

Lawrence Berkeley National Laboratory

LBL Publications

Title

A Statistician's Overview of Physics-Informed Neural Networks for Spatio-Temporal Data

Permalink

<https://escholarship.org/uc/item/7590w4xd>

Journal

Journal of the American Statistical Association, 121(554)

ISSN

0162-1459

Authors

Wikle, Christopher K

North, Joshua

Gopalan, Giri

et al.

Publication Date

2026-04-03

DOI

10.1080/01621459.2026.2625420

Copyright Information

This work is made available under the terms of a Creative Commons Attribution-NonCommercial-NoDerivatives License, available at

<https://creativecommons.org/licenses/by-nc-nd/4.0/>

Peer reviewed

A Statistician's Overview of Physics-Informed Neural Networks for Spatio-Temporal Data

Christopher K. Wikle*

University of Missouri
wiklec@missouri.edu

Joshua North

Lawrence Berkeley National Laboratory
jsnorth@lbl.gov

Giri Gopalan

Los Alamos National Laboratory
ggopalan@lanl.gov

Myungsoo Yoo

The University of Texas at Austin
myungsoo.yoo@austin.utexas.edu

ABSTRACT

The recent success of deep neural network models with physical constraints (so-called, Physics-Informed Neural Networks, PINNs) has led to renewed interest in the incorporation of mechanistic information in predictive models. Statisticians and others have long been interested in this problem, which has led to several practical and innovative solutions dating back decades. In this overview, we focus on the problem of data-driven prediction and inference of dynamic spatio-temporal processes that include mechanistic information, such as would be available from partial differential equations, with a strong focus on the quantification of uncertainty associated with data, process, and parameters. We give a brief review of several paradigms and focus our attention on Bayesian implementations given they naturally accommodate uncertainty quantification. We then show that it is straight-forward to include the Bayesian PINN (B-PINN) within the Bayesian hierarchical model (BHM) framework that has long been considered for modeling dynamic spatio-temporal processes. Such a BHM-PINN is illustrated via a simulation study in which a latent nonlinear Burgers' equation PDE governs the dynamics of Poisson distributed spatio-temporal data.

Keywords Bayesian, Burgers, Deep Learning, Hierarchical, Spatio-Temporal, Uncertainty Quantification

*Corresponding author

1 Introduction

Paraphrasing Galileo, the laws of nature are written in the language of mathematics. Indeed, much of our understanding of the world around us, including physical laws and principles, is codified in terms of mechanistic formulas, such as partial and ordinary differential equations. From a statistical perspective, one can attempt to infer mechanistic equations that describe the system under study. Moreover, one can use *a priori* knowledge of mechanistic equations describing a physical process of interest, along with data, for inference (e.g., learning the parameters that control the specified equations or initial conditions) or prediction (e.g., forecasting/interpolating).

When data are involved, one inherently has uncertainty when building models and optimization methods become an integral part of understanding the world around us. The classic example is when Carl Friedrich Gauss used the method of least squares in the early 1800s to successfully calculate the orbits of celestial bodies; importantly, Gauss used Kepler’s laws of celestial mechanics as constraints to his least squares solution (Gauss, 1809). Indeed, statisticians have long been interested in using proposed mathematical relationships in the presence of observations to motivate statistical models. For example, as described in Wikle & Hooten (2010), in 1927, Yule used pendulum motion differential equations to suggest an autoregressive time series for sunspot observations and Hotelling used approximations of differential equations to develop models for population growth in the United States. The integration of mechanistic information in statistical models is now a fundamental component of several sub-disciplines of statistics; e.g., statistical meteorology (Berliner et al., 1999), ecological statistics (Hefley et al., 2017), epidemiology (Leach et al., 2020), glaciology (Gopalan et al., 2018), statistical oceanography (Herbei et al., 2008; Milliff et al., 2011), pharmacokinetic and pharmacodynamic analysis in the development of pharmaceuticals (e.g., Gabrielsson & Weiner, 2001), to name a few.

With the advent of deep learning, the use of mechanistic information combined with deep neural network models has seen an exponential increase in applications across many disciplines. Such approaches, often known as Physics Informed Neural Networks (PINNs, Raissi et al., 2017a), are typically based on augmenting the neural network loss (e.g., mean squared error) with loss terms corresponding to the physical equations and boundary/initial conditions (e.g., Karniadakis et al., 2021). Although this is novel in the context of neural network models, augmenting loss functions to include mechanistic information when fitting models to data is not new. For example, Sasaki pioneered the approach in a variational solution to the data assimilation problem in meteorology beginning in the 1950s (e.g., Sasaki, 1958, 1970). Around the same time, control engineers were combining (linearized) dynamical equations within a filter that updated in the presence of new observations (e.g., Kalman, 1960). Additional approaches developed in engineering and statistics to “fuse” data, random processes, and deterministic models were developed starting in the 1990s and continue to this day (e.g., see overviews in Berliner, 2003; Cockayne et al., 2019; Meng et al., 2020).

From a statistical perspective, the goal is not to develop a “solver” for a known mechanistic set of equations as is often of interest in the PINN literature (Raissi et al., 2019), but rather to recognize, in the G. Box sense, that all our mechanistic models are in some way “wrong” when applied to real world scenarios, and that we need to account for the uncertainty in the model specification, parameters, and data. In this sense, the Bayesian hierarchical modeling

(BHM) approach has long been seen as the natural framework to account for all of these sources of uncertainty in physics-motivated statistical modeling (e.g., see [Berliner et al., 1999](#); [Cressie & Wikle, 2011](#); [Hennig et al., 2015](#)). The Bayesian view has also increasingly been shared in the PINN community (e.g., [Yang et al., 2021](#); [Meng et al., 2022](#); [Huang et al., 2023](#); [Psaros et al., 2023](#)). This has been an important development from a statistical perspective. Yet, some components of the BHM modeling paradigm are often not explicitly represented in these implementations, and there is still very little consideration of the importance of specifying dependent random effects, which has been shown to be very useful in statistical mechanistic modeling of spatio-temporal processes (e.g., [Wikle et al., 2001](#); [Cressie & Wikle, 2011](#); [Chkrebtii et al., 2016](#)).

Our goal in this paper is to present a statistically-minded overview of mechanistically-informed modeling in the presence of data and uncertainty to provide context in our discussion of PINNs. Note, although we generally use the PINN acronym due to its ubiquitous use in the literature, we prefer the term “mechanistically-informed” to “physics-informed” because many mechanistic models for non-physical processes (e.g., invasive species in ecology, [Wikle \(2003\)](#)) also follow this paradigm and it can be misleading to use the term “physics” for such processes. We will illustrate how it is natural to consider a PINN within a traditional BHM as is commonly considered for dynamic spatio-temporal statistical models (e.g., see [Cressie & Wikle, 2011](#)); we call this a BHM-PINN. We first provide a brief historical overview of how mechanistic process information has been combined with data in various modeling paradigms. Throughout, we discuss issues of interest to statisticians, such as uncertainty in data, parameters, and process knowledge, as well as computational tractability.

Section 2 discusses the general problem at hand and introduces the relevant notation. Then, in Section 3 we present a brief review of mechanistically-informed modeling from the perspective of variational data assimilation and optimal interpolation, as well as statistical approaches, with a focus on Bayesian and hierarchical Bayesian implementations to account for uncertainty. Section 4 describes PINNs and Bayesian-PINNs (B-PINNs) as well as connections to the aforementioned earlier literature. Section 5 discusses a how it is natural to include a B-PINN in a traditional BHM framework for dynamic spatio-temporal models, which we call a BHM-PINN. In Section 6, we apply the BHM-PINN framework to Poisson distributed spatio-temporal data where the latent process corresponds to a mechanistic PDE parameterized using Burgers’ equation. We conclude with a brief overview and discussion of challenges in Section 7.

2 Problem Outline

Our interest in mechanistically-informed models here concerns problems common to spatio-temporal statistics where one has potentially incomplete and noisy observations of a univariate or multivariate process at some locations in space and time, and seeks to: (a) determine an initial condition (i.e., the true process values at specified spatial locations at some initial time), (b) predict (interpolate) incomplete observations at specified locations in space and time, or (c) forecast at specified spatial locations at times beyond which we have observations. We generally assume that we have some *a priori* knowledge of the underlying spatio-temporal dynamics of the process, yet our knowledge is incomplete. That is, we may not know the parameters of the mechanistic model and/or the mechanistic model is an incomplete

representation of the process. An incomplete representation of dynamics could be due to many circumstances – for instance, due to approximations to the underlying partial differential equations, or in other instances due to missing mechanistic understanding. Next, we outline the notation to be used in this exposition.

2.1 Process and Data Notation

Assume our interest is in a K -variate process at spatial locations $\mathbf{s} \in \mathcal{D}$ (in $\mathbb{R}^1$, $\mathbb{R}^2$, or $\mathbb{R}^3$) and times $t \in [0, T + \tau]$, denoted by $\mathbf{u}_t(\mathbf{s}) \equiv (u_t^{(1)}(\mathbf{s}), \dots, u_t^{(K)}(\mathbf{s}))'$. Ultimately, we are interested in this process at some fixed set of n spatial locations, say $\mathcal{S} \equiv \{\mathbf{s}_1, \dots, \mathbf{s}_n\}$ and some discrete set of $n_{\mathcal{T}}$ times $\mathcal{T} \subseteq \{0, 1, \dots, T + \tau\}$ (although, the approach presented here can accommodate continuous time). If we stack these vectors, we can write the Kn -vector $\mathbf{u}_t \equiv (\mathbf{u}_t(\mathbf{s}_1)', \dots, \mathbf{u}_t(\mathbf{s}_n'))'$. For example, if the process is univariate then $K = 1$ and $\mathbf{u}_t \equiv (u_t(\mathbf{s}_1), \dots, u_t(\mathbf{s}_n))'$. More generally, observations could be available at different locations for each of the K variables of interest, but we do not add that additional notational complexity here.

Assume we have observations of elements of the process corresponding to some or all of the times in $\mathcal{T}$ at m_t spatial locations, which we denote by the m_t -vector $\mathbf{z}_t$. Some or all of these observation locations may coincide with the spatial locations in $\mathcal{S}$, but may not, as discussed below. Note, it may be the case that $m_t = 0$ for some times (i.e., there are no observations for certain times).

2.2 Mechanistic Model Notation

Here we focus our attention on mechanistic models that take the form of a time-dependent partial differential equation (PDE), but the ideas presented here can be easily adapted to other forms of mechanistic models (e.g., ordinary differential equations, or agent-based models). Assume we have a PDE that we believe is mechanistically informative to our latent process of interest, $\mathbf{u}_t(\mathbf{s})$. We denote this as

$$\frac{\partial \mathbf{v}_t(\mathbf{s})}{\partial t} = \mathcal{M}(\mathbf{v}_t(\mathbf{s}), \boldsymbol{\theta}_m), \quad (1)$$

where $\mathbf{v}_t(\mathbf{s})$ corresponds to the K -dimensional mechanistic process defined at $t \in [0, T + \tau]$ and $\mathbf{s} \in \mathcal{D}$ and $\mathcal{M}$ is an operator controlled by parameters $\boldsymbol{\theta}_m$. For example, we may believe that our latent process shows evidence of advective and diffusive dynamics, both of which can be represented through appropriate terms in $\mathcal{M}$. Note, we are deliberately making a distinction between the true latent process of interest, $\mathbf{u}_t(\mathbf{s})$, and the deterministic mechanistic model representation, denoted by $\mathbf{v}_t(\mathbf{s})$ - this distinction will be made more clear when we discuss models in Section 3 and 4. We also assume that the PDE has associated initial and boundary conditions, respectively,

$$\begin{aligned} \mathbf{v}_0(\mathbf{s}) &= \mathbf{g}_I(\mathbf{s}), \quad \mathbf{s} \in \mathcal{D}, \\ \mathbf{B}(\mathbf{v}_t(\mathbf{s})) &= 0, \quad t \in [0, T + \tau], \mathbf{s} \in \mathcal{D}_B, \end{aligned}$$

where $\mathbf{g}_I(\cdot)$ is some function defined at the initial time, 0, $\mathbf{B}(\cdot)$ is a function defined on the boundary $\mathcal{D}_B$.

Our goal is to perform inference on $\mathbf{u}_t(\mathbf{s})$ at the $\mathcal{S}$ spatial locations and at times $t \in \mathcal{T}$ while taking into account the information present in $\mathcal{M}$ and given the data $\mathbf{Z} \equiv \{\mathbf{z}_1, \dots, \mathbf{z}_T\}$. We can denote this as a posterior distribution (with

some abuse of notation) as $[\mathbf{u}_t : t \in \mathcal{T} | \mathbf{Z}, \mathcal{M}]$, where we use the bracket notation $([\])$ to represent a probability distribution, as is common in the Bayesian hierarchical modeling literature (e.g., [Gelfand & Smith, 1990](#); [Cressie & Wikle, 2011](#)); that is, $[a|b]$ corresponds to the distribution of a given (conditional upon) b , $[b]$ corresponds to the marginal distribution of b , and $[a, b]$ corresponds to the joint distribution of a and b . Typically, we are interested in learning the latent process at some initial time $\mathbf{u}_0$, interpolation at some time $t \in (0, T]$, $\mathbf{u}_t$, or forecasting at some time $T + \tau$, $\mathbf{u}_{T+\tau}$. This is generally termed the “forward problem” where $\mathcal{D}$, $\mathcal{D}_B$, $\mathbf{g}_I(\cdot)$, $\mathbf{B}(\cdot)$, and $\boldsymbol{\theta}_m$ are known and one seeks to find $\mathbf{u}_0(s)$ and $[\mathbf{u}_t]_{t \in (0, T+\tau]}$ given the dynamics suggested by $\mathcal{M}$. We may also be interested in learning parameters associated with the mechanistic PDE ($\boldsymbol{\theta}_m$) and various parameters associated with the underlying statistical or neural models used to represent the PDE. This is generally termed the “inverse problem,” where we may have partial observations of the process in the domain of interest. The primary goal of the methodology presented in this work is to infer the latent process $\{\mathbf{u}_t\}_{t \in [0, T+\tau]}$ on $\mathcal{D}$ as well as the $\boldsymbol{\theta}_m$ that satisfies $\mathcal{M}$ given observations $\mathbf{Z}$.

In the exposition that follows, we will sometimes refer to the numerical solution of the PDE in the short-hand notation,

$$\mathbf{v}_{t+1} = \mathcal{M}(\mathbf{v}_t; \boldsymbol{\theta}_m), \quad (2)$$

but recognize that there are PDE solvers that provide $\mathbf{v}_t$ for any and all times in $[0, T]$ that are not necessarily sequential in time (e.g., implicit solvers).

3 Historical Perspective

Fundamental advances in mechanistic-informed modeling in the presence of data originated with the operational need to assimilate observations with physical models of the atmosphere for weather forecasting (i.e., numerical solutions of simplified Navier-Stokes PDEs). The main purpose of such “data assimilation” is to generate initial conditions for the numerical forecast models that are consistent with the dynamics of the system, but these initial conditions have incomplete observations in space (e.g., we cannot observe the atmosphere everywhere). We discuss the data assimilation approach in Section 3.1. The connection to more statistical approaches and further developments are discussed in Section 3.2.

3.1 Variational Data Assimilation

It was known shortly after L.F. Richardson’s failed attempt at the first numerical weather forecast in the early 1920s ([Richardson, 1922](#)) that the numerical solution of PDEs with initial conditions given by data would quickly become unstable due to lack of consistency with the governing equations. Thus, by the time that computers were actually capable of generating a realistic forecast in real time, it was known that initial condition fields of weather variables had to be smoothed in a way that respected dynamical constraints ([Charney et al., 1950](#); [Phillips, 1956](#)). Two main approaches were developed in the 1950s and 1960s to do this, one deterministic based on the calculus of variations and one stochastic, based on what would come to be known as Gaussian processes (GPs). In this section, we briefly discuss

these approaches as they provide the foundation for mechanistically-informed models for spatio-temporal processes in statistics.

