Bayesian Neural Networks With Physics-Informed Priors With Application to Boundary Layer Velocity

Luca Menicali, David H. Richter, and Stefano Castruccio

***Abstract*—One of the most popular recent areas of machine learning predicates the use of neural networks (NNs) augmented by information about the underlying process in the form of partial differential equations (PDEs). These physics-informed NNs (PINNs) are obtained by penalizing the inference with a PDE and have been cast as a minimization problem currently lacking a formal approach to quantify the uncertainty. In this work, we propose a novel model-based framework that regards the PDE as a prior information of a deep Bayesian NN (BNN), physics-informed prior (PIP)-BNN. The prior is calibrated without data to resemble the PDE solution in the prior mean, while our degree of confidence in the PDE with respect to the data is expressed in terms of the prior variance. The information embedded in the PDE are then propagated to the posterior yielding physics-informed forecasts with uncertainty quantification. We apply our approach to a simulated viscous fluid and to an experimentally obtained turbulent boundary layer velocity in a water tunnel using an appropriately simplified Navier–Stokes (NS) equation. Our approach requires very few observations to produce physically consistent forecasts as opposed to nonphysical forecasts stemming from noninformed priors, thereby allowing forecasting complex systems, where some amount of data as well as some contextual knowledge is available.**

***Index Terms*—Bayesian neural network (BNN), Navier–Stokes (NS), physics-informed priors (PIPs), variational inference.**

I. Introduction

OVER the last two decades, data-rich problems have become increasingly more relevant due to the extraordinary increase in the volume, velocity, and variety of data available. The simultaneous growth in computational capabilities has also allowed researchers to implement highly complex models, especially nonparametric constructs such as neural networks (NNs), able to capture complex patterns in the data. This has facilitated new findings, such as predicting long-term trends in the El Niño-Southern Oscillation [1], constructing optimal wind farms given scattered observational data [2], and understanding complex flows in computational fluid dynamics [3].

Received 26 September 2024; revised 4 February 2025 and 1 May 2025; accepted 2 June 2025. The work of Stefano Castruccio was supported in part by NSF Directorate for Computer and Information Science and Engineering (CISE) Office of Advanced Cyberinfrastructure (OAC) under Grant 2347239 and in part by University of Notre Dame Provost Strategic Framework Grants. *(Corresponding author: Luca Menicali.)*

Luca Menicali and Stefano Castruccio are with the Department of Applied and Computational Mathematics and Statistics, University of Notre Dame, Notre Dame, IN 46556 USA (e-mail: lmenical@nd.edu; scastruc@nd.edu).

David H. Richter is with the Department of Civil and Environmental Engineering, University of Notre Dame, Notre Dame, IN 46556 USA (e-mail: david.richter.26@nd.edu).

This article has supplementary downloadable material available at https://doi.org/10.1109/TNNLS.2025.3577508, provided by the authors.

Digital Object Identifier 10.1109/TNNLS.2025.3577508

While NNs have unequivocally presented new opportunities for data-rich problems, in their original formulation they discard any contextual information that may be provided by a theoretical model. This is a major limitation of these constructs, as a significant amount of applied problems consists of some data as well as some information in the form of a partial differential equation (PDE). In their original formulation, NNs are unsuitable to account for additional context, as parameters are learned by minimizing a loss function which is agnostic of the specific problem [4]. A rapidly growing branch of machine learning has been focusing on NNs that can account for the physical laws governing a given system of interest. These physics-informed NNs (PINNs) incorporate this knowledge as an additional PDE constraint on the objective function, so that the inference steers the model toward predictions that are physically consistent [5]. PINNs have been shown to perform better than NNs in many applications, where data are difficult to collect but information on the process in the form of a PDE are available [2], [6], [7]. Alternative approaches to incorporate physical knowledge in deep NNs are Hamiltonian NNs [8], Lagrangian NNs [9], and Newtonian NNs [10]. While promising in their own right, each of the three models ultimately aligns with the standard PINN framework introduced in [5], i.e., none of them are Bayesian models, and physical knowledge is incorporated via a physics-augmented loss function.

While PINNs have demonstrated more flexibility than standard NNs, to the best of the authors' knowledge, they have always been regarded as algorithmic minimization problems rather than an outcome of an inferential process from a statistical model. As such, a formal approach to quantify uncertainty in PINNs has not yet been agreed upon [11]. In the context of NNs, uncertainty quantification can be assessed using dropout, a bootstrapping-inspired method [12], [13], or in a Bayesian framework by imposing (often vague) prior distributions, thereby introducing parameter uncertainty [14]. In the context of PINNs, Bayesian NNs (BNNs) have been used to augment PINNs' predictions in the identification of the system inertia and damping in power systems [15] as well as in the estimation of hypocenter location during seismic activity [16]. Even though these studies' results do include measures of uncertainty, they do so in a similar fashion as classical data-driven PINN approaches in forward problems and, most importantly, are cast as PDE solvers [17]. Previous attempts in fact focus on PINN-driven approximations to PDEs to construct Bayesian priors and eventually augment those deterministic approximations [11], [17]. In particular, the B-PINN model [17] is a Bayesian PDE solver, which assumes

that the data are noisy observations centered around the latent solution. To the best of the authors' knowledge, a unified model-based approach aimed at presenting PINNs as outcomes from a statistical model has not been developed yet.

