = 2$  and  $\lambda_{kl}, \gamma_{kl} \sim \mathcal{N}(0, 1)$ .

We use PhiFlow (Holl et al., 2020) to solve the PDEs. The system is simulated on a time interval  $[0, 1]$  and on a spatial domain  $[0, 1]^2$  with periodic boundary conditions. The solution is evaluated at randomly selected spatial locations and time points. We use 1089 spatial locations and 25 time points. The spatial and temporal grids are the same for all trajectories. The dataset contains 1000 trajectories.

Since the data is discrete, we define continuous field  $\mathbf{u}(t, \mathbf{x})$  by linearly interpolating the data across space and time. Next, we use the interpolant to define the intensity function

$$\lambda(\mathbf{u}(t, \mathbf{x})) = 2000|c(t, \mathbf{x})|, \quad (39)$$

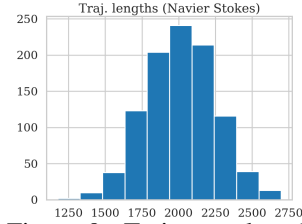


Figure 9: Trajectory lengths for Navier-Stokes dataset.

indicating that measurement intensity is proportional to the species concentration. We use the intensity function to simulate the non-homogeneous Poisson process using the thinning algorithm (Lewis & Shedler, 1979). As the result, for each of the trajectories, we obtain a sequence of time points and spatial locations  $\{t_i, \mathbf{x}_i\}_{i=1}^N$ , where  $N$  can be different for each trajectory. Finally, we compute the observations as

$$\mathbf{y}_i = c(t_i, \mathbf{x}_i), \quad (40)$$

observing only the species concentrations.

We use the first 0.5 seconds as the context for the initial state inference, and further filter the dataset to ensure that each trajectory has at least 10 points in the context, resulting in all 1000 trajectories satisfying this condition. We use 80%/10%/10% split for training/validation/testing.

#### A.4 SCALAR FLOW

To demonstrate capabilities of our model on a real-world process, we use the Scalar Flow Dataset (Eckert et al., 2019). This dataset consists of videos of smoke plumes rising in hot air. Each video has 150 frames recorded over 2.5 seconds, and each frame has the resolution of 1080x1920 pixels. The observations are postprocessed camera images of the smoke plumes taken from multiple views. For simplicity, we use only the front view. We downsample the original frames from 1080x1920 to 120x213 pixels and smooth out the high-frequency noise by applying Gaussian smoothing with the standard deviation of 1.5.

Then, we emulate a real-world measurement process where each "sensor" (pixel) makes a measurement with the probability directly proportional to the observed smoke density (higher density implies more frequent measurements). In particular we set the observation probability to  $0.03 \times$  smoke density, resulting in 2000 observations per trajectory on average. As a result, we have a sequence of smoke density observations made at different time points and spatial locations. Note that we do not make any Poisson process assumptions here. Instead, we use a realistic measurement process where sensors are utilized only when some action in the measured process starts to happen, which is the application we are targeting in our work.

We use the first half (1.25 sec.) of each trajectory as the context for the initial state inference. In total, the dataset contains 100 trajectories. We train our and other models on 80 trajectories, and use 10 trajectories for validation, and 10 for testing.

## B MODEL, POSTERIOR, AND ELBO

### B.1 MODEL

As described in Section 4.1 our model is defined as

$$\mathbf{z}_1 \sim p(\mathbf{z}_1), \quad \frac{d\mathbf{z}(\tau)}{dt} = f(\mathbf{z}(\tau)), \quad (41)$$

$$\tilde{\mathbf{z}}(t) = \text{interpolate}(t; \mathbf{z}(\tau_1), \dots, \mathbf{z}(\tau_n)), \quad (42)$$

$$\mathbf{u}(t, \mathbf{x}) = \phi(\tilde{\mathbf{z}}(t), \mathbf{x}), \quad (43)$$

$$t_i, \mathbf{x}_i \sim \text{nhpp}(\lambda(\mathbf{u}(t, \mathbf{x}))), \quad i = 1, \dots, N, \quad (44)$$

$$\mathbf{y}_i \sim p(\mathbf{y}_i | g(\mathbf{u}(t_i, \mathbf{x}_i))), \quad i = 1, \dots, N.. \quad (45)$$