The initial idea in data assimilation, that one could minimize an objective function containing the sum of squared differences between the observations and the “true” initial field while simultaneously satisfying a constraint that incorporated the PDE dynamics, was presented in [Sasaki \(1958\)](#). Although he formulated this in continuous space, practical uses of the methodology considered discrete space (e.g., [Thompson, 1969](#); [Sasaki, 1970](#); [Stephens, 1970](#)).

Following the notation in Section 2.1, say we have m_t -dimensional data vectors, $\mathbf{z}_t$, and numerical solutions of the deterministic PDE state process at time t denoted by $\mathbf{v}_t$ given by (2). Here, without loss of generality, we assume $K = 1$ (a univariate process) for notational simplicity. Critically, for deterministic systems, if one knows the initial state $\mathbf{v}_0$ then (2) determines the sequence of future states, $\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_t, \dots\}$. In Sasaki’s variational approach, one seeks an initial state $\mathbf{v}_0$ that combines the physics from the PDE and observations to obtain the initial condition. Importantly, in the context of weather forecasting, it is this estimated initial condition that is used to generate forecasts from the deterministic model, $\{\mathbf{v}_t, t = 1, \dots\}$.

Sasaki’s approach estimates the initial state, $\mathbf{v}_0$, by minimizing an objective function, $J_0(\mathbf{v}_0)$, subject to the constraint that the state evolution follows the numerical PDE solver exactly. That is, (e.g., see [Lewis & Lakshmivarahan, 2008](#)) the optimization problem is to minimize

$$J_0(\mathbf{v}_0) = \frac{1}{2} \sum_{t=1}^T (\mathbf{z}_t - \mathbf{h}(\mathbf{v}_t))' \mathbf{R}_t^{-1} (\mathbf{z}_t - \mathbf{h}(\mathbf{v}_t)), \quad (3)$$

subject to the constraint of (2) – i.e., $\mathbf{v}_{t+1} = \mathcal{M}(\mathbf{v}_t; \boldsymbol{\theta}_m)$ for $t = 0, \dots, T - 1$; this constraint allows each of $\mathbf{v}_1, \dots, \mathbf{v}_T$ to be written as a function of $\mathbf{v}_0$ by iterating application of $\mathcal{M}(\cdot; \boldsymbol{\theta}_m)$. Additionally, $\mathbf{h}(\cdot)$ is a (known) mapping function that takes the n -dimensional deterministic model state vector $\mathbf{v}_t$ and maps to the m_t -dimensional observation vector at time t (e.g., an interpolator if data are not coincident with model locations, or perhaps the identity if they are; this function can also take into account nonlinear transformations, see [Daley \(1993\)](#)). In addition, $\mathbf{R}_t$ is the time-varying $m_t \times m_t$ observation error covariance matrix (representing measurement error and other sources of observational error), also typically assumed known. [Lewis & Lakshmivarahan \(2008\)](#) also describe how this constrained optimization problem can be solved via the method of Lagrange multipliers. We emphasize that this approach is considered a “strong” constraint because the state solution must follow the deterministic model exactly ([Sasaki, 1970](#)).

3.1.1 Operational Variational Data Assimilation (4DVar)

In practice, variational data assimilation evolved away from Lagrange multipliers to an unconstrained approach known as “4DVar” (“4D” because it includes 3 dimensions in space and time). As with the Sasaki approach, one considers data in the interval $[1, T]$ and seeks the initial condition, $\mathbf{v}_0$. The observation component of the objective function is the same as (3) above where again the $\mathbf{v}_t$ are functions of the initial condition $\mathbf{v}_0$ and obtained deterministically from the numerical solution of the PDE (2), which is assumed to be error free. In this approach, one then assumes that there is some background “estimate” of the initial state condition, say $\mathbf{v}_b$, such as a forecast from the previous assimilation

cycle, or some historical average. Then, the background term of the objective function is given by,

$$J_{1b}(\mathbf{v}_0) = \frac{1}{2}(\mathbf{v}_0 - \mathbf{v}_b)' \mathbf{C}_{b_0}^{-1}(\mathbf{v}_0 - \mathbf{v}_b), \quad (4)$$

where $\mathbf{C}_{b_0}$ is the background error covariance matrix (a spatial covariance matrix for the errors between the true initial state and the previous forecast of the initial state). The optimization finds $\mathbf{v}_0$ that minimizes the sum of $J_0(\mathbf{v}_0) + J_{1b}(\mathbf{v}_0)$ (i.e., equation (3) plus (4)). In practice, the optimization is facilitated by using an adjoint method (e.g., [Lewis & Derber, 1985](#); [Errico, 1997](#)).

3.1.2 Weak Constraint 4DVar

One can consider weak-constraint variational assimilation by incorporating an explicit model error term in the objective function (e.g., [Bennett, 1992](#), via the method of representors). For example, say we have the model error term, $\boldsymbol{\eta}_t$, as in

$$\mathbf{u}_t = \mathcal{M}(\mathbf{u}_{t-1}; \boldsymbol{\theta}_m) + \boldsymbol{\eta}_t, \quad (5)$$

where $\boldsymbol{\eta}_t \sim \mathcal{N}(\boldsymbol{\mu}_\eta, \mathbf{Q})$, and the mean $\boldsymbol{\mu}_\eta$ vector and error covariance matrix $\mathbf{Q}$ are assumed known or estimated offline. In this section, we prefer to denote the process by $\mathbf{u}_t$ instead of $\mathbf{v}_t$ since $\mathbf{u}_t$ is not constrained to follow the numerical model exactly, as in the previous subsections.

Then, the extra term in the objective function in the 4DVar case is given by (e.g. [Laloyaux et al., 2020](#)),

$$J_2(\boldsymbol{\eta}) = \frac{1}{2} \sum_{t=1}^T (\boldsymbol{\eta}_t - \boldsymbol{\mu}_\eta)' \mathbf{Q}^{-1} (\boldsymbol{\eta}_t - \boldsymbol{\mu}_\eta), \quad (6)$$

where $\boldsymbol{\eta}_t \equiv \mathbf{u}_t - \mathcal{M}(\mathbf{u}_{t-1}; \boldsymbol{\theta}_m)$ from (5) and $\boldsymbol{\eta} = \{\boldsymbol{\eta}_1, \dots, \boldsymbol{\eta}_T\}$. Minimizing the loss function $J_0(\mathbf{u}_0, \boldsymbol{\eta}) + J_{1b}(\mathbf{u}_0) + J_2(\boldsymbol{\eta})$ (i.e., equation (3) plus (4) plus (6)) in terms of $\mathbf{u}_0$ and $\boldsymbol{\eta}$ is constrained by the numerical model but can deviate from an exact solution of the model because of $\boldsymbol{\eta}$, resulting in a weak constraint. Note that J_0 is written as a function of both $\mathbf{u}_0$ and $\boldsymbol{\eta}$ because both are necessary to determine $\mathbf{u}_1, \dots, \mathbf{u}_T$ as per (5).

3.1.3 Optimal Interpolation

Around the same time that the variational approach to meteorological data assimilation was being developed, other researchers were developing statistical optimization approaches based on stochastic processes. A brief history of these approaches is presented in [Lewis & Lakshmivarahan \(2008\)](#). Although there is some debate about its origins, [Lewis & Lakshmivarahan \(2008\)](#) describe the first practical approach to come from [Eliassen \(1954\)](#). More formal development was given by [Gandin \(1963\)](#). Gandin's approach is widely considered to have been equivalent to the formalization of kriging by Matheron ([Matheron, 1963](#)), which is the essence of geostatistical modeling ("kriging") in spatial statistics; see [Cressie \(1990\)](#) for a discussion of the origins of kriging. In retrospect, all these approaches simply use Gaussian processes to perform interpolation (spatial prediction) that is optimal in the sense of minimizing the expected squared error loss (i.e., risk).

In the context of meteorological data assimilation, one interpolates the residuals by removing a background field (e.g., the previous forecast) from observations. So, given an observation vector, $\mathbf{z}_t$, and a forecast from the previous time, $\mathbf{u}_b$

(where again, b stands for “background”; note, keeping with the notation used in this paper, when this forecast is from the numerical solution of a PDE, we could denote the background as $\mathbf{v}_b$), the kriging approach gives the latent state vector $\mathbf{u}_t$ from:

$$\mathbf{u}_t = \mathbf{u}_b + \mathbf{K}_t(\mathbf{z}_t - \mathbf{H}_t\mathbf{u}_b),$$

where $\mathbf{H}_t$ is an $m_t \times n$ matrix that maps the state process to the observations at time t , and $\mathbf{K}_t$ is the $n \times m_t$ matrix of interpolation weights given by

$$\mathbf{K}_t = \mathbf{C}_b\mathbf{H}_t'(\mathbf{H}_t\mathbf{C}_b\mathbf{H}_t' + \mathbf{R}_t)^{-1},$$

where $\mathbf{C}_b$ is the $n \times n$ spatial covariance matrix of the background (forecast) error. This matrix must be specified in terms of a covariance function, and because one may need to develop predictions from locations that are not observed, it comes from a covariance function associated with a Gaussian process (e.g., see [Cressie, 1993](#); [Cressie & Wile, 2011](#)).

There are different ways that physical constraints are included in this framework. First, note that the forecast for time t , which we are denoting by $\mathbf{u}_b$ here, comes from a forecast given by the dynamic model (2) and the previous estimate of the state, say $\mathbf{u}_{t-1}$. Alternatively, one can specify covariance functions that satisfy dynamical constraints, which is especially useful in multivariate applications of this methodology (e.g., see [Daley, 1993](#)).

Note, this formulation does not explicitly account for multiple observations over times; i.e., there is no dependence specified across time. In addition, this optimal interpolation approach for spatial fields can also be formulated from a variational perspective as in Section 3.1.1, known as “3DVar” in the data assimilation literature. Spatial optimal interpolation and 3DVar are equivalent if the observation mapping function is linear, i.e., $\mathbf{h}(\mathbf{u}_t) = \mathbf{H}_t\mathbf{u}_t$ (e.g., [Lewis & Lakshmivarahan, 2008](#)).

3.1.4 Connections to Filtering and Bayesian Estimation

About the same time that Sasaki was developing the variational approach to combining data and mechanistic model information, control engineers were working on the same problem from an online filtering perspective. The classic example of this is the Kalman filter ([Kalman, 1960](#)), which has become ubiquitous in control engineering and statistical state-space time-series and spatio-temporal analysis (e.g., see [Shumway & Stoffer, 2000](#); [Wikle & Berliner, 2007](#)). In the control engineering case, one typically assumes they have an initial condition and initial variance/covariance matrix and proceeds to generate a forecast for the next time; this forecast is then updated as new observations come online, and this updated state estimate is used to generate another forecast, etc. The filtering equations explicitly account for model error, as in the weak constraint 4DVar case.

It has long been known that there are connections between the variational optimization approach in data assimilation and Bayesian inference (e.g., [Lorenc, 1986](#); [Talagrand, 1997](#); [Wikle & Berliner, 2007](#)). This is not surprising to statisticians, as the maximum *a posteriori* (MAP) estimator in a Bayesian model is equivalent to a regularized optimization. That is, say we have data $\mathbf{D}$ and parameters $\boldsymbol{\theta}$, then the MAP estimator is given by maximizing the numerator in Bayes formula

with respect to θ

$$\begin{aligned}\hat{\theta}_{MAP} &= \arg \max_{\theta} ([D|\theta][\theta]) \\ &= \arg \max_{\theta} (\log[D|\theta] + \log[\theta]),\end{aligned}$$

which of course is equivalent to minimizing $(-\log[D|\theta] - \log[\theta])$. The connection is then obvious in that the first term (Bayesian likelihood) corresponds to a loss function such as (3) and the second term (Bayesian prior) corresponds to a regularization term such as (4). A classic example in statistics is ridge regression, where a least squares regression is regularized by an ℓ_2 penalty on the parameters, which is equivalent to a multivariate normal prior on the parameters. Typically, assuming one has a reasonably strong belief that their models are correct, the Bayesian approach is often preferred because one gets natural uncertainty quantification. However, in the context of high-dimensional nonlinear data and process models that are used in operational data assimilation, the variational approach has been an effective and reasonably efficient assimilation approach where fully Bayesian estimation would be computationally prohibitive.

Similarly, the sequential estimation in the Kalman filter procedure is also naturally considered from a sequential Bayesian perspective (e.g., Talagrand, 1997; Wikle & Berliner, 2007). That is, the forecast distribution for time $t + 1$ can be considered as the “prior” distribution for updating the latent state for time $t + 1$ when new data become available for that time. For linear, normal data and process models, these update equations are available in closed form and are equivalent to the variational approach, but when these distributions are non-normal or the dynamics are nonlinear, then sequential Monte Carlo approaches such as particle filtering or ensemble Kalman filtering must be used (e.g., see the tutorial in Katzfuss et al., 2016).

One of the limitations of considering the equivalence of the variational and Bayesian perspective is that it assumes the model parameters (and error variances/covariances) are known (or they are somehow estimated off-line) and that one’s models are correct. In the case of online sequential filtering, one can augment the state vector with parameters and update them along with the state process. Alternatively, in variational approaches it is certainly possible to add another term to the objective function that accounts for parameter uncertainty. This is equivalent to assigning a prior distribution to the parameters. When practical, it is often more flexible to specify prior distributions explicitly in case non-normal priors are warranted. The development of mechanistically-motivated spatio-temporal dynamic models (described in the next section) evolved from this hierarchical Bayesian perspective.

3.2 Statistical Approaches for Adding Mechanistic Information

There have been several approaches to combine mechanistic information in statistical spatial and spatio-temporal models. For example, the development of spatial and spatio-temporal covariance functions from analytical or numerical solutions to PDEs has long been considered an efficient way of incorporating mechanistic information in spatial and spatio-temporal models. PDE-derived covariance functions have been used for spatial prediction, such as for kriging models (e.g., Heine, 1955; Whittle, 1954), and have recently become quite popular with the stochastic-PDE (SPDE) approach (e.g., Lindgren et al., 2011, 2022). Indeed, as mentioned in Section 3.1.3, this idea of using PDE-derived covariance functions for optimal interpolation has also been used in the early data assimilation literature (e.g., Daley,

1993). Generally, this approach is useful for developing covariance functions, but is quite limited in the complexity of the underlying mechanistic model that can be effectively considered (see the overview in [Cressie & Wikle, 2011](#)).

As an alternative, in the functional data and non-parametric regression analysis literature, it has long been common to use function derivatives to regularize a regression model such as in smoothing splines for spatial data (e.g., [Wahba, 1990](#); [Eubank, 1999](#); [Ramsay, 2002](#)). Over the last 10 years or so, this has been extended to the physics-informed regularization case in spatial and spatio-temporal regression (e.g., see the review in [Sangalli, 2021](#)). These methods are very much in the spirit of the [Sasaki \(1958\)](#) approach described in Section 3.1, where one adds a PDE regularization term to the loss function that compares the non-parametric model of interest to the data. As described in [Tomasetto et al. \(2024\)](#), point estimates of parameters can be estimated in this framework. These approaches can be quite effective for certain problems, and in fundamental ways the Physics-Informed Neural Network (PINN) approach described below in Section 4.1 can be thought of as a generalization of this approach with more flexible semi-parametric (deep neural) models and the use of automatic differentiation to allow more complex PDE regularization. Bayesian-PINNs (described in Section 4.2), mechanistically-motivated statistical models (described in Section 3.2.1), and probabilistic numerical methods (described in Section 3.2.2) operate similarly, but can accommodate more principled uncertainty quantification across all phases of model development via a hierarchical Bayesian framework. Given that uncertainty quantification is a major focus for modern spatio-temporal modeling applications, we describe the statistical approaches in more detail in the following subsections, and PINNs and B-PINNs in Section 4.

3.2.1 Mechanistically-Motivated (Physical-Statistical) Models

Before introducing the mechanistically-motivated framework, we first briefly review Bayesian hierarchical models for spatio-temporal processes. Adding additional sources of data and accounting for uncertainty in observation model parameters and state-process model parameters is often needed when modeling complex spatio-temporal data. In some cases, one can gain much more expressiveness in their models if they add dependence structure (time, space) to the parameters as well by considering them to be spatial or temporal random processes and/or to be related to time or space-varying covariates. Bayesian hierarchical models are natural extensions of the Bayesian paradigm that allow for more complete accommodation of uncertainty and this additional model flexibility.