In this work, we propose a novel Bayesian approach that incorporates contextual knowledge in the prior stage of a deep BNN. Our construction of PDE-informed priors builds upon a well-established tradition of embedding domain knowledge into Bayesian models. In econometrics, structured priors have been shown to improve forecasting under uncertainty in Bayesian vector autoregression models [18], [19], [20]. Similarly, in spatial statistics, stochastic PDEs define Gaussian process priors through differential operators, enabling scalable inference for continuous spatial fields [21]. For inverse problems, PDE constraints serve as natural regularizers, ensuring well-posed posterior inferences [22]. Recent advances in machine learning rigorously characterize the role of latent variable nonidentifiability in variational autoencoders [23], proposing hierarchical priors that mitigate posterior collapse by enforcing meaningful latent structure. Diffusion models, inspired by stochastic differential equations, employ learned noise schedules that mimic PDE-driven dynamics, improving sample quality in generative tasks [24].

Our approach is fundamentally different from the classical PINN paradigm, which instead adapts the loss function to a set of governing equations [5], and from previous Bayesian-inspired approaches to PINNs, whose main focus remains uncertainty quantification of latent PDE solutions [11], [17]. At the core of our proposed method is the acknowledgment that an established theory effectively constitutes our prior beliefs on a given system's behavior, which is then updated with some data, and is, thus, an inherently Bayesian approach. This work also addresses the issue of uncertainty quantification in PINNs which is naturally derived from the posterior distribution. In practice, we propose to calibrate the parameters' prior distribution such so the prior mean is close to the PDE solution, while the prior variance reflects our degree of confidence in the PDE with respect to the data. From an epistemological perspective, our approach regards PINNs as a Bayesian construct, where the prior calibration occurs before any observation is made, so that prior and data are modeled independently of each other to obtain physically consistent posteriors with uncertainty quantification. The large parameter space of BNNs implies a computationally challenging inference, so we approximate it with a variational inference approach [14]. Our methodology is very general as it can be adapted to any system for which an NN (or more generally any model) is to be applied along with some contextual information in the form of a PDE and prescribes a formal approach to incorporate both in the analysis. Henceforth, we will refer to our method as a BNN with physics-informed priors (PIPs)-BNN for short.

We consider an application of a turbulent boundary layer in a water tunnel—the flow that develops as fluid flows over a wall beneath. The governing set of PDEs, the Navier–Stokes (NS) equations, dictate the conservation of mass and momentum in three spatial dimensions and are generally able to describe accurately the turbulent, time-dependent velocity in boundary layers. For such turbulent flows, however, the goal is rarely to predict the instantaneous velocity field. Instead, the equations are typically averaged so as to predict the mean velocity field, which requires certain modeling assumptions. We show how physics-informed priors acknowledge some degree of variation in the mean water velocity, as expected from the experimental configuration. When augmented with sufficiently flexible prior variance and observations, our approach produces realistic forecasts even with a small number of data. Noninformed priors instead result in nonphysical predictions.

The remainder of this article is organized as follows. Section II outlines the proposed Bayesian model and the algorithm chosen to estimate its parameters. Section III presents a simulation study with a viscous fluid to compare our approach with a standard BNN with vague priors. Section IV reports the proposed model's results when implemented on the boundary layer velocity application. Section V summarizes this article findings and discusses future work.

II. Methods

This section formally introduces the general Bayesian framework PIP-BNN and proceeds as follows. We first review deep BNNs in Section II-A. Then, we introduce our novel BNN prior physical calibration in Section II-B. Due to the high number of parameters in the BNN, deriving the posteriors is computationally challenging, so Section II-C presents a variational inference approach to approximate the posteriors. Finally, Section II-D highlights the differences between PIP-BNN and existing methods.

A. Bayesian Neural Network

We consider a spatiotemporal process $u(\boldsymbol{s}, t)$, where $(\boldsymbol{s}, t) \in \Omega \times [0, T] \subset \mathbb{R}^d \times [0, \infty)$ and d is the number of spatial dimensions. We model u with an L-layer deep BNN parameterized by $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_L) \in \Theta$, where $\boldsymbol{\theta}_l = \{\boldsymbol{W}_l, \boldsymbol{b}_l\}$ is the parameters in the lth layer, comprised of its weights $\boldsymbol{W}_l$ and bias terms $\boldsymbol{b}_l$, and Θ is the parameter space. Formally, we have

$$\boldsymbol{H}_0 = (\boldsymbol{s}, t) \tag{1a}$$

$$\boldsymbol{H}_l = g\left(\boldsymbol{W}_l \boldsymbol{H}_{l-1} + \boldsymbol{b}_l\right), \quad l \in \{1, \ldots, L\} \tag{1b}$$

$$\boldsymbol{H}_L = \left(\mu_u(\boldsymbol{s}, t), \sigma_u^2(\boldsymbol{s}, t)\right) \in \mathbb{R} \times [0, \infty) \tag{1c}$$

$$u(\boldsymbol{s}, t) \sim \mathcal{N}\left(\mu_u(\boldsymbol{s}, t), \sigma_u^2(\boldsymbol{s}, t)\right) \tag{1d}$$

where the Lth layer is 2-D, every hidden layer $l \in \{1, \ldots, L-1\}$ has the same number of nodes K (although this assumption can be relaxed), and $u(\boldsymbol{s}, t)$ is distributed as a normal distribution with mean $\mu_u(\boldsymbol{s}, t)$ and variance $\sigma_u^2(\boldsymbol{s}, t)$. Each realization of u is, therefore, assumed to be conditionally independent given the mean and variance. The nonlinear activation function g is assumed to be the same for all l, and common choices for g are the rectified linear unit, sigmoid, or hyperbolic tangent function [25].