The corresponding joint distribution is

$$p(\mathbf{z}_1, \{t_i, \mathbf{x}_i, \mathbf{y}_i\}_{i=1}^N) = p(\mathbf{z}_1) p(\{t_i, \mathbf{x}_i\}_{i=1}^N | \mathbf{z}_1) \prod_{i=1}^N p(\mathbf{y}_i | \mathbf{z}_1, t_i, \mathbf{x}_i), \quad (46)$$

where

$$p(\mathbf{z}_1) = \mathcal{N}(\mathbf{0}, I), \quad (47)$$

$$p(\{t_i, \mathbf{x}_i\}_{i=1}^N | \mathbf{z}_1) = \prod_{i=1}^N \lambda(\mathbf{u}(t_i, \mathbf{x}_i)) \exp\left(-\int \lambda(\mathbf{u}(t, \mathbf{x})) d\mathbf{x} dt\right), \quad (48)$$

$$p(\mathbf{y}_i | \mathbf{z}_1, t_i, \mathbf{x}_i) = \mathcal{N}(g(\mathbf{u}(t_i, \mathbf{x}_i)), \sigma_{\mathbf{y}}^2 I), \quad (49)$$

where  $\mathcal{N}$  is the normal distribution,  $\mathbf{0}$  is a zero vector, and  $I$  is the identity matrix. We set  $\sigma_{\mathbf{y}}$  to  $10^{-3}$ .



## B.2 POSTERIOR

We define the approximate posterior with  $\psi = [\psi_\mu, \psi_{\sigma^2}]$  as

$$q(\mathbf{z}_1; \psi) = \mathcal{N}(\mathbf{z}_1 | \psi_\mu, \text{diag}(\psi_{\sigma^2})), \quad (50)$$

where  $\mathcal{N}$  is the normal distribution, and  $\text{diag}(\psi_{\sigma^2})$  is a matrix with vector  $\psi_{\sigma^2}$  on the diagonal.

As discussed in Section 4.3, the encoder maps the context  $\{t_i, \mathbf{x}_i, \mathbf{y}_i\}_{i=1}^{N_{\text{ctx}}}$  to the local variational parameters  $\psi$  (which we break up into  $[\psi_\mu, \psi_{\sigma^2}]$ ). The encoder first converts the context sequence  $\{t_i, \mathbf{x}_i, \mathbf{y}_i\}_{i=1}^{N_{\text{ctx}}}$  to a sequence of vectors  $\{\mathbf{b}_i\}_{i=1}^{N_{\text{ctx}}}$  and then adds an aggregation token, which gives the following input sequence:  $\{\{\mathbf{b}_i\}_{i=1}^{N_{\text{ctx}}}, [\text{AGG}]\}$ . We pass it through the transformer encoder and read the value of the aggregation token at the last layer, which we denote by  $[\text{AGG}]$ . Then, we map  $[\text{AGG}]$  to  $[\psi_\mu, \psi_{\sigma^2}]$  as follows:

$$\psi_\mu = \text{Linear}(\overline{[\text{AGG}]}), \quad (51)$$

$$\psi_{\sigma^2} = \exp(\text{Linear}(\overline{[\text{AGG}]}) ), \quad (52)$$

where Linear are separate linear mappings.