The Bayesian hierarchical (deep) modeling paradigm for spatio-temporal data (e.g., [Berliner, 1996](#); [Cressie & Wikle, 2011](#)) considers breaking the joint distribution of data, process, and parameters into a series of conditional models (distributions). In that context, the posterior distribution is then given by the following general decomposition

$$\begin{aligned} [process, parameters \mid data] &\propto [data \mid process, data parameters] \\ &\times [process \mid process parameters] \\ &\times [parameters]. \end{aligned}$$

Importantly, each of these stages can have sub-stages as described below in Section 3.2.1.

In the context of the dynamic spatio-temporal models (DSTMs) of interest here, the basic BHM might take the following form (e.g., see [Cressie & Wikle, 2011](#)).

$$\begin{aligned}
 \text{Data Model: } z_t &= \mathbf{h}_t(\mathbf{u}_t) + \epsilon_t, \quad \epsilon_t \stackrel{\text{ind.}}{\sim} \mathbf{N}(\mathbf{0}, \mathbf{R}_t(\boldsymbol{\theta}_z)), \\
 \text{Process (Dynamic) Model: } \mathbf{u}_t &= \mathcal{M}(\mathbf{u}_{t-1}, \mathbf{B}(\mathbf{u}_t; \boldsymbol{\theta}_b); \boldsymbol{\theta}_m) + \boldsymbol{\eta}_t, \quad \boldsymbol{\eta}_t \stackrel{\text{ind.}}{\sim} \mathbf{N}(\mathbf{0}, \mathbf{Q}_t(\boldsymbol{\theta}_q)), \\
 \text{Process Initial Value: } \mathbf{u}_0 &\sim \mathbf{N}(\mathbf{a}, \boldsymbol{\Sigma}_0) \\
 \text{Parameter Models: } &[\boldsymbol{\theta}_z, \boldsymbol{\theta}_m, \mathbf{a}, \boldsymbol{\Sigma}_0, \boldsymbol{\theta}_b, \boldsymbol{\theta}_q],
 \end{aligned}$$

where $\mathbf{h}_t(\cdot)$ is a mapping function between the latent process and observations, $\mathbf{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is a multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, and $\mathbf{B}(\mathbf{u}_t; \boldsymbol{\theta}_b)$ represents a boundary condition and $\mathbf{u}_0$ is the initial condition, both of which depend on parameters that are given prior distributions in general. The unknown parameters are given by $\boldsymbol{\theta}_z, \boldsymbol{\theta}_m, \mathbf{a}, \boldsymbol{\Sigma}_0, \boldsymbol{\theta}_b, \boldsymbol{\theta}_q$. As mentioned above, the parameter models can also have dependence (in space and/or time) or depend on other covariates, which gives this model a great amount of generalizability and can make it much “deeper” in terms of having more levels of conditioning (e.g., see *Process* description below).

Including Mechanistic Information: The BHM framework naturally allows one to include mechanistic information in each model layer. Here, we briefly describe how mechanistic information can be incorporated into the data, process, and parameter levels.

Data: A strength of the BHM approach is that one can easily specify multiple data sources conditioned on the latent state process, e.g., $[Z_1, Z_2|u] = [Z_1|u][Z_2|u]$ (assuming conditional, although this assumption can be relaxed). In statistics, a common “*data fusion*” approach since the late 1990s has been to combine *in situ* data, such as satellite observations (Z_1) and mechanistic model outputs such as weather reanalysis fields (Z_2), to incorporate mechanistic information (e.g., [Berliner et al., 1999](#); [Wikle et al., 2001](#); [Berliner, 2003](#); [Fuentes & Raftery, 2005](#); [Wikle & Berliner, 2005](#)). This is a type of weak constraint mechanistic modeling that has the advantage of being able to account for additive errors (to facilitate uncertainty quantification).

Process: As with the data assimilation literature, one can specify $\mathcal{M}(\cdot)$ to be a known mechanistic model. Alternatively, in the physical-statistical (mechanistically-motivated) DSTM paradigm, one can specify statistical models for the transition operator, say $\mathcal{M}(\cdot; \boldsymbol{\theta}_m)$, that are “motivated” by mechanistic processes, but are conditioned on parameters $\boldsymbol{\theta}_m$ that typically have spatial or temporal structure (e.g. spatially-varying diffusion parameters); see [Wikle & Hooten \(2010\)](#) and [Cressie & Wikle \(2011\)](#) for numerous examples. For example, one may discretize (or write in spectral/finite-element space) a reaction-diffusion PDE or integro-difference equation and condition this on the diffusion/growth parameters that may vary with space as a GP, or in terms of covariates, which are specified at the next level of the hierarchy (e.g., [Wikle, 2003](#); [Xu et al., 2005](#); [Wikle & Berliner, 2007](#); [Hooten & Wikle, 2008](#)).

Parameters: In some cases, mechanistic information can be imparted to the parameters (say, $\boldsymbol{\theta}_m$) through the prior specification; e.g., via regularization or informed by mechanistic information. For example, [Wikle \(2003\)](#) allowed covariates and a latent Gaussian process to model the spatially-varying diffusion rate in reaction-diffusion model for

spread of an invasive species and [Wikle et al. \(2001\)](#) included a known turbulent scaling relationship as a prior on the variances of tropical surface winds to add realistic structure. Alternatively, [Leeds et al. \(2014\)](#) used pre-training/meta-learning of prior distributions with model output obtained from a coupled ocean-ecosystem model to inform a quadratic nonlinear mechanistically-motivated statistical model for data assimilation of ocean color measurements.

As a concrete example illustrating the process and parameter stages of a mechanistically motivated DSTM, consider the latent dynamic process v in one-dimensional space, which is assumed to approximately follow the advection–diffusion equation

$$\frac{\partial v}{\partial t} = a \frac{\partial v}{\partial s} + \lambda \frac{\partial^2 v}{\partial s^2}. \quad (7)$$

As shown in [Wikle & Berliner \(2007\)](#), a simple finite difference numerical solution suggests the following difference equation

$$v_t(s) = \theta_1 v_{t-\Delta_t}(s) + \theta_2 v_{t-\Delta_t}(s + \Delta_s) + \theta_3 v_{t-\Delta_t}(s - \Delta_s),$$

where the coefficients θ_i are functions of the spatial discretization, Δ_s , time discretization, Δ_t , and advection and diffusion parameters, a and λ , respectively. Denoting the discretized process at n spatial locations by the vector $\mathbf{v}_t = [v_t(s_1), \dots, v_t(s_n)]'$, and the boundary process by $\mathbf{v}_t^b = [v_t(s_0), v_t(s_{n+1})]'$, we can write the vectorized difference equation for the solution of the PDE as $\mathbf{v}_t = \mathbf{M} \mathbf{v}_{t-\Delta_t} + \mathbf{M}^b \mathbf{v}_{t-\Delta_t}^b$, where $\mathbf{M}$ and $\mathbf{M}^b$ are functions of $\theta_0, \theta_1, \theta_2$; e.g., $\mathbf{M}$ is a tridiagonal matrix (e.g., with θ_0 on the main diagonal and θ_1, θ_3 on the off-diagonals; e.g., see [Wikle & Berliner \(2007\)](#)). Now, a mechanistically motivated model for the latent process of interest, say $\mathbf{u}_t$, could be written as $\mathbf{u}_t = \mathbf{M}_u \mathbf{u}_{t-1} + \boldsymbol{\eta}_t$, where $\mathbf{M}_u$ might be a tridiagonal matrix, but where the main diagonal is spatially varying, e.g., $\boldsymbol{\theta}_0 \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Sigma}_\theta)$, where $\boldsymbol{\theta}_0$ is an $n \times 1$ vector and $\mathbf{X}$ represents covariates and $\boldsymbol{\Sigma}_\theta$ imparts spatial dependence. With a data model that relates observations to the latent process, $\mathbf{u}_t$, this corresponds to a simple version of the BHM DSTM described above. The aforementioned references contain specific examples and extensions of such a mechanistically motivated model. For example, if desired, one can include hard constraints (i.e., “mass balance”) in the physical-statistical framework (e.g. [Berliner et al., 2008](#); [Zhuang & Cressie, 2014](#); [Gopalan et al., 2018](#)).

3.2.2 Probabilistic Numerical Methods

Over the last decade, there has been growing interest in modernizing the ideas associated with treating differential equation solutions as random processes (e.g., [Larkin, 1972](#); [Diaconis, 1988](#)). These ideas, which come under a broad set of methods called *probabilistic numerical methods* (PNMs), have provided a comprehensive toolkit to consider uncertainty quantification in the solution of differential equations (e.g., see [Hennig et al., 2015](#); [Chkrebtii et al., 2016](#); [Cockayne et al., 2019](#), and references therein). These methods have the advantage of being rigorous in their convergence guarantees and are straightforward to implement.

Many PNM methods for solving PDEs with uncertainty assume the solution has a GP prior. Then observations of boundary conditions and forcing terms of the PDE at a finite set of (collocation) points act as data that allow for a posterior distribution over the PDE solution. For a simple illustration, in the space-only case, we might have the

following GP prior:

$$u_t(s) \sim GP(m(s), k(s, s'; \theta_k)),$$

where $m(s)$ is a mean function and $k(s, s'; \theta_k)$ is a covariance function between two spatial locations s and s' , defined by parameters θ_k . Thus, this prior accommodates belief about the inherent smoothness of the PDE solution. In the case of a linear map (say, $\mathcal{M}(u(s)) = f(s)$), the implied GP prior for the PDE solution is

$$\mathcal{M}(u(s)) \sim GP(\mathcal{M}(m(s)), \mathcal{M}\mathcal{M}'(k(s, s'; \theta_k))),$$

where $\mathcal{M}\mathcal{M}'(k(s, s'; \theta_k))$ is the quadratic form associated with the linear operation on the prior covariance function (i.e., the kernel derived based on the fact that $\mathbf{A}\Sigma\mathbf{A}^T$ is the variance-covariance matrix of random vector $\mathbf{A}\mathbf{x}$ where the variance-covariance matrix of $\mathbf{x}$ is Σ). One then treats forcing terms as data, e.g.,

$$f(s_i) = \mathcal{M}(u(s_i)) + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma_\epsilon^2),$$

where σ_ϵ^2 controls the fidelity of the PDE solution. Given the prior and data models are normal, the posterior distribution of $u(s)$ given $f(s_1), \dots, f(s_m)$ are available in closed form via a normal distribution. In practice, one includes a boundary condition process and, for time-varying PDEs, an initial condition. In addition, one typically chooses the collocation points at which to obtain “observations” $f(s_i)$ to facilitate computation. This approach is often embedded in a BHM with real world observations to learn the parameters of the underlying PDE (i.e., the inverse problem).

The strength of the PNM approach is its principled uncertainty quantification, which comes from the numerical approximation (e.g., the $f(s_i)$ observations) and, in the case of inverse problems, real-world observations. The inherent regularization of the Bayesian prior also provides some robustness to ill-posedness. Perhaps most importantly, PNM’s strong connections to traditional numerical methods have been established for special cases (e.g., linear PDEs and GP priors). That said, these methods are much more difficult to implement for nonlinear PDEs, do not scale well for very high-dimensional problems, and have not been used as much in deep BHMs with multiple sources of process uncertainty, as in the physical-statistical paradigm. However, the Bayesian PINN framework described in the next section might be considered a neural network based PNM.

4 Physics-Informed Neural Networks (PINNs)

In the spirit of the Sasaki’s variational approach described above in Section 3.1, consider fitting a deep neural network (NN) but constraining the solution to follow a PDE. This is the essence of PINNs (Raissi et al., 2017a,b; Raissi, 2018; Raissi et al., 2019, 2020). Such approaches can be used in the context of fitting a PDE in the presence of data (as in data assimilation), or solving a PDE explicitly (presumably for the sake of efficiency). We are interested here in the situation where we have data corresponding to the latent process of interest, initial conditions, or boundary conditions. In this section, we provide a brief overview of the PINN approach and its Bayesian counterpart (B-PINN).

4.1 PINN Formulation

PINNs, first proposed by Raissi et al. (2017a,b), mainly consist of two components: (1) a neural network for prediction and (2) a physically-inspired equation (e.g., PDE) to impose constraints in the loss function. We denote the deep neural network solution to the PDE at time t and location $\mathbf{s}$ as $\mathbf{v}_{NN,t}(\mathbf{s}; \boldsymbol{\theta}_W)$, where $\boldsymbol{\theta}_W$ represents the neural network parameters (weights and biases). Note, $\mathbf{v}_{NN,t}$ is an approximation of the PDE solution $\mathbf{v}_t$. Indeed, since this is not a perfect representation it is best to make the distinction as follows by defining a PDE “residual” (as in the PNM approach described in Section 3.2.2). Let $\boldsymbol{\theta}_m$ correspond to PDE parameters as in (1), and write

$$\mathbf{r}_t(\mathbf{s}; \boldsymbol{\theta}_W, \boldsymbol{\theta}_m) \equiv \frac{\partial \mathbf{v}_{NN}(\boldsymbol{\theta}_W)}{\partial t} - \mathcal{M}(\mathbf{v}_{NN}(\boldsymbol{\theta}_W); \boldsymbol{\theta}_m), \quad (8)$$

for the “residuals” of the PDE evaluated at time t and location $\mathbf{s}$ (note, in some cases where there is an exogenous forcing term in the PDE, this “residual” could be thought of as a representation of the forcing). In practice, as in the PNM approach, one considers N_c so-called *collocation points* at which to evaluate the PDE, say $\{\mathbf{s}_c^i, t_c^i\}_{i=1}^{N_c}$. Note, there can be different numbers of spatial locations at different times, so the N_c collocation points include the complete set of space/time locations to train the neural net on the PDE (although we typically are interested in the solution at other locations in time and space). The loss function associated with the PDE is then given by the mean squared error (MSE),

$$J_P(\boldsymbol{\theta}_W) = \frac{1}{N_c} \sum_{i=1}^{N_c} \|\mathbf{r}_{t_c^i}(\mathbf{s}_c^i; \boldsymbol{\theta}_W)\|_2^2, \quad (9)$$

where $\|\cdot\|_2$ represents the ℓ_2 norm and, for now, we are assuming the PDE parameters $\boldsymbol{\theta}_m$ are known, and so exclude them notationally. We typically would also include similar terms for the initial and boundary conditions, i.e., where there are N_{IC} initial condition collocation points and N_{BC} boundary condition collocation points. Denote these initial and boundary collocation points as $\{\mathbf{s}_{IC}^i\}_{i=1}^{N_{IC}}$ and $\{\mathbf{s}_{BC}^i, t_{BC}^i\}_{i=1}^{N_{BC}}$, respectively; then the corresponding loss functions can be written as:

$$J_{IC}(\boldsymbol{\theta}_W) = \frac{1}{N_{IC}} \sum_{i=1}^{N_{IC}} \|\mathbf{v}_{NN,0}(\mathbf{s}_{IC}^i; \boldsymbol{\theta}_W) - \mathbf{g}_I(\mathbf{s}_{IC}^i)\|_2^2,$$

$$J_{BC}(\boldsymbol{\theta}_W) = \frac{1}{N_{BC}} \sum_{i=1}^{N_{BC}} \|\mathbf{B}(\mathbf{v}_{NN,t_{BC}^i}(\mathbf{s}_{BC}^i; \boldsymbol{\theta}_W))\|_2^2.$$

If one is simply interested in finding the solution of the PDE given the initial and boundary conditions and PDE parameters $\boldsymbol{\theta}_m$, then one finds the $\boldsymbol{\theta}_W$ that minimize the combined objective function,

$$J(\boldsymbol{\theta}_W) = a_{IC} J_{IC}(\boldsymbol{\theta}_W) + a_{BC} J_{BC}(\boldsymbol{\theta}_W) + a_P J_P(\boldsymbol{\theta}_W),$$

where the specified non-negative constants $\{a_{IC}, a_{BC}, a_P\}$ allow one to focus more on different components of the solution if desired. The solution in this case does not require measurements of the process besides initial and boundary condition training data, so the PINN serves as an alternative to the traditional numerical solver (e.g., finite difference, finite element or spectral scheme). In this context, note a well-known limitation of the PINN approach is it can have difficulty in simulating multi-scale, chaotic, or turbulent behavior because the explicit forward (“causal”) nature of

time is not considered (e.g., [Wang et al., 2022](#)). This limitation can be mediated in a number of ways, for example, by adding a time weighting function in the loss function (9). Regardless, some recent comparison studies have shown that traditional finite element solutions of PDEs work better and are as or more efficient than PINNs for most problems (e.g., [Grossmann et al., 2024](#)).