The parameters are assumed to have a prior distribution

$$p(\boldsymbol{\theta}_l) = \mathcal{N}\left(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l\right), \quad l \in \{1, \ldots, L\}. \tag{2}$$

Then, if we denote with $\boldsymbol{u} = \{u(\boldsymbol{s}_1, t_1), \ldots, u(\boldsymbol{s}_n, t_n)\}$ the observations collected at n spatiotemporal locations and

$p(\boldsymbol{\theta}) = \prod_{l=1}^{L} p(\boldsymbol{\theta}_l)$ the prior of $\boldsymbol{\theta}$ (i.e., $\boldsymbol{\theta}_l$ is a priori independent), then, we can write the posterior as $p(\boldsymbol{\theta} \mid \boldsymbol{u}) \propto p(\boldsymbol{u} \mid \boldsymbol{\theta})p(\boldsymbol{\theta})$, where $p(\boldsymbol{u} \mid \boldsymbol{\theta})$ is the Gaussian likelihood implied by (1). While $p(\boldsymbol{\theta} \mid \boldsymbol{u})$ is formally straightforward, the dimensionality of $\boldsymbol{\theta}$ comprising, among others, of all entries in the weight matrices $\boldsymbol{W}_l$, is such that in practice this expression is computationally intractable. In fact, except for some degenerate cases, e.g., a single-layer BNN with an identity activation function that degenerates into a Bayesian linear regression model, conjugate priors in deep BNNs do not generally exist and the large parameter space makes posterior sampling numerically infeasible. For this reason, variational inference techniques are used to approximate the true posterior and generate forecasts [14], as we show in Section II-C.

B. Physics-Informed Priors

The focus of this work lies in the choices of prior mean $\boldsymbol{\mu} = (\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_L)$ and variance $\boldsymbol{\Sigma} = (\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_L)$ in (2) so that the priors are informed by the physics of the problem. To this end, we assume the existence of a theoretical model in the form of a spatiotemporal PDE for u

$$\frac{\partial u(\boldsymbol{s},t)}{\partial t} + \mathcal{N}[u(\boldsymbol{s},t)] = 0, \quad \boldsymbol{s} \in \Omega, \ t \in [0,T] \tag{3a}$$

$$u(\boldsymbol{s},0) = h_{\mathrm{IC}}(\boldsymbol{s}), \quad \boldsymbol{s} \in \Omega \tag{3b}$$

$$u(\boldsymbol{s},t) = h_{\mathrm{BC}}(\boldsymbol{s},t), \quad \boldsymbol{s} \in \partial\Omega, \ t \in [0,T] \tag{3c}$$

where $\mathcal{N}[\cdot]$ is a (possibly nonlinear) differential spatial operator that may contain derivatives with respect to $\boldsymbol{s}$, (3b) is the initial condition, (3c) is the boundary condition, and $\partial\Omega$ is the boundary of Ω. Our objective is to calibrate the parameters $\boldsymbol{\mu}$ so that a prior for u captures as much information carried by the PDE (3) as possible.

A deep NN, such as (1), can be augmented to yield prior means that are compliant with physical laws [5]. Let $\hat{u}_{\boldsymbol{\mu}}(\boldsymbol{s},t)$ be the estimate of $u(\boldsymbol{s},t)$ as induced by $\boldsymbol{\mu}$, the prior mean of $\boldsymbol{\theta}$ at spatiotemporal location $(\boldsymbol{s},t)$. Well-calibrated prior means are such that $\hat{u}_{\boldsymbol{\mu}}(\boldsymbol{s},t)$ conforms with the physical model (3). Given any input $(\boldsymbol{s},t)$, we can find values of $\boldsymbol{\mu}$ that minimize

$$\hat{f}_{\boldsymbol{\mu}}(\boldsymbol{s},t) := \frac{\partial \hat{u}_{\boldsymbol{\mu}}(\boldsymbol{s},t)}{\partial t} + \mathcal{N}\left[\hat{u}_{\boldsymbol{\mu}}(\boldsymbol{s},t)\right] \tag{4}$$

while also being compliant with the initial conditions in (3b) and boundary conditions in (3c). We discretize the domain over which (3) exists with N_{PDE}, N_{IC}, and N_{BC} points, representing the number of grid points in $\Omega \times [0,T]$, $\Omega \times \{0\}$, and in $\partial\Omega \times [0,T]$, identifying each of the PDE conditions (3a)–(3c), respectively. Then, we can obtain $\boldsymbol{\mu}$ as follows:

$$\begin{aligned}
\boldsymbol{\mu} := \hat{\boldsymbol{\mu}}_{\mathrm{physics}} &= \arg\min_{\boldsymbol{\mu}} \mathrm{mse}(\boldsymbol{\mu}) \\
&= \arg\min_{\boldsymbol{\mu}} \{\mathrm{mse}_{\mathrm{PDE}}(\boldsymbol{\mu}) + \mathrm{mse}_{\mathrm{IC}}(\boldsymbol{\mu}) + \mathrm{mse}_{\mathrm{BC}}(\boldsymbol{\mu})\} \\
&= \arg\min_{\boldsymbol{\mu}} \left\{ \int_{\Omega\times[0,T]} \left|\hat{f}_{\boldsymbol{\mu}}(\boldsymbol{s},t)\right|^2 \mathrm{d}\boldsymbol{s}\ \mathrm{d}t \right. \\
&\qquad + \int_{\Omega} \left|h_{\mathrm{IC}}(\boldsymbol{s}) - \hat{u}_{\boldsymbol{\mu}}(\boldsymbol{s},0)\right|^2 \mathrm{d}\boldsymbol{s} \\
&\qquad \left. + \int_{\partial\Omega\times[0,T]} \left|h_{\mathrm{BC}}(\boldsymbol{s},t) - \hat{u}_{\boldsymbol{\mu}}(\boldsymbol{s},t)\right|^2 \mathrm{d}\boldsymbol{s}\ \mathrm{d}t \right\} \\
&\approx \arg\min_{\boldsymbol{\mu}} \left\{ \frac{1}{N_{\mathrm{PDE}}} \sum_{i=1}^{N_{\mathrm{PDE}}} \left|\hat{f}_{\boldsymbol{\mu}}(\boldsymbol{s}_i,t_i)\right|^2 \right. \\
&\qquad + \frac{1}{N_{\mathrm{IC}}} \sum_{i=1}^{N_{\mathrm{IC}}} \left|h_{\mathrm{IC}}(\boldsymbol{s}_i) - \hat{u}_{\boldsymbol{\mu}}(\boldsymbol{s}_i,0)\right|^2 \\
&\qquad \left. + \frac{1}{N_{\mathrm{BC}}} \sum_{i=1}^{N_{\mathrm{BC}}} \left|h_{\mathrm{BC}}(\boldsymbol{s}_i,t_i) - \hat{u}_{\boldsymbol{\mu}}(\boldsymbol{s}_i,t_i)\right|^2 \right\}.
\end{aligned} \tag{5}$$