## B.3 ELBO

Given definitions in the previous sections, we can write the ELBO as

$$\mathcal{L} = \int q(\mathbf{z}_1) \ln \frac{p(\mathbf{z}_1, \{t_i, \mathbf{x}_i, \mathbf{y}_i\}_{i=1}^N)}{q(\mathbf{z}_1)} d\mathbf{z}_1 \quad (53)$$

$$= \int q(\mathbf{z}_1) \ln \frac{\prod_{i=1}^N p(\mathbf{y}_i | \mathbf{z}_1, t_i, \mathbf{x}_i) p(\{t_i, \mathbf{x}_i\}_{i=1}^N | \mathbf{z}_1) p(\mathbf{z}_1)}{q(\mathbf{z}_1)} d\mathbf{z}_1 \quad (54)$$

$$= \int q(\mathbf{z}_1) \ln \prod_{i=1}^N p(\mathbf{y}_i | \mathbf{z}_1, t_i, \mathbf{x}_i) d\mathbf{z}_1 \quad (55)$$

$$+ \int q(\mathbf{z}_1) \ln p(\{t_i, \mathbf{x}_i\}_{i=1}^N | \mathbf{z}_1) d\mathbf{z}_1 \quad (56)$$

$$+ \int q(\mathbf{z}_1) \ln \frac{p(\mathbf{z}_1)}{q(\mathbf{z}_1)} d\mathbf{z}_1 \quad (57)$$

$$= \underbrace{\sum_{i=1}^N \mathbb{E}_{q(\mathbf{z}_1)} [\ln p(\mathbf{y}_i | \mathbf{z}_1, t_i, \mathbf{x}_i)]}_{\mathcal{L}_1} \quad (58)$$

$$+ \underbrace{\mathbb{E}_{q(\mathbf{z}_1)} \left[ \sum_{i=1}^N \ln \lambda(\mathbf{u}(t_i, \mathbf{x}_i)) - \int \lambda(\mathbf{u}(t, \mathbf{x})) d\mathbf{x} dt \right]}_{\mathcal{L}_2} \quad (59)$$

$$- \underbrace{\text{KL}[q(\mathbf{z}_1) || p(\mathbf{z}_1)]}_{\mathcal{L}_3}. \quad (60)$$

$$= \mathcal{L}_1 + \mathcal{L}_2 - \mathcal{L}_3. \quad (61)$$

**Computing ELBO.** We compute the ELBO as follows:

1. Compute local variational parameter from the context as  $\psi_\mu, \psi_{\sigma^2} = \text{Encoder}(\{t_i, \mathbf{x}_i, \mathbf{y}_i\}_{i=1}^{N_{\text{ctx}}})$
2. Sample the latent initial state as  $\mathbf{z}_1 \sim \mathcal{N}(\mathbf{z}_1 | \psi_\mu, \text{diag}(\psi_{\sigma^2}))$  using reparameterization
3. Evaluate  $\mathbf{z}(\tau_1), \dots, \mathbf{z}(\tau_n)$  by solving  $\frac{d\mathbf{z}(\tau)}{d\tau} = f(\mathbf{z}(\tau))$  with  $\mathbf{z}_1$  as the initial condition. We solve the ODE using `torchdiffeq` package (Chen et al., 2018).

4. Compute the latent states  $\tilde{z}(t_1), \dots, \tilde{z}(t_N)$  via interpolation as  $\tilde{z}(t) = \text{interpolate}(t; \mathbf{z}(\tau_1), \dots, \mathbf{z}(\tau_n))$
5. Compute  $\mathbf{u}(t_1, \mathbf{x}_1), \dots, \mathbf{u}(t_N, \mathbf{x}_N)$  as  $\mathbf{u}(t_i, \mathbf{x}_i) = \phi(\tilde{z}(t_i), \mathbf{x}_i)$
6. Compute  $\mathcal{L}_1$  using Monte Carlo integration
7. Compute  $\mathcal{L}_2$  using Monte Carlo integration, with the intensity integral computed using `torchquad` (Gómez et al., 2024) package.
8. Compute  $\mathcal{L}_3$  analytically

Sampling is done using reparametrization (Kingma & Welling, 2013) to allow for exact gradient evaluation via backpropagation. Monte Carlo integration is done using sample size of one.