4.1.1 PINNs with Observational Data

As mentioned previously, in most cases in statistical applications of mechanistic-informed models with neural networks, our interest is to fit a deep neural network model with PDE constraints to data, which in the PINN literature is considered to be “data-driven discovery” of the PDE, by which they mean to learn the PDE parameters, θ_m (i.e., an inverse problem). In this case, a data component is added to the objective function as in the variational approach and the PDE parameters can be estimated jointly with neural network parameters by minimizing the associated loss; the loss function J_P in (9) is a function of θ_m in addition to θ_W due to the dependence of $r_t(\cdot)$ on θ_m . Moreover, additional data observations of the process can be used; i.e., given observations, $\{z_{ti}(\mathbf{s}^i) : i = 1, \dots, N\}$,

$$J_D(\theta_W) = \frac{1}{N} \sum_{i=1}^N \|z_{ti}(\mathbf{s}^i) - v_{NN,ti}(\mathbf{s}^i; \theta_W)\|_2^2. \quad (10)$$

Note, in the original presentation of data-driven discovery, there is no measurement error associated with the observation in (10). The solution proceeds by optimizing the weighted combination

$$J(\theta_W, \theta_m) = a_{IC}J_{IC}(\theta_W) + a_{BC}J_{BC}(\theta_W) + a_P J_P(\theta_W, \theta_m) + a_D J_D(\theta_W),$$

where we now explicitly note that J_P depends on the PDE parameters θ_m . After this optimization, one has learned the PDE parameters and the associated neural network can be used to obtain an estimate of the latent process of interest, $u_t(\mathbf{s})$, at desired locations in space and time.

Importantly, “data-driven discovery of PDEs” does not mean finding a functional form of PDEs in this case, but rather indicates estimating parameters within PDEs. The use of statistical models and deep neural networks to learn the functional form of mechanistic models is a growing area of research, but beyond the scope of this article. For recent overviews, see [North et al. \(2023\)](#), [Sahimi \(2024\)](#), and [North et al. \(2025\)](#).

4.2 Bayesian PINNs

The PINN approach described above does not explicitly account for uncertainty in the data, model, or parameters. It is quite straightforward to account for measurement error as in the variational approach presented earlier; we can modify the data portion of the loss function (10) by assuming that the measurement errors are normal. For example, we might assume that we have observation models of a univariate process (although this can be generalized to a multivariate process) $z_{ti}(\mathbf{s}^i) = v_{NN,ti}(\mathbf{s}^i; \theta_W) + \epsilon_{ti}(\mathbf{s}^i)$, where $\epsilon_{ti}(\mathbf{s}^i) \sim N(0, \sigma_z^2)$. Similarly, in principle we could have observations of $r_t(\mathbf{s})$ (the PDE residual term), boundary condition $B(\cdot)$ and initial conditions $g_I(\cdot)$ that also have independent normal errors and associated error variances. All these distributions are conditioned on the solution of the PDE and PDE model parameters, which in the PINN context are completely determined given the NN and PDE

model parameters, θ_W, θ_m . Note, in principle, one could have dependent errors as described in the variational and physical-statistical approaches above.

It is then natural, following from the Bayesian NN literature (e.g., [Neal, 2012](#)), to specify prior distributions on the NN weight and bias parameters, θ_W ; known as a B-PINN (e.g., [Yang et al., 2021](#)). That is, assume we generically have conditional data distributions for the process, PDE residual, boundary and initial conditions given by $[D_z|\sigma_z^2, \theta_W]$, $[D_r|\sigma_r^2, \theta_W, \theta_m]$, $[D_{bc}|\sigma_{bc}^2, \theta_W]$ and $[D_{ic}|\sigma_{ic}^2, \theta_W]$, respectively. Note, we use D_z to represent the observations $\{z_{t^i}(\mathbf{s}^i)\}$ described above to provide a more general and consistent notation for the observations. Each of these terms is implicitly conditioned on $v_{NN,t}$, but as mentioned above, since the NN is completely determined conditioned on its weights and biases, it is not necessary to write this explicitly. In addition, as in the case of PNMs, unless we have a forcing term in the PDE, we assume implicitly that “observations” of the residuals are equal to zero on average – that is, conditioned on the parameters, the relationship in (1) holds perfectly up to some arbitrary white noise (represented by σ_r^2). This has the effect of including a term equivalent to (9) in the un-normalized posterior representation (12) below. That is, we can either write this as a “pseudo data model” with known zero observations or as a regularization prior term. For the latter, we might write the prior distribution as:

$$\frac{\partial v_{NN}(\theta_W)}{\partial t} \sim \mathcal{N}(\mathcal{M}(v_{NN}(\theta_W); \theta_m), \sigma_r^2), \quad (11)$$

which corresponds to a regularization term and discrepancy parameter, σ_r^2 , that effectively specifies *a priori* belief in the representativeness of the PDE for this process.

Assuming conditional independence of the different data types, one can express the conditional data distribution as

$$[D_z, D_r, D_{bc}, D_{ic}|\theta_W, \theta_m, \theta_D] = [D_z|\sigma_z^2, \theta_W][D_r|\sigma_r^2, \theta_W, \theta_m][D_{bc}|\sigma_{bc}^2, \theta_W][D_{ic}|\sigma_{ic}^2, \theta_W],$$

where $\theta_D \equiv (\sigma_z^2, \sigma_r^2, \sigma_{bc}^2, \sigma_{ic}^2)'$. The parameter θ_m is only included for the conditional distribution of D_r because PDE residuals are the only quantities that depend on the physics parameters; the other quantities’ conditional distributions are only dependent on the neural network, and consequently only θ_W . Now, assume we have prior distributions $[\theta_W]$, $[\theta_m]$, and $[\theta_D]$. Then, Bayes’ rule gives

$$[\theta_W, \theta_m, \theta_D|D] \propto [D|\theta_W, \theta_m, \theta_D][\theta_W][\theta_m][\theta_D], \quad (12)$$

where $D \equiv \{D_z, D_r, D_{bc}, D_{ic}\}$. The intractable posterior (12) is typically sampled from using MCMC methods (e.g., Hamiltonian Monte Carlo, Langevin Dynamics), variational methods, or ensemble methods (see the comprehensive review in [Psaros et al., 2023](#)).

A significant challenge with Bayesian NNs (and, hence, B-PINNs) is the specification of reasonable prior distributions for the NN parameters θ_W . In general, such parameters are not identifiable, which can make MCMC sampling challenging without extra constraints (e.g., [McDermott & Wikle, 2019](#)), the specification of informative priors (e.g., through empirical Bayes or meta-learning; e.g., [Meng et al. \(2022\)](#)), or special classes of priors (e.g., [Meng et al., 2022](#); [Sell & Singh, 2023](#); [Castillo & Egels, 2024](#)).

Note that the B-PINN framework presented here does not include an explicit term for model error, although σ_r^2 does account for minor discrepancies in the PDE. But, this does not respect the fact that the PDE dynamics are not usually sufficient to model the real world process. In other words, the real-world dynamical process is latent and is motivated by the PDE, but is typically more complicated. This is exactly the motivation for denoting a distinction between the true latent process, u , and the PDE solution v in Section 2.2 and the NN solution v_{NN} here. Such discrepancy terms are common in the computer model experiment literature (e.g., Gramacy, 2020). In the context of B-PINNs, a discrepancy term can be added in the data model, e.g., $z_{t^i}(s^i) = v_{NN,t^i}(s^i; \theta_W) + \delta_{t^i}(s^i; \theta_\delta) + \epsilon_{t^i}(s^i)$, where $\delta_{t^i}(s^i; \theta_\delta)$ is a discrepancy term that depends on parameters θ_δ . In the statistical surrogate and mechanistically-motivated statistics literature, this discrepancy is typically modeled as a GP, a functional (basis) expansion, or a machine learning surrogate. In the B-PINN context, it is natural to model the discrepancy as an additional (separate) neural network (e.g., Zou et al., 2024), with the discrepancy used to correct for the misspecification of the PDE when observing PDE residuals. In either case, one must be careful to ensure that there is sufficient prior information regarding the spatial (and possibly temporal) scale of the discrepancy process to ensure identifiability of the solution (see our example in Section 6 below for more discussion).

Although the B-PINN framework is an increasingly useful paradigm to account for uncertainty in PINN applications, there are still potential limitations in its lack of an explicit hierarchical formulation, lack of consideration of covariate effects, lack of accommodation of dependent random effects, focus on normal data distributions, and lack of more complicated measurement/sampling errors. In the next section, we show how the B-PINN fits into a more general deep BHM framework for dynamic spatio-temporal processes that can accommodate these concerns.

5 Bayesian Hierarchical Model PINNs (BHM-PINNs)

A strength of an overview is drawing generalizations and connections to a larger body of work. As part of this overview of PINNs, we show here that the B-PINN (Yang et al., 2021) fits naturally into the BHM framework that has long been used to model dynamic spatio-temporal processes in statistics (e.g., see Cressie & Wike, 2011, for a review). It is important to note that this is not a new methodology, but simply a new use case of the spatio-temporal BHM framework. We denote this a *BHM-PINN* and show it is a flexible approach to bring mechanistic-inspired PDE dynamics to spatio-temporal modeling in a paradigm that can account for non-normal data distributions, covariate effects, and residual spatial or spatio-temporal processes. This is analogous to how mechanistic-informed models are included in physical-statistical BHMs. Thus, the BHM-PINN described in the following subsections generalizes both the traditional BHM spatio-temporal physical-statistical modeling framework and the PINN modeling framework.

5.1 Data Models

Assume we have up to p different datasets at time t given by $\{z_{1t}, \dots, z_{pt}\}$ that are all thought to be informative about our latent process of interest, u_t . As is common in spatio-temporal statistics, we assume these datasets are conditionally independent given the latent spatio-temporal process. Specifically, $z_{jt}|u_t, \theta_{zj} \sim \mathcal{D}(h_j(u_t); \theta_{zj})$, where for $j = 1, \dots, p$, $\mathcal{D}(h_j(u_t); \theta_{zj})$ is some data distribution (e.g., normal, Poisson, log-normal, etc.) that is conditioned

on the latent spatio-temporal process, $\mathbf{u}_t$, and parameters θ_{zj} . Here, $h_j(\cdot)$ is a mapping function between the latent process and the data that can account for missing observations, nonlinear transformations, and change-of-support (spatial resolution) (for more background see [Cressie & Wike, 2011](#)). For instance, in the example of Section 6, the function $h(\cdot)$ implicitly is a composition of an incidence matrix (to account for missing data locations) and exponentiation (to provide a link function for the Poisson distribution, which requires a positive rate parameter); in other words, this mapping can accommodate a link function in a generalized linear model context. As discussed in the context of physical-statistical models in Section 3.2.1, data models can accommodate multiple observational data sources as well as mechanistic model output. In addition, as discussed in the context of B-PINNs in Section 4.2, these data models can accommodate “pseudo-data” for PDE residuals, boundary conditions, and initial conditions.

5.2 Process Models

We let the latent process part of the model be decomposed into three terms; an overall time-varying spatial mean, a PINN dynamic process, and a residual (discrepancy) process, respectively as

$$\mathbf{u}_t = \boldsymbol{\mu}_t + \mathbf{v}_{NN,t} + \boldsymbol{\delta}_t, \quad (13)$$

where we are interested in modeling the latent process at a fixed set of n locations given by the $n \times 1$ vector $\mathbf{u}_t$. Then, $\boldsymbol{\mu}_t$ is the n -dimensional time-varying spatial mean, $\mathbf{v}_{NN,t}$ is the PINN at the n locations of interest at time t , and $\boldsymbol{\delta}_t$ is the discrepancy term. Each of these terms can be specified in multiple ways as described below.

5.2.1 Mean

The process mean $\boldsymbol{\mu}_t$ can accommodate covariates or other “fixed effects”; e.g., $\boldsymbol{\mu}_t = \mathbf{X}_t \boldsymbol{\beta}_t$, where $\mathbf{X}_t$ is an $n \times q$ covariate matrix with potentially time-varying covariates, and $\boldsymbol{\beta}_t$ is a q -dimensional vector (that may be time varying). In this case, we would specify a prior distribution $[\boldsymbol{\beta}_t]$ (which depends on the application; e.g., see the simulation example in the next section). We use the term “covariates” quite loosely here, and note that $\mathbf{X}_t$ can accommodate seasonality, temporal trends, one hot encoding, exogenous covariate effects, etc. In some cases, one may not need this term in the model, or it might not depend on time (e.g., it could represent a simple constant offset).

5.2.2 PINN Dynamic Process

The dynamic component of the model is the most important as it represents the underlying spatio-temporal mechanism driving the process. This is the main place where we impart mechanistic information via the PINN as described in Section 4.1 or B-PINN as in Section 4.2, denoted as $\mathbf{v}_{NN,t}$. Note that we are not explicitly writing out the initial and boundary conditions, but those can be incorporated in the PINN as described previously. Recall, the PDE constraint can be incorporated into “observations” of the PDE residuals (8) as before (e.g., we can include a data model for these residuals and specify the observations to be zero unless there is a known PDE forcing term) or we can specify a prior on the time derivative of the PDE. This then accommodates the PDE dynamical structure and we note there are additional θ_m parameters that come in the residual process from the PDE. As with the B-PINN framework, the residual variance term (σ_r^2) allows one to control the degree to which the neural network is constrained to follow the underlying PDE.

5.2.3 Discrepancy Process

The discrepancy process term δ_t is important because it respects the fact that our latent dynamical model represented by the process $v_{NN,t}$ surely does not completely account for all of the latent spatio-temporal dependence in real-world data (i.e., “all models are wrong”). This term could be accommodated by another NN, say $\delta_t = \delta_{NN,t}(\mathcal{S}; \theta_\delta)$ as in [Zou et al. \(2024\)](#). Alternatively, it could be a time-varying stochastic process, e.g., $\delta_t = f_\delta(\delta_{t-1}; \theta_\delta)$ such as the residual autoregressive process to accommodate small scale dynamics in [Wikle et al. \(2001\)](#), or simply a time-independent spatial GP. More generally, this term could be represented by any spatio-temporal random effect distribution, say $\delta_t \sim [\theta_\delta]$.

5.3 Parameter Models

Depending on the models selected in the data and process stages, one can have a variety of parameters that require prior distributions, e.g., $[\beta_t | \theta_\beta][\theta_w][\theta_\delta]$, etc. Typically, we require regularization for these parameters as there are likely to be quite a lot of them, or as mentioned in the context of physical-statistical models, we might have mechanistic information for processes (e.g., parameters must obey a turbulent scaling relationship as in [Wikle et al. \(2001\)](#)). In general, many of the prior specifications that have been used for mechanistically motivated BHM models (for some examples, see [Wikle \(2003\)](#); [Pagendam et al. \(2014\)](#); [Zammit-Mangion et al. \(2014\)](#)) carry over to BHM-PINNs because the BHM-PINN takes the same form at the parameter level. In the B-PINN context, [Zhang et al. \(2025\)](#) consider a hierarchical structure on parameters for a gear-grinding physics model in which correlation is allowed between parameters (we note, however, that this hierarchical framework is solely for the parameters of the physics model, distinct from the BHM-PINN approach).

Deep BHMs as described here have multiple hyperparameters. These correspond to specifications of the parameters for the priors in the previous stages as well as model architecture (e.g., number of hidden units, levels, etc.) in the case of B-PINNs. Depending on the application, these hyperparameters can be given vague proper prior distributions or are fixed (i.e., point mass priors) and the sensitivity to the choice is examined.