In summary, we assign values to the prior means $\boldsymbol{\mu}$ as determined by the minimizers of (5), so that the prior is compliant with the physical model (3). It is important to note that we are not interested in solving the PDE (3). Rather, this formulation yields prior means $\hat{\boldsymbol{\mu}}_{\mathrm{physics}}$ that carry physical meaning, specifically one that expresses actual prior beliefs, i.e., our best guess for the observed behavior of u resembling the physical model (3).

In spite of its apparent similarities to a PDE solver, the prior calibration is fundamentally different. If the primary goal were to solve the PDE, we would look for the best (possibly NN) model, sometimes cast as a function approximator for $u(\boldsymbol{s},t)$, so that $\hat{u}(\boldsymbol{s},t)$ satisfies physical constraints. However, from the perspective of the BNN, that is the model for $u(\boldsymbol{s},t)$, the calibration seeks to find a set of parameters (in our case prior means) given physical constraints.

The prior variance $\boldsymbol{\Sigma}_l$ in (2), $l = 1, \ldots, L$, is a de facto measure of confidence in the prior means $\hat{\boldsymbol{\mu}}_{\mathrm{physics}} = (\hat{\boldsymbol{\mu}}_{1,\mathrm{physics}}, \ldots, \hat{\boldsymbol{\mu}}_{L,\mathrm{physics}})$. Here, we assume $\boldsymbol{\Sigma}_{l,ij} = 0$ for $i \neq j$, i.e., a priori independence for the elements of $\boldsymbol{\theta}_l$ for all l and denote the diagonal of $\boldsymbol{\Sigma}_l$ by $\boldsymbol{\sigma}^2_{l,\mathrm{physics}}$, where $\boldsymbol{\sigma}^2_{\mathrm{physics}} = (\boldsymbol{\sigma}^2_{1,\mathrm{physics}}, \ldots, \boldsymbol{\sigma}^2_{L,\mathrm{physics}})$. For a given $C > 0$, expressing our relative degree of confidence in the theoretical model against the data, and N_l, the number of elements in $\hat{\boldsymbol{\mu}}_{l,\mathrm{physics}}$, we let

$$\boldsymbol{\sigma}^2_{l,\mathrm{physics}} := \left(\frac{C}{N_l}\hat{\boldsymbol{\mu}}_{l,\mathrm{physics}}\right)^2, \quad l \in \{1, \ldots, L\}$$

so that the variance of each $\boldsymbol{\theta}_l$ is relative to its mean $\hat{\boldsymbol{\mu}}_{l,\mathrm{physics}}$ and, from (2), we have

$$p(\boldsymbol{\theta}_l) = \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l) \tag{6a}$$

$$\boldsymbol{\mu}_l = \hat{\boldsymbol{\mu}}_{l,\mathrm{physics}}, \quad \boldsymbol{\Sigma}_l = \mathrm{diag}\left(\boldsymbol{\sigma}^2_{l,\mathrm{physics}}\right), \quad l \in \{1, \ldots, L\}. \tag{6b}$$

If we let $\hat{u}_{\hat{\boldsymbol{\mu}}_{\mathrm{physics}}}(\boldsymbol{s},t)$ be the estimate of $u(\boldsymbol{s},t)$ as induced by $\hat{\boldsymbol{\mu}}_{\mathrm{physics}}$, previous work [11], [26] suggests a possible choice for C determined by a function of the mean squared error between $\hat{u}_{\hat{\boldsymbol{\mu}}_{\mathrm{physics}}}(\boldsymbol{s},t)$ and $u(\boldsymbol{s},t)$. However, such a metric would employ data and the priors would no longer reflect prior beliefs. In general, values of C close to 0 imply a high level of confidence in $\hat{\boldsymbol{\mu}}_{\mathrm{physics}}$ and, by extension, a posterior that will skew in the direction the physical model (3). On the other hand, large values of C grant the likelihood more flexibility to drive posterior inference, rendering the model similar to one with vague priors. In both cases, the uncertainty encoded in the physics-informed priors propagates through the BNN and the influence that the physics places on the posterior, and ultimately the forecasts, is rooted in an established theoretical model.