## C EXPERIMENTS SETUP

### C.1 DATASETS

See Appendix A for all details about dataset generation. For all datasets we use 80%/10%10% train/validation/test splits. The context size  $N_{\text{ctx}}$  is determined for each trajectory separately based on the time interval that we consider to be the context, for Burgers’ and Navier-Stokes it is the first 0.5 seconds (out of 2 seconds), while for Shallow Water it is 0.1 seconds (out of 1 second). Since the number of events occurring within this fixed time interval is different for each trajectory, the context size consequently differs across trajectories. See encoder architecture description below for details on how different context lengths are processed.

### C.2 TRAINING, VALIDATION, AND TESTING

We use AdamW (Loshchilov & Hutter, 2019) optimizer with constant learning rate  $3e-4$  (we use linear warmup for first 250 iterations). We train for 25000 iterations, sampling a random minibatch from the training dataset at every iteration. We do not use any kind of scaling for the ELBO terms.

We compute validation error (MAE on observations) on a single random minibatch from the validation set every 125 iterations. We save the current model weights if the validation error averaged over last 10 validation runs is the smallest so far.

To simulate the model’s dynamics we use differentiable ODE solvers from `torchdiffeq` package (Chen et al., 2018). In particular, we use the `dopri5` solver with  $\text{rtol} = \text{atol} = 10^{-5}$  without the adjoint method. To evaluate the intensity integral  $\int \lambda(\mathbf{u}(t, \mathbf{x})) d\mathbf{x} dt$  we use Monte Carlo integration from the `torchquad` (Gómez et al., 2024) package with 32 randomly sampled points for training, and 256 randomly sampled points for testing, we found these sample sizes to be sufficient for producing small variance of the integral estimation in our case.

### C.3 ENCODER AND MODEL ARCHITECTURES

**Encoder.** We use padding to efficiently handle input contexts of varying length. The time points, coordinates, and observations are linearly projected to the embedding space (128-dim for Burgers’ and Shallow Water, and 192-dim for Navier-Stokes). Then, the projections are summed and passed through a stack of transformer encoder layers (4 layers for Burgers’ and Shallow Water, and 5 layers for Navier-Stokes, 4 attention heads were used in all cases). We introduce a learnable aggregation token which we append at the end of each input sequence. The final layer’s output is read from the last token of the output sequence (corresponds to the aggregation token) and is mapped to the local variational parameters as discussed in Appendix B.

**Model.** The dynamics function  $f(\mathbf{z}(t))$  is an MLP with 3 layers 368 neurons each, and GeLU (Hendrycks & Gimpel, 2016) nonlinearities. We set resolution of the uniform temporal grid  $\tau_1, \dots, \tau_n$  to  $n = 50$ . The dynamics function maps the current latent state  $\mathbf{z}(t)$  to its time derivative  $\frac{d\mathbf{z}(t)}{dt}$ , both are  $d_{\mathbf{z}}$ -dimensional vectors.

The function  $\phi(\tilde{z}(t), \mathbf{x})$  is an MLP with 3 hidden layers with width 368 for Burgers’, 256 for Shallow Water, and 512 for Navier-Stokes. We use GeLU (Hendrycks & Gimpel, 2016) nonlinearities. As the

input the MLP takes the sum of  $\tilde{z}(t) \in \mathbb{R}^{d_z}$  with  $\text{proj}(\mathbf{x}) \in \mathbb{R}^{d_z}$ , where  $\text{proj}$  is a linear projection, and maps the sum to the output  $\mathbf{u}(t, \mathbf{x}) \in \mathbb{R}^{d_u}$ .

The intensity function  $\lambda(\mathbf{u}(t, \mathbf{x}))$  is an MLP with 3 layers, 256 neurons each, and GeLU (Hendrycks & Gimpel, 2016) nonlinearities. We further exponentiate the output of the MLP and add a small constant (1e-4) to it to ensure the intensity is positive and to avoid numerical instabilities.

We set the observation function  $g(\mathbf{u}(t, \mathbf{x}))$  to the identity function, thus  $g(\mathbf{u}(t, \mathbf{x})) = \mathbf{u}(t, \mathbf{x})$ .