5.4 Implementation Considerations

As with many mathematical and statistical models, model identifiability must be addressed when implementing a BHM-PINN, specifically with regard to the NN and discrepancy terms. For example, in (13), if δ_t is parameterized using a GP, there is potential for confounding between this term and the PINN term, $v_{NN,t}$. One way to mitigate this issue is to restrict the spatial (temporal, or spatio-temporal) range of the GP to operate on a different scale than the dynamics imposed on $v_{NN,t}$. This can be accomplished by an informative prior distribution on the argument(s) of the kernel function for the GP.

All the implementation challenges for B-PINNs are present for BHM-PINNs. For example, it is not clear how to select the NN architecture in general, but see [Toscano et al. \(2025\)](#) for a review of various NN architectures that have been considered in the context of PINNs. Simple NNs, such as narrow 3-layer feed-forward NNs, are often used to fit B-PINNs. Not only is this more computationally tractable than a more complex model, it is also related to the

bias-variance trade-off, where wide and deep NNs can be over-parameterized and have the potential to overfit. Empirical results often suggest narrow and shallow feed-forward NNs result in the best performance, although this may not be the case for complex dynamics.

The primary computational tool that has been used for posterior inference in Bayesian hierarchical models is MCMC, such as Gibbs sampling (and variants like Metropolis-within-Gibbs sampling) (Brooks et al., 2011). Modern gradient-based methods, such as Hamiltonian Monte Carlo (HMC; Neal et al., 2011) and extensions like the No-U-Turn Sampler (Hoffman & Gelman, 2014), have become prevalent in general for Bayesian inference. In the B-PINN literature (Yang et al., 2021), the two primary computational tools are HMC and variational inference (Blei et al., 2017). In the case of variational Bayes, the mean-field approximation (i.e., product of independent normals) for posterior neural network weights might not be realistic in the sense that correlations (which are likely to exist in neural network parameters) are neglected, although variational GPs (Tran et al., 2015) can help alleviate this issue. Yang et al. (2021) found gradient-based MCMC methods often result in the best performance for B-PINNs and for our example presented below in Section 6, we use this approach.

In the context of the discrepancy term, we note that GP models can be computationally difficult due to cubic complexity of inversion and determinant calculations. This is a well-researched topic in spatial statistics and there are several ways to mitigate this issue (e.g., see the review in Heaton et al., 2019).

6 Simulation Experiment: BHM-PINN

We consider a simulation experiment to illustrate a specific BHM-PINN. In particular, we consider a latent mechanistic dynamic process described by a one-dimensional (in space) Burgers' equation (Bateman, 1915; Burgers, 1948), which is a classic nonlinear PDE given by

$$\frac{\partial v(s, t)}{\partial t} + v(s, t) \frac{\partial v(s, t)}{\partial s} - \lambda \frac{\partial^2 v(s, t)}{\partial s^2} = 0, \quad (14)$$

where λ is the diffusion coefficient. The Burgers' equation is often used to illustrate dynamics because it is a very simple nonlinear advection-diffusion representation of a fluid dynamics problem that can model features like nonlinear wave propagation and shock formation while remaining relatively easy to analyze and solve. Here, we also include a fixed-effects mean term in the process representation, along with a Gaussian process discrepancy to reflect a situation where the true process is not just a known mechanistic model. We then assume this process corresponds to the latent mean of a Poisson count data generating model. The details of the simulation are given below.

1. **Domain:** We consider a one-dimensional equally spaced spatial grid with $n = 51$ locations in the domain $s \in [-\pi, \pi]$ and $T = 25$ equally distributed time points, $t \in [0, 5]$.
2. **Process Mean ($\mu(s, t)$):** We consider two covariates in the mean process, $\mathbf{x}(s, t) = (x_1(s, t), x_2(s, t))^T$, which are simulated from a zero-mean spatio-temporal GP with exponential covariance function $k_x((s, t), (s', t')) = 0.1 \cdot \exp\left(-\frac{\sqrt{(s-s')^2 + (t-t')^2}}{\ell_s}\right)$ where ℓ_s is set to $1/20$ the maximum observed

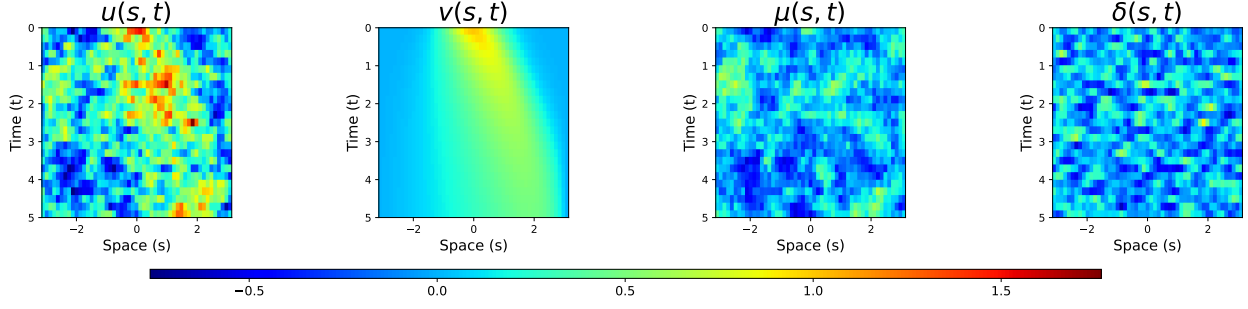
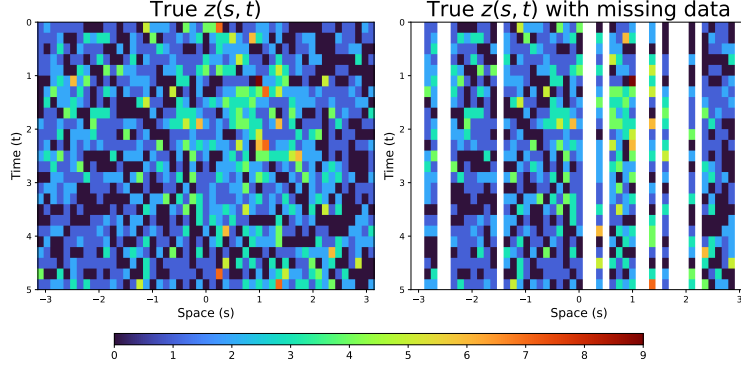


Figure 1: Simulated latent process $u(s, t)$ (left) and components over one-dimensional space and time: Burgers' equation $v(s, t)$ (second from left), mean process $\mu(s, t)$ (second from right), and discrepancy process $\delta(s, t)$ (right).

spatio-temporal distance across all location-time pairs. Specifically, we choose $\ell_s = 8.03/20$ and let $\beta = (0.6, -0.4)^\top$. This is shown in the third panel of Figure 1.

3. **Latent Dynamic Process ($v(s, t)$):** For each spatial and temporal location, we generated the PDE solution to Burgers' equation (14) numerically (with a spectral solver), denoted as $v(s, t)$. This numerical solution used initial condition $v(s, 0) = \exp(-s^2)$ for $s \in (-\pi, \pi)$ and 0 otherwise, with boundary conditions $v(-\pi, t) = v(\pi, t) = 0$. This is shown in the second panel of Figure 1.
4. **Discrepancy Process ($\delta(s, t)$):** For each time point a spatial discrepancy field was drawn as an independent, zero-mean Gaussian process, i.e., $\delta(s, t) \sim \text{GP}(0, k_\delta(s, s'))$, where $k_\delta(s, s') = \sigma_\delta^2 \exp\left(-\frac{(s-s')^2}{2\ell_\delta^2}\right)$ is a squared-exponential covariance function. Specifically, we select $\sigma_\delta^2 = 0.05$, and $\ell_\delta = 0.15$. This is shown in the fourth panel of Figure 1.
5. **Simulated Process ($u(s, t)$):** The simulated process at each spatial and temporal coordinate is then given as the sum of the mean, latent dynamic process, and discrepancy: $u(s, t) := \mathbf{x}(s, t)^\top \beta + v(s, t) + \delta(s, t)$. This is shown in the fourth panel of Figure 1.
6. **Simulated Observations ($z(s, t)$):** We then simulated observations from a random Poisson distribution, $z(s, t) \sim \text{Poisson}(\text{rate} = \exp(u(s, t)))$ (Figure 2 left panel). We assumed that observations at 30% of the space coordinates are missing at random for all times (Figure 2 right panel). Note, data missing over space for all times is a realistic scenario for simulating environmental processes with incomplete observation networks and can be more challenging to model than cases where data are missing completely at random in space and time.


 Figure 2: Left: Simulated z_t , Right: Missing data in z_t highlighted in white.

We consider the following BHM to fit these simulated data. First, the data, process, and discrepancy models are given by:

$$\begin{aligned}
 \text{Data Model : } z_t &\sim \text{Poisson}(\exp(\mathbf{H}_t \mathbf{u}_t)) \\
 \text{True Process : } u(s, t) &= \mathbf{x}(s, t)^\top \boldsymbol{\beta} + v_{NN}(s, t) + \delta(s, t) \\
 \text{Latent dynamics } (v_{NN}(s, t; \boldsymbol{\theta}_W) : &\frac{\partial v_{NN}(s, t)}{\partial t} \sim \mathcal{N}(v_{NN}(s, t) \frac{\partial v_{NN}(s, t)}{\partial s} - \lambda \frac{\partial^2 v_{NN}(s, t)}{\partial s^2}, \sigma_r^2) \\
 &v_{NN}(s, 0) \sim \mathcal{N}(\exp(-s^2), \sigma_{ic}^2) \quad (\text{Initial condition}) \\
 &v_{NN}(-\pi, t), v_{NN}(\pi, t) \sim \mathcal{N}(0, \sigma_{bc}^2) \quad (\text{Boundary condition}) \\
 \text{Discrepancy process : } \delta(s, t) &\sim \text{GP}(0, k_\delta(s, s')) \\
 &k_\delta(s, s') = \sigma_\delta^2 \exp\left(-\frac{(s - s')^2}{2\ell_\delta^2}\right).
 \end{aligned}$$

In this case, $\mathbf{H}_t$ is an incidence matrix that accommodates missing observations (e.g., see the discussion in [Wikle et al., 2019](#)). In addition, the following prior distributions are specified:

$$\begin{aligned}
 \boldsymbol{\beta} &\sim \mathcal{N}(\boldsymbol{\mu}_\beta, c_\beta \mathbf{I}), & \lambda &\sim \mathcal{N}(\mu_\lambda, \sigma_\lambda^2) \cdot \mathbf{I}(0 \leq \lambda < \infty) \\
 \ell_\delta &\sim \mathcal{N}(\mu_\ell, \sigma_\ell^2) \cdot \mathbf{I}(\ell_\ell \leq \ell_\nu < u_\ell), & \sigma_\delta^2 &\sim \mathcal{N}(\mu_\delta, \gamma_\delta^2) \cdot \mathbf{I}(\ell_\delta < \sigma_\delta^2 < u_\delta) \\
 \boldsymbol{\theta}_W &\sim \mathcal{N}(\boldsymbol{\mu}_W, c_W \mathbf{I}),
 \end{aligned}$$

where relatively informative priors are assigned for σ_δ^2 and ℓ_δ with $\mu_\delta = \mu_\ell = 0$, $\gamma_\delta^2 = \sigma_\ell^2 = 1$, and truncation bounds $\ell_\delta = 0.02$, $u_\delta = 0.1$, $\ell_\ell = 0.05$, and $u_\ell = 0.2$. These informative priors on the discrepancy parameters address the potential identifiability issue between the NN and discrepancy as well as the ability to recover Gaussian process covariance parameters in a fixed domain setting (an issue which has been frequently noted in the spatial statistics literature, e.g., [Gelfand et al. \(2010\)](#)). In contrast, we assign relatively non-informative priors to $\boldsymbol{\beta}$, λ , and the neural network weight and bias parameters $\boldsymbol{\theta}_W$, with $\boldsymbol{\mu}_\beta = \boldsymbol{\mu}_W = \mathbf{0}$, $c_W = \sigma_\lambda^2 = 1$, $\mu_\lambda = 0$, and $c_\beta = 10$. The neural network for $v_{NN,t}$ consists of $L = 3$ hidden layers, each with 16 hidden units.

As discussed in Section 4.2, the specification of the B-PINN model in this BHM uses “pseudo-data” for the PDE residuals, boundary conditions, and initial conditions. Specifically, we can write the following conditional data

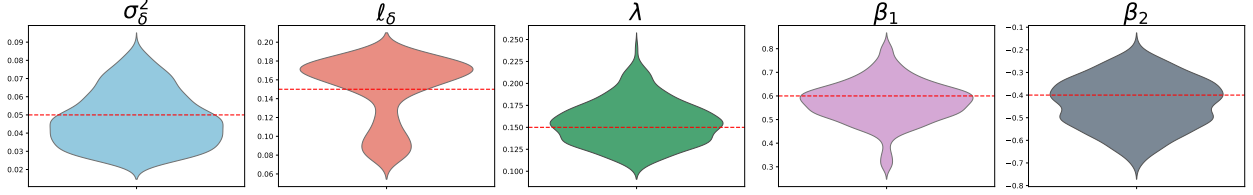


Figure 3: Violin plots of the BHM model parameter posterior distributions, with the true value indicated by a dashed red line. From left to right, the plots correspond to σ_δ^2 , ℓ_δ , λ , σ_z^2 , β_1 , and β_2 .

distributions that have the effect of updating the prior distributions for weights θ_W .

$$[D_r | \sigma_r^2, \theta_W, \lambda] = N(0; [\frac{\partial v_{NN}}{\partial t} + v_{NN} \frac{\partial v_{NN}}{\partial s} - \lambda \frac{\partial^2 v_{NN}}{\partial s^2}](s, t; \theta_W), \sigma_r^2), \quad (\text{Burgers' equation})$$

$$[D_{bc+} | \sigma_{bc}^2, \theta_W] = N(0; v_{NN}(\pi, t; \theta_W), \sigma_{bc}^2), \quad (\text{Right boundary condition})$$

$$[D_{bc-} | \sigma_{bc}^2, \theta_W] = N(0; v_{NN}(-\pi, t; \theta_W), \sigma_{bc}^2), \quad (\text{Left boundary condition})$$

$$[D_{ic} | \sigma_{ic}^2, \theta_W] = N(\exp(-s^2); v_{NN}(s, 0; \theta_W), \sigma_{ic}^2), \quad (\text{Initial condition}),$$

where σ_r^2 , σ_{bc}^2 , and σ_{ic}^2 govern the extent to which the neural network v_{NN} is constrained to satisfy Burgers' equation along with the initial conditions and boundary conditions. Assuming reasonably strong prior information for the PDE, boundary conditions, and initial conditions, we set $\sigma_r^2 = 0.05^2$ and $\sigma_{bc}^2 = \sigma_{ic}^2 = 0.01^2$. We re-emphasize that assuming the residual data are zero (e.g., $D_r = 0$) further constrains the prior on θ_W and these residuals are not actual measurements of the process.

After specifying the prior distributions and the neural network architecture, we draw 2,000 posterior samples using HMC with the No-U-Turn Sampler (Hoffman & Gelman, 2014), discarding the first 1,000 as burn-in and using a tree depth of 10. This is implemented using the “NumPyro” (Phan et al., 2019) and “JAX” (Bradbury et al., 2018) packages in Python and takes approximately 45 minutes on an Apple MacBook M4 Pro with 128GB RAM.

6.1 Results

We first evaluate whether the posterior samples generated from the No-U-Turn sampler produce accurate estimates of the true simulation parameters. Figure 3 presents violin plots for the parameters, excluding the neural network weights and biases θ_W , which are not the primary focus. Although some parameters show slight estimation bias, their credible intervals include the true values and are indicative of Bayesian learning.