The formula for the prior variance aims to quantify a subjective issue (i.e., how does one quantify confidence in a PDE?), a challenge compounded by the limited interpretability of BNN parameters. The Bayesian Lasso paper [26], which establishes the equivalence between Bayesian linear regression with Laplace priors and Lasso regression, also proposes the use of hyperpriors as a strategy to adaptively learn prior scales. Following seminal works in economic forecasting [18], [19], [20], domain-informed hierarchical priors have been shown to improve the robustness of uncertainty quantification and facilitate automatic selection of prior variances [27]. Such an approach also helps account for varying coefficient magnitudes across different regions of the parameter space. Nonetheless, the number of parameters in deep learning models is often orders of magnitude larger than in classical parametric models, and implementing full-hierarchical priors may incur prohibitive computational costs. The introduction of additional hyperparameters can also exacerbate issues of identifiability and increase the risk of overfitting, particularly when the amount of data relative to the number of parameters is limited, as is often the case in physics-informed settings. For these reasons, we maintain that the best practical strategy in PIP-BNN is to determine prior variances via cross validation.

The proposed approach can also be extended to allow for PDEs, where the physical parameters need to be estimated. In other terms, we could assume that the physical parameters in the differential operator $\mathcal{N}[\cdot]$ are not fixed to their true value and are instead unknown. In such cases, a natural extension is to incorporate physical parameter uncertainty by estimating their posterior distribution before the prior on the PIP-BNN is calibrated. The supplementary material includes a simulation study where this approach is shown.

C. Posterior Inference

In this work, we approximate the true posterior $p(\boldsymbol{\theta} \mid \boldsymbol{u})$ with variational inference methods, which we summarize here [28], [29], [30]. We chose this posterior estimation approach since it is the most common in the BNN literature (see the review paper in [31]). The main advantage over other methods, mainly Markov chain Monte Carlo (MCMC), is that variational inference casts the estimation of the posterior as an optimization problem for which stochastic gradient descent works well. While MCMC (asymptotically) provides samples from the true posterior, it may pose significant challenges, in terms of runtime and convergence, when the number of parameters in the BNN is large, as also pointed out by Papamarkou et al. [32]. On the other hand, variational inference compromises some amount of accuracy for the sake of computational demand and scalability. Further technical details on the mathematical formulation of the chosen approach are available in the supplementary materials.

We approximate the posterior using a distribution $q(\boldsymbol{\theta};\boldsymbol{\eta}) \in \mathcal{Q}$, where $\mathcal{Q}$ is a family of Gaussian distributions with dimensionality equal to that of $p(\boldsymbol{\theta} \mid \boldsymbol{u})$. The variational parameters are $\boldsymbol{\eta} = (\boldsymbol{\mu}_*, \boldsymbol{\rho}_*)$, where $\boldsymbol{\rho}_* = \log[\exp(\boldsymbol{\sigma}_*) - 1]$ and $q(\boldsymbol{\theta};\boldsymbol{\eta})$ has mean $\boldsymbol{\mu}_*$ and diagonal covariance $\mathrm{diag}(\boldsymbol{\sigma}_*^2)$. We want to find the density $\hat{q}(\boldsymbol{\theta};\hat{\boldsymbol{\eta}}) \in \mathcal{Q}$ that is a suitable approximation for $p(\boldsymbol{\theta} \mid \boldsymbol{u})$, which we accomplish by minimizing the Kullback–Leibler (KL) divergence between the variational and true posterior [33]. The variational approximation $\hat{q}(\boldsymbol{\theta};\hat{\boldsymbol{\eta}})$ also indirectly depends on the observations $\boldsymbol{u}$ through the KL divergence and the shape of $q(\boldsymbol{\theta};\boldsymbol{\eta})$ is determined by $\boldsymbol{\eta}$ [29]. The optimal choice of $q(\boldsymbol{\theta};\boldsymbol{\eta})$, defined as follows:

$$\begin{aligned}
\hat{q}\left(\boldsymbol{\theta};\hat{\boldsymbol{\eta}}\right) &= \arg\min_{q\in\mathcal{Q}} \mathrm{KL}\left(q\left(\boldsymbol{\theta};\boldsymbol{\eta}\right) \,\|\, p\left(\boldsymbol{\theta} \mid \boldsymbol{u}\right)\right) \\
&= \arg\min_{q\in\mathcal{Q}} \int_\Theta q\left(\boldsymbol{\theta};\boldsymbol{\eta}\right) \log\left(\frac{q\left(\boldsymbol{\theta};\boldsymbol{\eta}\right)}{p\left(\boldsymbol{\theta} \mid \boldsymbol{u}\right)}\right) \,\mathrm{d}\boldsymbol{\theta} \\
&= \arg\min_{q\in\mathcal{Q}} \int_\Theta \Bigg[\underbrace{\log q\left(\boldsymbol{\theta};\boldsymbol{\eta}\right)}_{\substack{\text{Variational}\\ \text{distribution}}} - \underbrace{\log p\left(\boldsymbol{\theta}\right)}_{\substack{\text{Physics}\\ \text{prior}}} \\
&\qquad - \underbrace{\log p\left(\boldsymbol{u} \mid \boldsymbol{\theta}\right)}_{\text{Likelihood}} \Bigg] q\left(\boldsymbol{\theta};\boldsymbol{\eta}\right) \,\mathrm{d}\boldsymbol{\theta} \qquad (7)
\end{aligned}$$

where the last equation holds because of the Bayes theorem. The minimization can be achieved to computing gradients of the cost function