## C.4 MODEL COMPARISON

### C.4.1 DYNAMIC SPATIOTEMPORAL MODELS

In this section we discuss neural spatiotemporal models we used for the comparisons. But first, we discuss data preprocessing that we do in order to make these models applicable.

**Data Preprocessing.** All models in this section assume that dense observations are available at every time point, and some further assume the observations are located on a uniform spatial grid. To make our data satisfy these requirements we use time binning and spatial interpolation. First, we use time binning and divide the original time grid into 10 bins and group all time points into 10 groups based on which bin they belong to. We used 10 time bins as it ensured that there was a sufficient number of time points in each bin while also being sufficient to capture changes of the system state. We denote by  $t_i$  the time point that represents the bin  $i$ . Next, we use all observations that belong to bin  $i$  and interpolate them to a uniform 32x32 spatial grid. In many cases some points of the spatial grid were outside of the convex hull of the interpolation points, so we set values of such points to -1 (values of the state range from 0 to 1). We used linear interpolation. We denote the resulting interpolated values at time  $t_i$  as  $\mathbf{u}_i$  which is the interpolated system state at the 32x32 spatial grid. When we compute the loss between  $\mathbf{u}_i$  and the corresponding model prediction we use only those spatial locations that had proper interpolation values (i.e. were inside the convex hull of the interpolation points).

**Median predictor.** This is a simple baseline where we compute the median of the observations  $\mathbf{y}_i$  on the training data and use it as the prediction on the test data.

**CNN-ODE.** This is another relatively simple baseline based on latent neural ODEs (Chen et al., 2018). It operates in some sense similarly to our model as it takes a sequence of initial observations (context)  $\mathbf{y}_1, \dots, \mathbf{y}_{N_{\text{ctx}}}$  and uses a CNN encoder to map the context to the latent initial state  $\mathbf{z}_1$ . Then it maps  $\mathbf{z}_1$  to  $\mathbf{z}_2, \dots, \mathbf{z}_N$  via a latent ODE, and finally decodes  $\mathbf{z}_i$  to obtain  $\mathbf{u}_i$  via a CNN decoder. In our case we used the first two observations  $\mathbf{u}_1$  and  $\mathbf{u}_2$  as the context.

**FEN.** This model discretizes the spatial domain into a mesh and uses a graph neural network to model the dynamics of the state at each node of the mesh. The model starts at  $\mathbf{u}_1$  and maps it directly to  $\mathbf{u}_2, \dots, \mathbf{u}_N$  via an ODE solver without using any context. We simulated the dynamics as stationary and autonomous, with the free-form term active. The dynamics function is a 2-layer MLP with the width of 512 neurons and ReLU nonlinearities. All other hyper-parameters were left as defaults as we found that changing them did not improve the results. We used the official code from Lienen & Günnemann (2022).

**NSPDE.** This model applies a point-wise transformation to map the system state at every spatial node to a high-dimensional latent representation, then uses the fixed-point method to simulate the dynamics at every node, and finally maps the high-dimensional latent states back to the observation space to obtain the predictions. We used the fixed-point method with maximum number of available modes and a single iteration as we found that larger number of iterations does not improve the results. All other hyper-parameters were left as defaults as we found that changing them did not improve the results. We used the official code from Salvi et al. (2022).

### C.4.2 NEURAL SPATIOTEMPORAL PROCESSES

**Constant Intensity Baseline.** As a simple baseline we use the following intensity function  $\lambda(t, \mathbf{x}) = c$ , where  $c$  is a learnable parameter that we fit on the training data.

$d_z$	$h_\phi$	$h_f$			
		23	92	368	512
40	128	0.090	0.091	0.084	0.081
40	512	0.075	0.073	0.074	0.072
368	128	0.084	0.083	0.084	0.082
368	512	0.072	0.073	0.073	0.072

Table 4: Dependence of the test MAE on latent space dimension  $d_z$ , dynamics complexity  $h_f$  and decoder complexity  $h_\phi$ .