Figures 4 and 5 compare the true values and predictions of v_t and δ_t , respectively. As shown in Figure 4, the predicted v_t space-time field and spatial and temporal marginal plots align well with the true values, despite the substantial amount of missing data. This is one of the primary advantages of incorporating mechanistic information into the model. In contrast, Figure 5 demonstrates that accurately predicting the discrepancy term, δ_t , is more challenging and subject to greater uncertainty, although the posterior mean (dashed line in the spatial and temporal slice plots) still reasonably matches the true process (solid line). Combining posterior samples from each term, Figure 6 shows the posterior mean

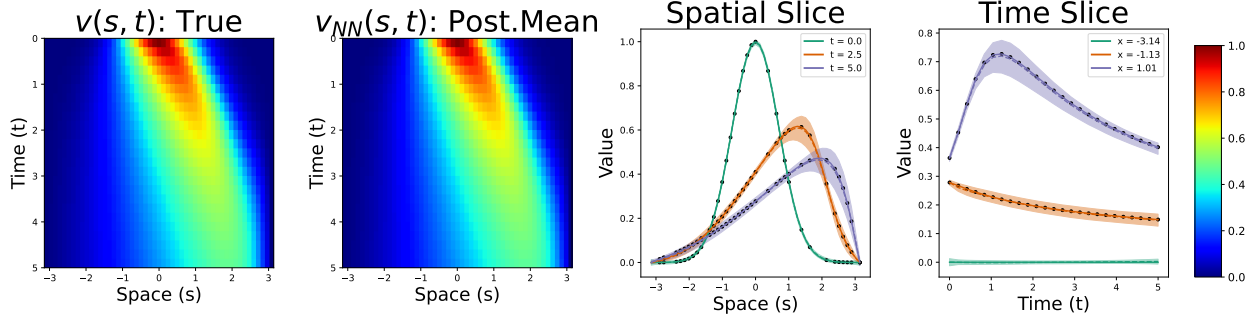


Figure 4: Left: true v_t ; Second from the left: predicted $v_{NN}(s, t)$; Second from the right: posterior means (dashed line) and corresponding 95% credible intervals (shaded region) for predicted values of a marginal spatial slice of $v_{NN,t}$ at times $t = 0, 2.5$, and 5 compared to the true process (solid line); Right: Posterior means (dashed line) and corresponding 95% credible intervals (shaded region) for a marginal time slice of the predicted values of v_t at spatial locations $s = -3.14, -1.13$, and 1.01 compared to the true process (solid line). In the right two plots, the dots correspond to spatial locations or time points when data are observed.

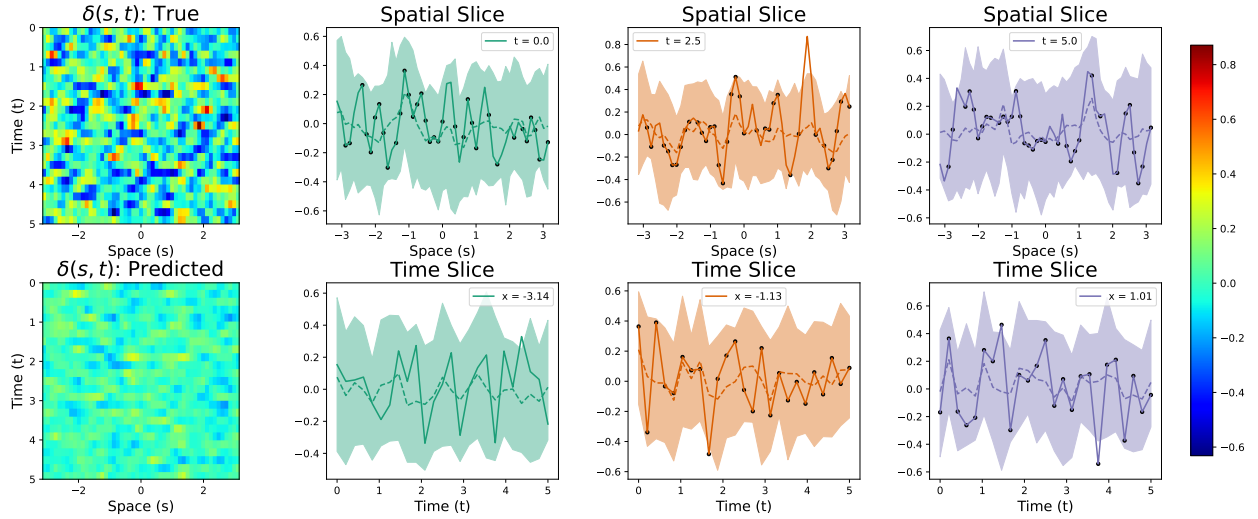


Figure 5: First Row, First Column: true values of the spatio-temporal discrepancy term δ_t ; Second Row, First Column: corresponding posterior prediction of δ_t ; First Row, Columns 2-4: spatial slice of posterior mean (dashed line) and 95% credible intervals (shaded region) for times $t = 0, 2$, and $t = 5$ compared to the true process (solid line); Second Row, Columns 2-4: time slice of posterior mean (dashed line) and 95% credible intervals (shaded region) for spatial locations $s = -3, -1$, and 1 compared to the true process (solid line). For all plots in columns 2-4, the dots correspond to spatial locations or time points when data is observed.

of the overall process u_t . While there are slight deviations from the true u_t , the posterior mean successfully captures the underlying spatio-temporal pattern, demonstrating the flexibility of the model.

We demonstrate a similar BHM-PINN for the case of normally-distributed data in the supplementary material. The results are quite comparable, if not slightly better, than those presented here for the Poisson data.

7 Discussion

The incorporation of mechanistic information in data-driven models has been of interest to scientists for decades. There are numerous paradigms for such data-driven approaches, ranging from variational data assimilation, online filtering,

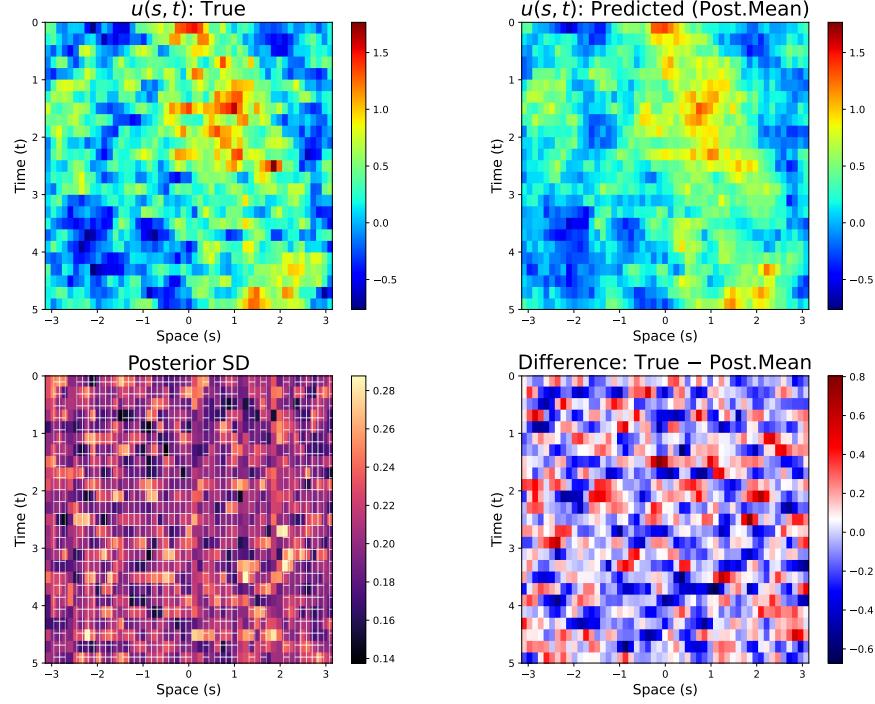


Figure 6: Top Left: true simulated $\exp(u_t)$; Top Right: predicted posterior mean of $\exp(u_t)$; Bottom Left: posterior standard deviation, with white rectangles indicating locations where data are observed; Bottom Right: difference between the true u_t and the posterior mean.

physical-statistical models and probabilistic numerical methods. Recently, there has been a resurgence of interest in these approaches because of the success of using deep neural networks with physical constraints (Raissi et al., 2020; Yang et al., 2021; Zou et al., 2024). Here, we review these PINN approaches and show some connections to the statistical approaches, followed by a generalization of B-PINNs to show that a general BHM framework can accommodate the B-PINN dynamical process but with the added flexibility to incorporate non-normal data models, covariate effects, and discrepancy processes. We illustrate such a BHM-PINN with the simulation of Poisson spatio-temporal data that are informed by the classic Burgers' nonlinear PDE, a covariate-based mean process, and a GP discrepancy term, and demonstrate we can recover important components of the model given uncertain and missing data.

The main contribution of this work is to place the PINN and B-PINN in the context of more traditional statistical approaches to incorporate mechanistic information in spatio-temporal models. Importantly, as a generalization, we show that a B-PINN naturally fits into a BHM framework (i.e., a BHM-PINN), which can accommodate non-normal data, covariates, and random discrepancy terms.

In some applications it may be preferable to embed a mechanistically-informed dynamical random spatio-temporal dynamical process as opposed to a neural network as in traditional physical-statistical BHMs. For example, Wikle et al. (2001) model 2-D spatio-temporal wind velocity components as a vector-autoregressive process that is motivated by an underlying set of (linear) PDEs, and Gopalan et al. (2019) motivates the latent process by a numerical PDE solution summed with a stochastic discrepancy process. However, for general nonlinear PDEs there may not be a convenient

autoregressive or general quadratic nonlinear (Wikle & Hooten, 2010) representation of the latent dynamics. In such cases, a mechanistic-BHM using a B-PINN approach may be preferred since automatic differentiation can be used for general, nonlinear PDEs. This could also be an advantage of using a model such as the one in Section 5 over PNM methods that lack a neural network representation, in which case it is not obvious that automatic differentiation would apply.

Related to the previous point, in the classical DSTM BHM (e.g., Cressie & Wikle, 2011) it is often useful to consider the latent dynamics in a dimension reduced framework such as $v_t = \Phi \alpha_t$ instead of neural network dynamics, $v_{NN,t}$. In such a case, v_t is the n -dimensional latent dynamic process and Φ is an $n \times n_\alpha$ dimensional basis function matrix with α_t an n_α -dimensional dynamic process where $n_\alpha < n$. For certain basis functions (e.g., Fourier) and systems, one can specify underlying mechanistic differential equations that govern the dynamics of α_t (e.g., Cressie & Wikle, 2011). This spectral approach could provide additional computational efficiency if the dynamics can be represented in a lower-dimensional manifold (reduced rank) basis expansion.

From our exploration of these methods, in particular, the incorporation of a B-PINN within a more general BHM, there are several challenges with PINN-based models that suggest additional research. Of primary concern is the challenge of computational efficiency when scaling to a realistic spatio-temporal application. Although automatic differentiation is quite helpful for implementing PDE models, it can be slow, and there are well-known challenges to Bayesian learning of deep neural network parameters. In addition, real-world processes are often described by more complex dynamics than can be parameterized by a PDE (all models are wrong), and modeling the discrepancy becomes crucial. Our simulation studies have shown there can be significant identifiability issues with a flexible discrepancy process and the underlying mechanistic process without rough *a priori* guidance as to at least the space and time scales corresponding to each. It is also the case the data volume and orientation can make a difference in accurately estimating the various model components. Lastly, as with all deep neural implementations, there is typically no *a priori* guidance to inform the model architecture. These are very active areas of research and we are optimistic that these challenges can be met and these methods can be scaled to many realistic spatio-temporal problems.

Acknowledgments

JSN was supported by the Director, Office of Science, Office of Biological and Environmental Research of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231 and by the Regional and Global Model Analysis Program area within the Earth and Environmental Systems Modeling Program as part of the Calibrated and Systematic Characterization, Attribution, and Detection of Extremes (CASCADE) project.

GG was supported by the Verification and Validation subprogram of the Advanced Simulation and Computing Program at the Los Alamos National Laboratory.

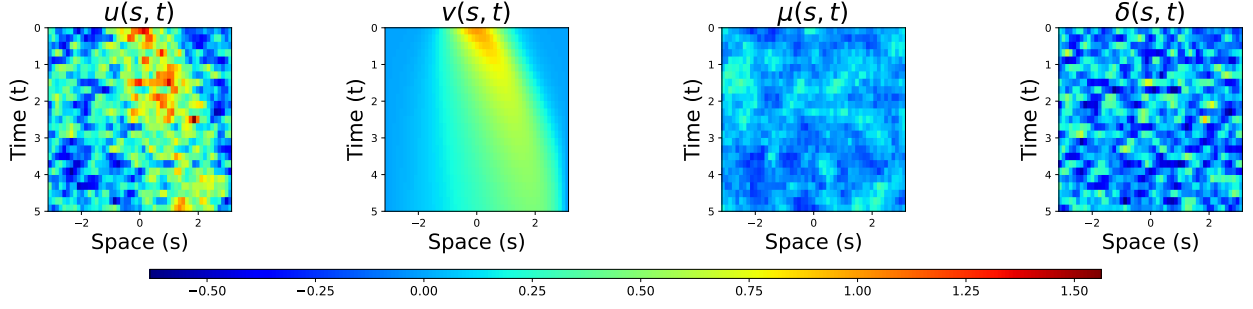


Figure 7: Simulated latent process $u(s, t)$ (left) and components over one-dimensional space and time: Burgers' equation $v(s, t)$ (second from left), mean process $\mu(s, t)$ (second from right), and discrepancy process $\delta(s, t)$ (right).

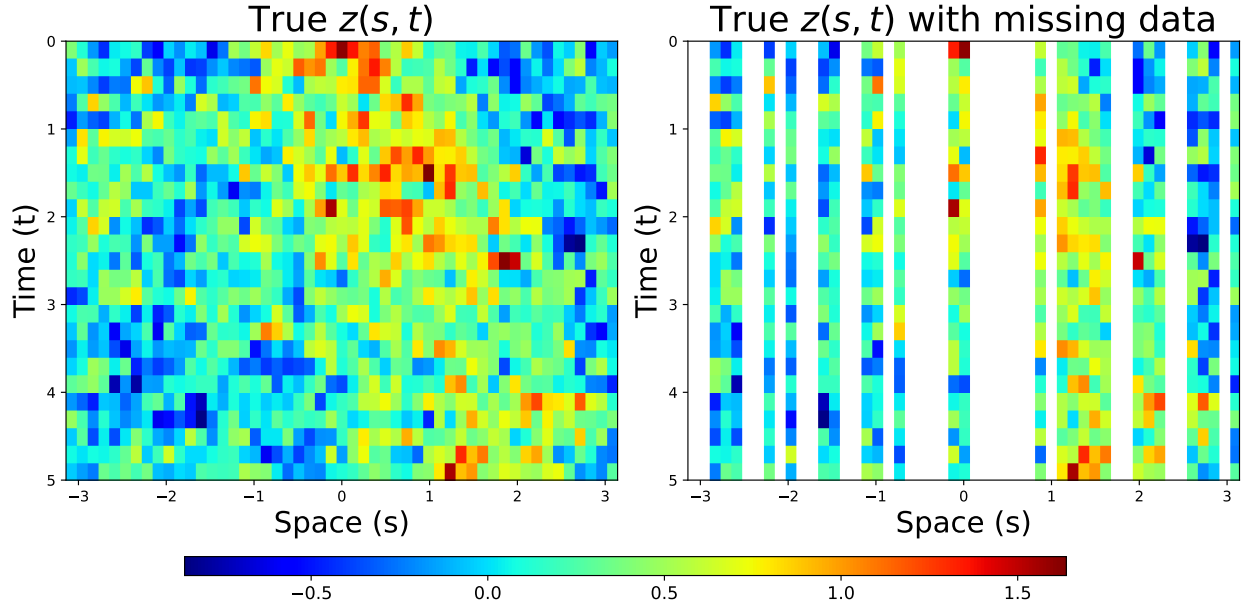
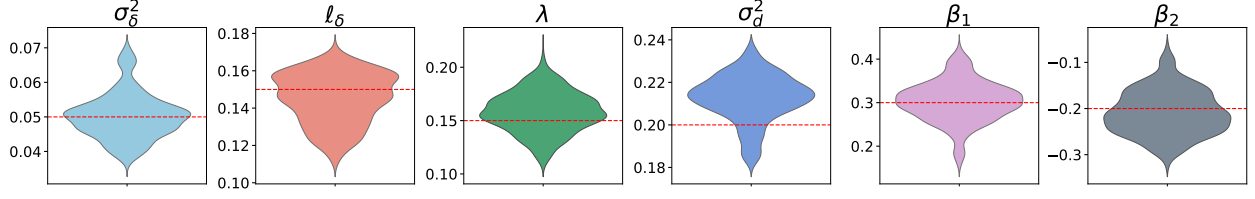


Figure 8: Left: Simulated z_t , Right: Missing data in z_t highlighted in white.