$$\begin{aligned}
\mathcal{F}\left(\boldsymbol{\eta},\boldsymbol{u}\right) &:= \int_\Theta \left[\log q\left(\boldsymbol{\theta};\boldsymbol{\eta}\right) - \log p\left(\boldsymbol{\theta}\right) \right. \\
&\qquad \left. - \log p\left(\boldsymbol{u} \mid \boldsymbol{\theta}\right)\right] q\left(\boldsymbol{\theta};\boldsymbol{\eta}\right) \,\mathrm{d}\boldsymbol{\theta} \\
&= \mathrm{KL}\left(q\left(\boldsymbol{\theta};\boldsymbol{\eta}\right) \,\|\, p\left(\boldsymbol{\theta}\right)\right) \\
&\qquad - \int_\Theta \log p\left(\boldsymbol{u} \mid \boldsymbol{\theta}\right) q\left(\boldsymbol{\theta};\boldsymbol{\eta}\right) \,\mathrm{d}\boldsymbol{\theta} \qquad (8)
\end{aligned}$$

with respect to $\boldsymbol{\eta}$. While the first integral in (8) can be computed analytically, as $q(\boldsymbol{\theta};\boldsymbol{\eta})$ and $p(\boldsymbol{\theta})$ are both Gaussian, the second integral is such that minimizing (8) naïvely is computationally demanding [28], [34]. Indeed, standard Monte Carlo approximation, i.e., drawing samples $\boldsymbol{\theta}^{(j)}$ from $q(\boldsymbol{\theta};\boldsymbol{\eta})$ directly, $j = 1,\ldots,E$, leads to a gradient estimator of the likelihood term in (8)

$$\begin{aligned}
&\frac{1}{E}\sum_{j=1}^{E} \log p\left(\boldsymbol{u} \mid \boldsymbol{\theta}\right) \nabla_{q(\boldsymbol{\theta}^{(j)})} \log q\left(\boldsymbol{\theta}^{(j)};\boldsymbol{\eta}\right) \\
&\qquad \approx \nabla_{\boldsymbol{\eta}} \left[\int_\Theta \log p\left(\boldsymbol{u} \mid \boldsymbol{\theta}\right) q\left(\boldsymbol{\theta};\boldsymbol{\eta}\right) \,\mathrm{d}\boldsymbol{\theta}\right]
\end{aligned}$$

with very high variance when E is low and significant computational burden when E is high [35]. Such issues arise since sampling from $q(\boldsymbol{\theta};\boldsymbol{\eta})$ directly would yield samples that depend on $\boldsymbol{\eta}$, with respect to which we are trying to optimize. Here, we estimate $\hat{q}(\boldsymbol{\theta};\hat{\boldsymbol{\eta}})$ using the Bayes By Backprop (BBB) optimization algorithm, an adaptation of stochastic gradient descent applied to variational inference [28], which addresses this issue. We describe the approximate minimization of $\mathcal{F}(\boldsymbol{\eta},\boldsymbol{u})$ in general, i.e., assuming that the first term in (8) may not be computed analytically, even though in our work we use its exact KL divergence.

BBB predicates estimation of $\mathcal{F}(\boldsymbol{\eta},\boldsymbol{u})$ (and ensuing gradients) using an auxiliary variable $\boldsymbol{\epsilon} \sim q(\boldsymbol{\epsilon}) = \mathcal{N}(\mathbf{0},\boldsymbol{I}) \in \mathcal{Q}$. Given any (fixed) value of the variational parameters $\boldsymbol{\eta}^{(i)}$, we sample E realizations from $q(\boldsymbol{\epsilon})$, $\{\boldsymbol{\epsilon}^{(1)},\ldots,\boldsymbol{\epsilon}^{(E)}\}$ and obtain samples from the variational distribution as follows:

$$\boldsymbol{\theta}^{(i,j)} := \boldsymbol{\mu}_*^{(i)} + \log\left(1 + \exp\left(\boldsymbol{\rho}_*^{(i)}\right)\right) \circ \boldsymbol{\epsilon}^{(j)}, \quad j = 1,\ldots,E$$

Algorithm 1 BBB

Given B, M, N_{BBB}, n
$i = 0$
$\boldsymbol{\eta} = (\boldsymbol{\mu}_*, \boldsymbol{\rho}_*)$ {Initialize $\boldsymbol{\eta}$}
for $i < N_{\text{BBB}}$ **do**
  $b = 0$
  for $b < B$ **do**
    Draw $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$
    $\boldsymbol{\theta} := \boldsymbol{\mu}_* + \log[1 + \exp(\boldsymbol{\rho}_*)] \circ \boldsymbol{\epsilon}$ {Reparameterization}
    $\phi(\boldsymbol{\theta}, \boldsymbol{\eta}) = \frac{M}{n}\text{KL}\left(q\left(\boldsymbol{\theta}; \boldsymbol{\eta}\right) \parallel p(\boldsymbol{\theta})\right) - \frac{1}{M}\sum_{j=1}^{M} \log p\left(u_j \mid \boldsymbol{\theta}\right)$
    $\nabla_{\boldsymbol{\mu}_*}\phi = \frac{\partial\phi(\boldsymbol{\theta},\boldsymbol{\eta})}{\partial\boldsymbol{\theta}} + \frac{\partial\phi(\boldsymbol{\theta},\boldsymbol{\eta})}{\partial\boldsymbol{\mu}_*}$ {Proposition S1}
    $\nabla_{\boldsymbol{\rho}_*}\phi = \frac{\partial\phi(\boldsymbol{\theta},\boldsymbol{\eta})}{\partial\boldsymbol{\theta}} \frac{\boldsymbol{\epsilon}}{1+\exp(-\boldsymbol{\rho}_*)} + \frac{\partial\phi(\boldsymbol{\theta},\boldsymbol{\eta})}{\partial\boldsymbol{\rho}_*}$ {Proposition S1}
    $\boldsymbol{\gamma} = (\nabla_{\boldsymbol{\mu}_*}, \nabla_{\boldsymbol{\rho}_*})$ {Unbiased MC gradients}
    $\boldsymbol{\eta} = \boldsymbol{\eta} - \alpha\boldsymbol{\gamma}$ {α is the learning rate}
    $b = b + 1$
  end for
  $i = i + 1$
end for
return $\hat{\boldsymbol{\mu}}_*$, $\hat{\boldsymbol{\rho}}_*$