**DSTPP.** This model uses a transformer-based encoder to condition a diffusion-based density model on the event history. We use 500 time steps and 500 sampling steps, and train for at most 12 hours. All other hyper-parameters were left as defaults as we found that changing them did not improve the results or made the training too slow. We used the official code from [Yuan et al. \(2023\)](#).

**NSTPP.** This model represents the vent history in terms of a latent state governed by a neural ODE, with the latent state being discontinuously updated after each event. The latent state is further used to represent the distribution of spatial locations via a normalizing flow model. We used the attentive CNF version of the model. All other hyper-parameters were left as defaults as we found that changing them did not improve the results. We used the official code from [Chen et al. \(2021\)](#).

**AutoSTPP.** This model uses dual network approach for flexible and efficient modeling of spatiotemporal processes. We used 5 product networks with 2 layers 128 neurons each and hyperbolic tangent nonlinearities. The step size was set to 20 and number of steps was set to 101 in each direction. All other hyper-parameters were left as defaults as we found that changing them did not improve the results. We used the official code from [Zhou & Yu \(2023\)](#).

## D EFFECTS OF DYNAMICS FUNCTION AND DECODER COMPLEXITY

Our model (similarly to previous competing models) consists of a composition of mappings that are parameterized by deep neural networks. Therefore, the role and importance of different model components depends on a number of hyperparameters. This, in turn, is related to the result reported in the main text (Figure 5), where we observed that, in the case of linear interpolation, decreasing the number of interpolation points did not decrease predictive performance of our model. Decreasing the number of interpolation points to  $n = 2$  corresponds to replacing non-linear dynamics function with a linear model, but this decrease in complexity of the dynamics model can, in some cases, be compensated by the capacity of decoder function. In this section, we conduct a series of experiments to better understand the role of the dynamics and decoder functions in our model. In particular, we look at the model’s predictive accuracy as a function of the latent space dimension  $d_z$ , and complexity of the dynamics function  $f$  and decoder  $\phi$ . Since both  $f$  and  $\phi$  are 3-layer MLPs, we define their complexities as the width of the hidden layers  $h_f$  and  $h_\phi$ , respectively. In particular, we look at low- and high-dimensional latent spaces ( $d_z = 40$  and  $d_z = 368$ ); we consider medium and large decoder function  $\phi$  (with  $h_\phi = 128$  and  $h_\phi = 512$ ), and dynamics functions of different complexities ranging from  $h_f = 23$  to  $h_f = 512$ . We use linear interpolation with 50 interpolation points to allow the model represent complex latent trajectories if it helps to fit the data better. We use the Navier-Stokes data (since this is the most complex synthetic dataset in our work) and report only the mean absolute error (as log-likelihood follows the same pattern). The results are in Table 4.

As can be seen, for the low-dimensional latent space ( $d_z = 40$ ) and medium-sized decoder (with  $h_\phi = 128$ ) increasing complexity of the dynamics function  $h_f$  leads to a meaningful reduction of MAE. However, after increasing complexity of the decoder from  $h_\phi = 128$  to  $h_\phi = 512$ , increasing  $h_f$  results in no significant improvements. Then, if we switch to a high-dimensional latent space ( $d_z = 368$ ), we see that complexity of the decoder  $h_\phi$  becomes the main determinant of the model’s predictive accuracy, with dynamics function complexity  $h_f$  having no observable effect.

In summary, we see that increasing complexity of the dynamics function of the neural ODE can be useful in some cases (low-dimensional latent space and medium-sized decoder), but with a sufficiently

large decoder and large latent spaces its benefits become less apparent, meaning that in such cases the dynamics can be modeled using simple function approximators. While the experimental results show that in some cases accurate predictions can be achieved with simple dynamics, enabling the model to represent complex dynamics via Neural ODEs can be useful also for cases where complex latent trajectories are expected or enforced.