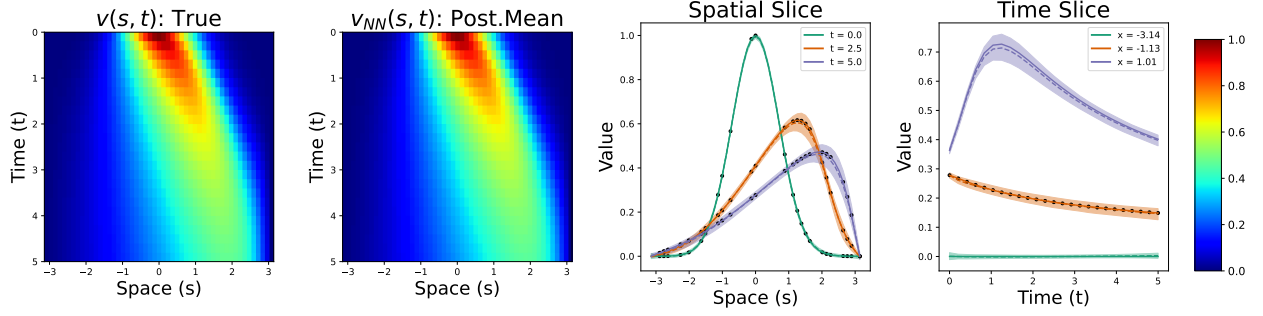
Supplementary Material

We present a simulation experiment where the data model follows a Normal distribution as opposed to Poisson as in the main paper. Specifically, the details of the simulation are the same as those described in the manuscript, except for:

1. **Process Mean ($\mu(s, t)$):** We consider two covariates in the mean process, $\mathbf{x}(s, t) = (x_1(s, t), x_2(s, t))^\top$, which are simulated from a zero-mean spatio-temporal GP with exponential covariance function $k_x((s, t), (s', t')) = 0.1 \cdot \exp\left(-\frac{\sqrt{(s-s')^2 + (t-t')^2}}{\ell_s}\right)$ where ℓ_s is set to 1/20 the maximum observed spatio-temporal distance across all location-time pairs. Specifically, we choose $\ell_s = 8.03/20$ and let $\boldsymbol{\beta} = (0.3, -0.2)^\top$.
2. **Simulated Observations ($z(s, t)$):** We then simulated observations from a Normal distribution, $z(s, t) \sim \mathcal{N}(u(s, t), \sigma_d^2)$, where $\sigma_d^2 = 0.2$. We assumed that 50% of the space coordinates are missing.



(a) Violin plots of the BHM model parameter posterior distributions, with the true value indicated by a dashed red line. From left to right, the plots correspond to σ_δ^2 , ℓ_δ , λ , σ_d^2 , β_1 , and β_2 .



(b) Left: true $v(s, t)$; Second from the left: Posterior mean of $v_{NN}(s, t)$; Second from the right: Posterior means (dashed line) and corresponding 95% credible intervals (shaded region) for predicted values of a marginal spatial slice of $v_{NN}(s, t)$ at times $t = 0, 2.5$, and 5 compared to the true process (solid line); Right: Posterior means (dashed line) and corresponding 95% credible intervals (shaded region) for a marginal time slice of the predicted values of $v_{NN}(s, t)$ at spatial locations $s = -3.14, -1.13$, and 1.01 compared to the true process (solid line). In the right two plots, the dots correspond to spatial locations or time points when data are observed.

Figure 9: Results

Figure 7 shows the simulated latent process and associated components and 8 shows the simulated data and the missingness used when fitting the model. Specifically, we consider the following BHM to fit these simulated data. First, using the notation defined in the main paper, the data and process models are given by

$$\text{Data Model : } \mathbf{z}_t = \mathbf{H}_t \mathbf{u}_t + \boldsymbol{\epsilon}_t, \quad \boldsymbol{\epsilon}_t \sim \mathcal{N}(\mathbf{0}, \sigma_d^2 \mathbf{I})$$

$$\text{True Process : } u(s, t) = \mathbf{x}(s, t)^\top \boldsymbol{\beta} + v_{NN}(s, t) + \delta(s, t),$$

where $\mathbf{H}_t$ is an incidence matrix. The latent dynamic and discrepancy processes are specified as

$$\text{Latent dynamic process } (v_{NN}(s, t; \boldsymbol{\theta}_W) : \frac{\partial v_{NN}(s, t)}{\partial t} \sim \mathcal{N}(v_{NN}(s, t) \frac{\partial v_{NN}(s, t)}{\partial s} - \lambda \frac{\partial^2 v_{NN}(s, t)}{\partial s^2}, \sigma_r^2)$$

$$v_{NN}(s, 0) \sim \mathcal{N}(\exp(-s^2), \sigma_{ic}^2) \quad (\text{Initial condition})$$

$$v_{NN}(-\pi, t), v_{NN}(\pi, t) \sim \mathcal{N}(0, \sigma_{BC}^2) \quad (\text{Boundary condition})$$

$$\text{Discrepancy process : } \delta(s, t) \sim \text{GP}(0, k(s, s'))$$

$$k(s, s') = \sigma_\delta^2 \exp\left(-\frac{(s - s')^2}{2\ell_\delta^2}\right).$$

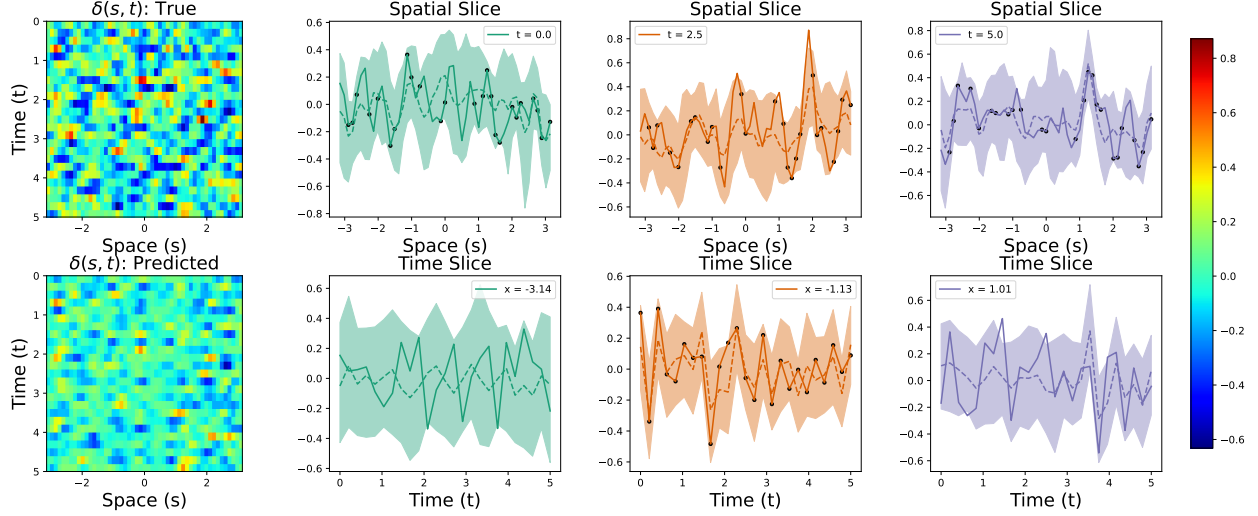


Figure 10: First Row, First Column: True values of the spatio-temporal discrepancy term δ_t ; Second Row, First Column: Corresponding posterior prediction of δ_t ; First Row, Columns 2-4: Spatial slice of posterior mean (dashed line) and 95% credible intervals (shaded region) for times $t = 0, 2$, and $t = 5$ compared to the true process (solid line); Second Row, Columns 2-4: Time slice of posterior mean (dashed line) and 95% credible intervals (shaded region) for spatial locations $s = -3, -1$, and 1 compared to the true process (solid line). For all plots in columns 2-4, the dots correspond to spatial locations or time points when data is observed.

The prior distributions are specified as

$$\begin{aligned} \sigma_d^2 &\sim \text{Cauchy}(\mu_d, \gamma_d) \cdot \mathbf{I}(\ell_d < \sigma_d^2 < u_d), & \beta &\sim \mathbf{N}(\mu_\beta, c_\beta \mathbf{I}) \\ \lambda &\sim \mathbf{N}(\mu_\lambda, \sigma_\lambda^2) \cdot \mathbf{I}(0 \leq \lambda < \infty), & \sigma_\delta^2 &\sim \text{Cauchy}(\mu_\delta, \gamma_\delta) \cdot \mathbf{I}(\ell_\delta < \sigma_\delta^2 < u_\delta) \\ \ell_\delta &\sim \mathbf{N}(\mu_\ell, \sigma_\ell^2) \cdot \mathbf{I}(\ell_\ell \leq \ell_\delta < u_\ell), & \theta_W &\sim \mathbf{N}(\mu_W, c_W \mathbf{I}), \end{aligned}$$

where we set $\mu_d = \mu_\delta = \mu_\ell = 0, \gamma_d = \gamma_\delta = \sigma_\ell^2 = 1$, and truncation bounds $\ell_d = 0.1, u_d = 0.3, \ell_\delta = 0.02, u_\delta = 0.1, \ell_\ell = 0.05$, and $u_\ell = 0.2$, in order to assign relatively informative priors for $\sigma_d^2, \sigma_\delta^2$, and ℓ_δ . For β, λ , and neural network weight and bias parameters θ_W , we set $\mu_\beta = \mu_W = \mathbf{0}, c_W = \sigma_\lambda^2 = 1, \mu_\lambda = 0$, and $c_\beta = 10$. The neural network for $\mathbf{u}_{NN,t}$ consists of $L = 3$ hidden layers, each with 16 hidden units. For PDE residuals, boundary, and initial condition data, we consider

$$\begin{aligned} [D_r | \sigma_r^2, \theta_W, \lambda] &= \mathbf{N}(0; [\frac{\partial v_{NN}}{\partial t} + v_{NN} \frac{\partial v_{NN}}{\partial s} - \lambda \frac{\partial^2 v_{NN}}{\partial s^2}](s, t; \theta_W), \sigma_r^2), & (\text{Burgers' equation}) \\ [D_{bc+} | \sigma_{bc}^2, \theta_W] &= \mathbf{N}(0; v_{NN}(\pi, t; \theta_W), \sigma_{bc}^2), & (\text{Right boundary condition}) \\ [D_{bc-} | \sigma_{bc}^2, \theta_W] &= \mathbf{N}(0; v_{NN}(-\pi, t; \theta_W), \sigma_{bc}^2), & (\text{Left boundary condition}) \\ [D_{ic} | \sigma_{ic}^2, \theta_W] &= \mathbf{N}(\exp(-s^2); v_{NN}(s, 0; \theta_W), \sigma_{ic}^2), & (\text{Initial condition}), \end{aligned}$$

where $\sigma_r^2 = 0.05^2$ and $\sigma_{bc}^2 = \sigma_{ic}^2 = 0.01^2$.

After specifying the prior distributions and the neural network architecture, we draw 2,000 posterior samples using HMC with the No-U-Turn Sampler (Hoffman & Gelman, 2014), discarding the first 1,000 as burn-in and using a tree

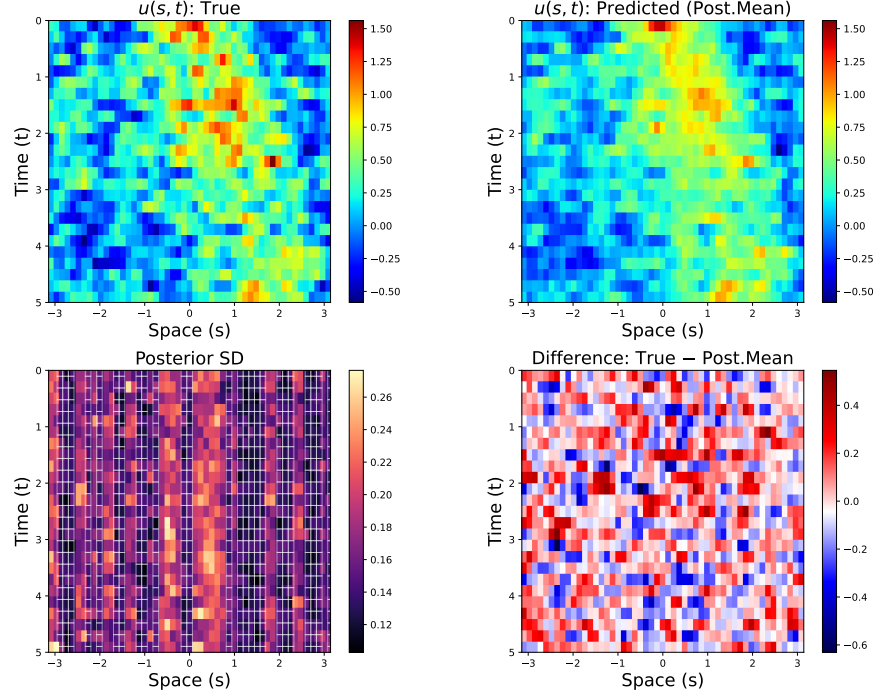


Figure 11: Top Left: True simulated u_t ; Top Right: Predicted posterior mean u_t ; Bottom Left: Posterior standard deviation, with white rectangles indicating locations where data are observed; Bottom Right: Difference between the true u_t and the posterior mean.

depth of 10. This is implemented using the “NumPyro” (Phan et al., 2019) and “JAX” (Bradbury et al., 2018) packages in Python and takes approximately 45 minutes on an Apple MacBook M4 Pro with 128GB RAM.

Figure 9a presents violin plots for the parameters, excluding the neural network weights and biases θ_W , for which the credible intervals contain the true values. Figure 9b show comparisons between $v(s, t)$ and the posterior mean of $v_{NN}(s, t)$, showing that the posterior mean closely aligns with the true $v(s, t)$ despite a substantial amount of missing data. Figure 10 compares the true δ_t with its predicted values, indicating that predicting the discrepancy term is more challenging subject to more uncertainty. Nevertheless, the posterior mean (dashed line in the spatial and temporal slice plots) provides a reasonably good match to the true process (solid line).

Figure 11 shows the posterior mean of the overall process u_t . While there are minor deviations from the true u_t , the posterior mean successfully captures the underlying spatio-temporal pattern, with the posterior standard deviation being higher in locations where data are missing. This demonstrates the flexibility of the model.

References

- Bateman, H. (1915). Some recent researches on the motion of fluids. *Monthly Weather Review*, 43(4), 163–170.
- Bennett, A. F. (1992). *Inverse Methods in Physical Oceanography*. Cambridge University Press.
- Berliner, L. M. (1996). Hierarchical Bayesian time series models. *Maximum entropy and Bayesian methods: Santa Fe, New Mexico, USA, 1995 Proceedings of the fifteenth international workshop on maximum entropy and Bayesian*

- methods*, 15–22.
- Berliner, L. M. (2003). Physical-statistical modeling in geophysics. *Journal of Geophysical Research: Atmospheres*, 108(D24).
- Berliner, L. M., Jezek, K., Cressie, N., Kim, Y., Lam, C. Q., & Van Der Veen, C. J. (2008). Modeling dynamic controls on ice streams: A Bayesian statistical approach. *Journal of Glaciology*, 54(187), 705–714.
- Berliner, L. M., Royle, J. A., Wikle, C. K., & Milliff, R. F. (1999). Bayesian methods in the atmospheric sciences. *Bayesian Statistics*, 6, 83–100.
- Blei, D. M., Kucukelbir, A., & McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518), 859–877.
- Bradbury, J., Frostig, R., Hawkins, P., Johnson, M. J., Leary, C., Maclaurin, D., Necula, G., Paszke, A., VanderPlas, J., Wanderman-Milne, S., & Zhang, Q. (2018). *JAX: composable transformations of Python+NumPy programs*. <http://github.com/jax-ml/jax>. version 0.3.13.
- Brooks, S., Gelman, A., Jones, G., & Meng, X.-L. (2011). *Handbook of Markov Chain Monte Carlo*. CRC press.
- Burgers, J. M. (1948). A mathematical model illustrating the theory of turbulence. *Advances in Applied Mechanics*, 1, 171–199.
- Castillo, I. & Egels, P. (2024). Posterior and variational inference for deep neural networks with heavy-tailed weights. *arXiv preprint arXiv:2406.03369*.
- Charney, J. G., Fjörtoft, R., & Neumann, J. v. (1950). Numerical integration of the barotropic vorticity equation. *Tellus*, 2(4), 237–254.
- Chkrebtii, O. A., Campbell, D. A., Calderhead, B., & Girolami, M. A. (2016). Bayesian solution uncertainty quantification for differential equations. *Bayesian Analysis*, 11(4), 1239–1267.
- Cockayne, J., Oates, C. J., Sullivan, T. J., & Girolami, M. (2019). Bayesian probabilistic numerical methods. *SIAM review*, 61(4), 756–789.
- Cressie, N. (1990). The origins of kriging. *Mathematical Geology*, 22, 239–252.
- Cressie, N. (1993). *Statistics for Spatial Data*. John Wiley & Sons.
- Cressie, N. & Wikle, C. K. (2011). *Statistics for Spatio-Temporal Data*. John Wiley & Sons.
- Daley, R. (1993). *Atmospheric Data Analysis*. Cambridge University Press.
- Diaconis, P. (1988). Bayesian numerical analysis. *Statistical Decision Theory and Related Topics IV*, 1, 163–175.
- Eliassen, A. (1954). Provisional report on calculation of spatial covariance and autocorrelation of the pressure field. *Report no 5, Videnskaps-Akademiets Institutt for Vaer-Og Klimaforskning, Oslo, Norway*, 12.

- Errico, R. M. (1997). What is an adjoint model? *Bulletin of the American Meteorological Society*, 78(11), 2577–2592.
- Eubank, R. L. (1999). *Nonparametric regression and spline smoothing*. CRC press.
- Fuentes, M. & Raftery, A. E. (2005). Model evaluation and spatial interpolation by Bayesian combination of observations with outputs from numerical models. *Biometrics*, 61(1), 36–45.
- Gabrielsson, J. & Weiner, D. (2001). *Pharmacokinetic and Pharmacodynamic Data Analysis: Concepts and Applications*. CRC press.
- Gandin, L. S. (1963). Objective analysis of meteorological fields. *Israel program for scientific translations*, 242.
- Gauss, C. F. (1809). *Theoria motus corporum coelestium in sectionibus conicis solem ambientium auctore Carolo Friderico Gauss*. sumtibus Frid. Perthes et IH Besser.
- Gelfand, A. E., Diggle, P., Guttorp, P., & Fuentes, M. (2010). *Handbook of Spatial Statistics*. CRC press.
- Gelfand, A. E. & Smith, A. F. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American statistical association*, 85(410), 398–409.
- Gopalan, G., Hrafnkelsson, B., Aðalgeirsdóttir, G., Jarosch, A. H., & Pálsson, F. (2018). A Bayesian hierarchical model for glacial dynamics based on the shallow ice approximation and its evaluation using analytical solutions. *The Cryosphere*, 12(7), 2229–2248.
- Gopalan, G., Hrafnkelsson, B., Wikle, C. K., Rue, H., Aðalgeirsdóttir, G., Jarosch, A. H., & Pálsson, F. (2019). A hierarchical spatiotemporal statistical model motivated by glaciology. *Journal of Agricultural, Biological and Environmental Statistics*, 24(4), 669–692.
- Gramacy, R. B. (2020). *Surrogates: Gaussian Process Modeling, Design, and Optimization for the Applied Sciences*. Chapman and Hall/CRC.
- Grossmann, T. G., Komorowska, U. J., Latz, J., & Schönlieb, C.-B. (2024). Can physics-informed neural networks beat the finite element method? *IMA Journal of Applied Mathematics*, hxae011.
- Heaton, M. J., Datta, A., Finley, A. O., Furrer, R., Guinness, J., Guhaniyogi, R., Gerber, F., Gramacy, R. B., Hammerling, D., Katzfuss, M., et al. (2019). A case study competition among methods for analyzing large spatial data. *Journal of Agricultural, Biological and Environmental Statistics*, 24, 398–425.
- Hefley, T. J., Hooten, M. B., Russell, R. E., Walsh, D. P., & Powell, J. A. (2017). When mechanism matters: Bayesian forecasting using models of ecological diffusion. *Ecology Letters*, 20(5), 640–650.
- Heine, V. (1955). Models for two-dimensional stationary stochastic processes. *Biometrika*, 42(1-2), 170–178.
- Hennig, P., Osborne, M. A., & Girolami, M. (2015). Probabilistic numerics and uncertainty in computations. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 471(2179), 20150142.

- Herbei, R., McKeague, I. W., & Speer, K. G. (2008). Gyres and jets: Inversion of tracer data for ocean circulation structure. *Journal of Physical Oceanography*, 38(6), 1180–1202.
- Hoffman, M. D. & Gelman, A. (2014). The No-U-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15(1), 1593–1623.
- Hooten, M. B. & Wikle, C. K. (2008). A hierarchical Bayesian non-linear spatio-temporal model for the spread of invasive species with application to the Eurasian Collared-Dove. *Environmental and Ecological Statistics*, 15, 59–70.
- Huang, J., Pang, Y., Liu, Y., & Yan, H. (2023). Posterior Regularized Bayesian Neural Network incorporating soft and hard knowledge constraints. *Knowledge-Based Systems*, 259, 110043.
- Kalman, R. (1960). A new approach to linear filtering and prediction problems. *Transactions of the ASME—Journal of Basic Engineering*, 82, 35–45.
- Karniadakis, G. E., Kevrekidis, I. G., Lu, L., Perdikaris, P., Wang, S., & Yang, L. (2021). Physics-informed machine learning. *Nature Reviews Physics*, 3(6), 422–440. <https://doi.org/10.1038/s42254-021-00314-5>
- Katzfuss, M., Stroud, J. R., & Wikle, C. K. (2016). Understanding the ensemble Kalman filter. *The American Statistician*, 70(4), 350–357.
- Laloyaux, P., Bonavita, M., Chrut, M., & Gürol, S. (2020). Exploring the potential and limitations of weak-constraint 4D-Var. *Quarterly Journal of the Royal Meteorological Society*, 146(733), 4067–4082.
- Larkin, F. (1972). Gaussian measure in Hilbert space and applications in numerical analysis. *The Rocky Mountain Journal of Mathematics*, 379–421.
- Leach, C. B., Hoeting, J. A., Pepin, K. M., Eiras, A. E., Hooten, M. B., & Webb, C. T. (2020). Linking mosquito surveillance to dengue fever through Bayesian mechanistic modeling. *PLOS Neglected Tropical Diseases*, 14(11), e0008868.
- Leeds, W. B., Wikle, C. K., & Fiechter, J. (2014). Emulator-assisted reduced-rank ecological data assimilation for nonlinear multivariate dynamical spatio-temporal processes. *Statistical Methodology*, 17, 126–138.
- Lewis, J. & Lakshmivarahan, S. (2008). Sasaki’s pivotal contribution: Calculus of variations applied to weather map analysis. *Monthly Weather Review*, 136(9), 3553–3567.
- Lewis, J. M. & Derber, J. C. (1985). The use of adjoint equations to solve a variational adjustment problem with advective constraints. *Tellus A*, 37(4), 309–322.
- Lindgren, F., Bolin, D., & Rue, H. (2022). The SPDE approach for Gaussian and non-Gaussian fields: 10 years and still running. *Spatial Statistics*, 50, 100599.

- Lindgren, F., Rue, H., & Lindström, J. (2011). 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.
- Lorenc, A. C. (1986). Analysis methods for numerical weather prediction. *Quarterly Journal of the Royal Meteorological Society*, 112(474), 1177–1194.
- Matheron, G. (1963). Principles of geostatistics. *Economic Geology*, 58(8), 1246–1266.
- McDermott, P. L. & Wikle, C. K. (2019). Bayesian recurrent neural network models for forecasting and quantifying uncertainty in spatial-temporal data. *Entropy*, 21(2), 184.
- Meng, T., Jing, X., Yan, Z., & Pedrycz, W. (2020). A survey on machine learning for data fusion. *Information Fusion*, 57, 115–129.
- Meng, X., Yang, L., Mao, Z., del Águila Ferrandis, J., & Karniadakis, G. E. (2022). Learning functional priors and posteriors from data and physics. *Journal of Computational Physics*, 457, 111073.
- Milliff, R. F., Bonazzi, A., Wikle, C. K., Pinardi, N., & Berliner, L. M. (2011). Ocean ensemble forecasting. Part I: Ensemble Mediterranean winds from a Bayesian hierarchical model. *Quarterly Journal of the Royal Meteorological Society*, 137(657), 858–878.
- Neal, R. M. (2012). *Bayesian learning for neural networks*, volume 118. Springer Science & Business Media.
- Neal, R. M. et al. (2011). MCMC using Hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, 2(11), 2.
- North, J. S., Wikle, C. K., & Schliep, E. M. (2023). A review of data-driven discovery for dynamic systems. *International Statistical Review*, 91(3), 464–492.
- North, J. S., Wikle, C. K., & Schliep, E. M. (2025). A Bayesian approach for spatio-temporal data-driven dynamic equation discovery. *Bayesian Analysis*, 20(2), 375–404.
- Pagendam, D., Kuhnert, P., Leeds, W., Wikle, C., Bartley, R., & Peterson, E. (2014). Assimilating catchment processes with monitoring data to estimate sediment loads to the great barrier reef. *Environmetrics*, 25(4), 214–229.
- Phan, D., Pradhan, N., & Jankowiak, M. (2019). Composable effects for flexible and accelerated probabilistic programming in numpyro. *arXiv preprint arXiv:1912.11554*.
- Phillips, N. A. (1956). The general circulation of the atmosphere: A numerical experiment. *Quarterly Journal of the Royal Meteorological Society*, 82(352), 123–164.
- Psaros, A. F., Meng, X., Zou, Z., Guo, L., & Karniadakis, G. E. (2023). Uncertainty quantification in scientific machine learning: Methods, metrics, and comparisons. *Journal of Computational Physics*, 477, 111902.
- Raissi, M. (2018). Deep hidden physics models: Deep learning of nonlinear partial differential equations. *Journal of Machine Learning Research*, 19, 1–24.

- Raissi, M., Perdikaris, P., & Karniadakis, G. (2019). Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378, 686–707. <https://doi.org/10.1016/j.jcp.2018.10.045>
- Raissi, M., Perdikaris, P., & Karniadakis, G. E. (2017a). Physics informed deep learning (Part I): Data-driven solutions of nonlinear partial differential equations. *arXiv preprint arXiv:1711.10561*, Part I, 1–22.
- Raissi, M., Perdikaris, P., & Karniadakis, G. E. (2017b). Physics informed deep learning (Part II): Data-driven discovery of nonlinear partial differential equations. *arXiv preprint arXiv:1711.10566*, Part II, 1–19.
- Raissi, M., Yazdani, A., & Karniadakis, G. E. (2020). Hidden fluid mechanics: Learning velocity and pressure fields from flow visualizations. *Science*, 367(6481), 1026–1030. <https://doi.org/10.1126/science.aaw4741>
- Ramsay, T. (2002). Spline smoothing over difficult regions. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 64(2), 307–319.
- Richardson, L. F. (1922). *Weather Prediction by Numerical Process*. University Press.
- Sahimi, M. (2024). Physics-informed and data-driven discovery of governing equations for complex phenomena in heterogeneous media. *Physical Review E*, 109(4), 041001.
- Sangalli, L. M. (2021). Spatial regression with partial differential equation regularisation. *International Statistical Review*, 89(3), 505–531.
- Sasaki, Y. (1958). An objective analysis based on the variational method. *Journal of the Meteorological Society of Japan. Ser. II*, 36(3), 77–88.
- Sasaki, Y. (1970). Some basic formalisms in numerical variational analysis. *Monthly Weather Review*, 98(12), 875–883.
- Sell, T. & Singh, S. S. (2023). Trace-class Gaussian priors for Bayesian learning of neural networks with MCMC. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(1), 46–66.
- Shumway, R. H. & Stoffer, D. S. (2000). *Time Series Analysis and its Applications*, volume 3. Springer.
- Stephens, J. (1970). Variational initialization with the balance equation. *Journal of Applied Meteorology and Climatology*, 9(5), 732–739.
- Talagrand, O. (1997). Assimilation of observations, an introduction (gtspecial issue\data assimilation in meteorology and oceanography: Theory and practice). *Journal of the Meteorological Society of Japan. Ser. II*, 75(1B), 191–209.
- Thompson, P. D. (1969). Reduction of analysis error through constraints of dynamical consistency. *Journal of Applied Meteorology and Climatology*, 8(5), 738–742.
- Tomasetto, M., Arnone, E., & Sangalli, L. M. (2024). Modeling anisotropy and non-stationarity through physics-informed spatial regression. *Environmetrics*, 35(8), e2889.

- Toscano, J. D., Oommen, V., Varghese, A. J., Zou, Z., Ahmadi Daryakenari, N., Wu, C., & Karniadakis, G. E. (2025). From PINNs to PIKANs: Recent advances in physics-informed machine learning. *Machine Learning for Computational Science and Engineering*, 1(1), 1–43.
- Tran, D., Ranganath, R., & Blei, D. M. (2015). The variational Gaussian process. *arXiv preprint arXiv:1511.06499*.
- Wahba, G. (1990). *Spline Models for Observational Data*. SIAM.
- Wang, S., Sankaran, S., & Perdikaris, P. (2022). Respecting causality is all you need for training physics-informed neural networks. *arXiv preprint arXiv:2203.07404*.
- Whittle, P. (1954). On stationary processes in the plane. *Biometrika*, 434–449.
- Wikle, C. K. (2003). Hierarchical Bayesian models for predicting the spread of ecological processes. *Ecology*, 84(6), 1382–1394.
- Wikle, C. K. & Berliner, L. M. (2005). Combining information across spatial scales. *Technometrics*, 47(1), 80–91.
- Wikle, C. K. & Berliner, L. M. (2007). A Bayesian tutorial for data assimilation. *Physica D: Nonlinear Phenomena*, 230(1-2), 1–16.
- Wikle, C. K. & Hooten, M. B. (2010). A general science-based framework for dynamical spatio-temporal models. *Test*, 19, 417–451. <https://doi.org/10.1007/s11749-010-0209-z>.
- Wikle, C. K., Milliff, R. F., Nychka, D., & Berliner, L. M. (2001). Spatiotemporal hierarchical Bayesian modeling tropical ocean surface winds. *Journal of the American Statistical Association*, 96(454), 382–397.
- Wikle, C. K., Zammit-Mangion, A., & Cressie, N. (2019). *Spatio-Temporal Statistics with R*. Chapman & Hall/CRC.
- Xu, K., Wikle, C. K., & Fox, N. I. (2005). A kernel-based spatio-temporal dynamical model for nowcasting weather radar reflectivities. *Journal of the American Statistical Association*, 100(472), 1133–1144.
- Yang, L., Meng, X., & Karniadakis, G. E. (2021). B-PINNs: Bayesian physics-informed neural networks for forward and inverse PDE problems with noisy data. *Journal of Computational Physics*, 425, 109913.
- Zammit-Mangion, A., Rougier, J., Bamber, J., & Schön, N. (2014). Resolving the antarctic contribution to sea-level rise: A hierarchical modelling framework. *Environmetrics*, 25(4), 245–264.
- Zhang, Q., Zhang, Q., Zhao, Y., Liu, Y., Wang, Z., & Ma, Y. (2025). Inverse solution of process parameters in gear grinding using hierarchical bayesian physics informed neural network (hbpinn). *Scientific Reports*, 15(1), 35397. <https://doi.org/10.1038/s41598-025-18005-x>
- Zhuang, L. & Cressie, N. (2014). Bayesian hierarchical statistical SIRS models. *Statistical Methods & Applications*, 23, 601–646.
- Zou, Z., Meng, X., & Karniadakis, G. E. (2024). Correcting model misspecification in physics-informed neural networks (PINNs). *Journal of Computational Physics*, 505, 112918.