where "$\circ$" indicates the pointwise multiplication. We then write the approximation

$$\widehat{\mathcal{F}\left(\boldsymbol{\eta}^{(i)}, \boldsymbol{u}\right)} = \left\{ \frac{1}{E}\sum_{j=1}^{E} \left[\log q\left(\boldsymbol{\theta}^{(i,j)}; \boldsymbol{\eta}^{(i)}\right) - \log p\left(\boldsymbol{\theta}^{(i,j)}\right) - \log p\left(\boldsymbol{u} \mid \boldsymbol{\theta}^{(i,j)}\right)\right] \right\} = \left\{ \frac{1}{E}\sum_{j=1}^{E} \phi\left(\boldsymbol{\theta}^{(i,j)}, \boldsymbol{\eta}^{(i)}, \boldsymbol{u}\right) \right\}. \tag{9}$$

Since the function $\phi(\cdot)$ is differentiable and $\boldsymbol{\theta}^{(i,j)}$ is not drawn directly from the variational distribution $q(\boldsymbol{\theta}; \boldsymbol{\eta})$, one can show (proof available in the supplementary material) that $\nabla_{\boldsymbol{\eta}}\mathbb{E}_{q(\boldsymbol{\theta};\boldsymbol{\eta})}[\phi(\boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{u})] = \mathbb{E}_{q(\boldsymbol{\epsilon})}[\nabla_{\boldsymbol{\eta}}\phi(\boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{u})]$, so we can use the gradient estimator

$$\nabla_{\boldsymbol{\eta}}\widehat{\mathcal{F}\left(\boldsymbol{\eta}^{(i)}, \boldsymbol{u}\right)} = \left(\frac{1}{E}\sum_{j=1}^{E} \nabla_{\boldsymbol{\eta}}\phi\left(\boldsymbol{\theta}^{(i,j)}, \boldsymbol{\eta}^{(i)}, \boldsymbol{u}\right) \right)$$

which is unbiased for $\nabla_{\boldsymbol{\eta}}\mathcal{F}(\boldsymbol{\eta}^{(i)}, \boldsymbol{u})$ [28], to update the variational parameters $\boldsymbol{\eta}$ until some convergence criterion is met. The minimization of the original KL divergence in (7) is now reduced to computing gradients of a differentiable function with respect to $\boldsymbol{\eta}$, as in the case of stochastic gradient descent for standard NNs.

We also set $E = 1$ and divide the n observations in $\boldsymbol{u}$ into B mini-batches, where each mini-batch $\boldsymbol{u}_b$ contains M observations (i.e., assume for simplicity $n = MB$), so that $\boldsymbol{u}_b = \{u_{b,1}, \ldots, u_{b,M}\}$ and $\boldsymbol{u} = \{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_B\}$, which is a common strategy to achieve faster convergence while lessening the computational burden [34]. The mini-batch optimization implemented in this work computes (9) for each $\boldsymbol{u}_b$ as follows:

$$\hat{\mathcal{F}}\left(\boldsymbol{\eta}^{(i)}, \boldsymbol{u}_b\right) = \frac{M}{n}\text{KL}\left(q\left(\boldsymbol{\theta}; \boldsymbol{\eta}^{(i)}\right) \parallel p(\boldsymbol{\theta})\right) - \log p\left(\boldsymbol{u}_b \mid \boldsymbol{\theta}^{(j)}\right)$$

so that the variational parameters are updated BN_{BBB} times during the optimization, where N_{BBB} denotes the total number of iterations. The entire optimization procedure is summarized in Algorithm 1 and further details on the KL divergence are available in the supplementary material.

The optimization yields the Gaussian distribution $\hat{q}(\boldsymbol{\theta}; \hat{\boldsymbol{\eta}})$ that balances the physics-informed priors and the likelihood, resulting in an approximation of the posterior that is informed by the theoretical and observed behavior of $u(\boldsymbol{s}, t)$. From the estimated $\hat{\boldsymbol{\eta}} = (\hat{\boldsymbol{\mu}}_*, \hat{\boldsymbol{\rho}}_*)$, we retrieve $\hat{\boldsymbol{\sigma}}_*^2 = \log(\exp(\hat{\boldsymbol{\rho}}_*) + 1)^2$ and have

$$p\left(\boldsymbol{\theta} \mid \boldsymbol{u}\right) \approx \hat{q}\left(\boldsymbol{\theta}; \hat{\boldsymbol{\eta}}\right) = \mathcal{N}\left(\hat{\boldsymbol{\mu}}_*, \text{diag}\left(\hat{\boldsymbol{\sigma}}_*^2\right)\right)$$

so that we can compute and sample from the posterior predictive distribution as follows:

$$\begin{aligned} p\left(\mu_u\left(\boldsymbol{s}, t\right) \mid \boldsymbol{u}\right) &= \int_{\Theta} p\left(\mu_u\left(\boldsymbol{s}, t\right) \mid \boldsymbol{\theta}\right) p\left(\boldsymbol{\theta} \mid \boldsymbol{u}\right) \, \mathrm{d}\boldsymbol{\theta} \\ &\approx \int_{\Theta} p\left(\mu_u(\boldsymbol{s}, t) \mid \boldsymbol{\theta}\right) \hat{q}\left(\boldsymbol{\theta}; \hat{\boldsymbol{\eta}}\right) \, \mathrm{d}\boldsymbol{\theta} \\ &\approx \frac{1}{M}\sum_{i=1}^{M} p\left(\mu_u(\boldsymbol{s}, t) \mid \boldsymbol{\theta}^{(i)}\right), \quad \boldsymbol{\theta}^{(i)} \sim \hat{q}\left(\boldsymbol{\theta}; \hat{\boldsymbol{\eta}}\right) \end{aligned} \tag{10}$$

and likewise from $p(\sigma_u^2(\boldsymbol{s}, t) \mid \boldsymbol{u})$, yielding physics-informed forecasts $\mathbb{E}(\mu_u(\boldsymbol{s}, t) \mid \boldsymbol{u})$ and physics-informed variance $\mathbb{E}(\sigma_u^2(\boldsymbol{s}, t) \mid \boldsymbol{u})$ [28]. The prior means $\hat{\boldsymbol{\mu}}_{\text{physics}}$ and prior variance $\boldsymbol{\sigma}_{\text{physics}}^2$ thus influence the posterior mean and posterior variance, yielding physics-informed forecasts and physics-informed variance.

D. Comparison With Other Methods

Our PIP-BNN model is a general approach to incorporating physical knowledge into an NN via a Bayesian framework. As such, it is not an alternative to, nor an improvement of the existing physics-informed methodologies discussed in Section I. Our proposed approach casts the PINN paradigm as a statistical problem, akin to how *The Bayesian Lasso* [26] demonstrated the parallel between Lasso regression and Bayesian linear regression with Laplace priors. We highlight the differences between PIP-BNN and the existing literature by focusing on the PINN approach introduced in [5] and the B-PINN model in [17], as they are the two methods from which PIP-BNN draws inspiration.

Standard PINNs, which are adopted in [1], [2], [3], [8], [9], and [10] predicate the use of a physics-augmented objective function, e.g., "Loss = $\text{Loss}_{\text{Data}} + \lambda\text{Loss}_{\text{PDE}}$," to estimate the parameters of a deep NN. This frequentist framework does not directly address uncertainty quantification, which PIP-BNN does by design. B-PINN is instead a Bayesian model inspired by PINN, but is aimed at solving the PDE rather than tradingoff the information from observations and physics. In

addition, B-PINN assumes that noisy observations of the PDE solution, boundary conditions, and the forcing term must be all available and that the noise in the data is known. In other words, PIP-BNN assumes a BNN (1) and aims at finding a set of parameters (in our case prior means) given physical constraints. B-BNN is instead looking for the best NN model so that $\hat{u}(\boldsymbol{s}, t)$ satisfies the physical constraints of the PDE.

In spite of the aforementioned fundamental differences, PIP-BNN shares enough similarities with standard PINNs (the incorporation of a PDE in the inferential process) and B-PINN (parameter uncertainty), that a comparison with each method is warranted. The results can be found in the supplementary material.

III. Simulation Study

We conduct a simulation study in a system that is fully explained by a physical model. In Section III-A, we report the simulation details. In Section III-B, we assess the performance of our general Bayesian framework with physics-informed priors against the noninformed alternative. In Section III-C, we conduct sensitivity analyses by considering different scenarios of data availability as well as specifications of the physical model.

A. Simulation Details

1) Governing PDE: We simulate data from a 1-D Burgers' equation, a fundamental nonlinear PDE that is often used as a prototype for conservation laws in compressible fluids [36]. Formally, the spatiotemporal process $u(\boldsymbol{s}, t)$ is the velocity of a fluid as a function of space and time, where $d = 1$ so that $\boldsymbol{s} = x$

$$(x, t) \in \Omega \times [0, T] = [-1, 1] \times [0, 2] \tag{11}$$

and the governing model is

$$\frac{\partial u(x,t)}{\partial t} + u(x,t)\frac{\partial u(x,t)}{\partial x} = \nu \frac{\partial^2 u(x,t)}{\partial x^2} \tag{12a}$$

$$u(x, 0) = -\sin(\pi x) \tag{12b}$$

$$u(-1, t) = u(1, t) = 0 \tag{12c}$$

where the nonlinear differential operator from (3) is $N[u(x,t)] = u(x,t)((\partial u(x,t))/\partial x) - \nu((\partial^2 u(x,t))/\partial x^2)$. The viscosity $\nu \in \{0.01, 0.05\}$ measures a fluid's resistance to flow. The initial condition (12b) describes a sinusoidal shock at $t = 0$ and the Dirichlet boundary condition (12c) enforces the resting velocity at the bounds of Ω. Fig. 1 shows the solution to Burgers' equation with a viscosity of $\nu = 0.05$, which will be used for the main comparison of the physics-informed approach with the noninformed alternative.

2) Data: We obtained solutions to Burgers' equation using the *PhiFlow* open-source simulation software (publicly available at github.com/tumpbs/PhiFlow). We discretized the domain (11) with an $N_s \times N_t$ grid, where $N_s = 100$ are points equally spaced in the spatial domain $[-1, 1]$ and $N_t = 1000$ are points equally spaced in the temporal domain $[0, 2]$, creating a grid of $N = N_s \times N_t$ points (as shown in Fig. S1 of the supplementary material), and solved for $u^*(x_i, t_i)$, $i \in \{1, \ldots, N\}$, where (x_i, t_i) are elements of the grid. In order to have more realistic data, we add Gaussian noise with variance $\sigma_D^2 = 0.1^2$

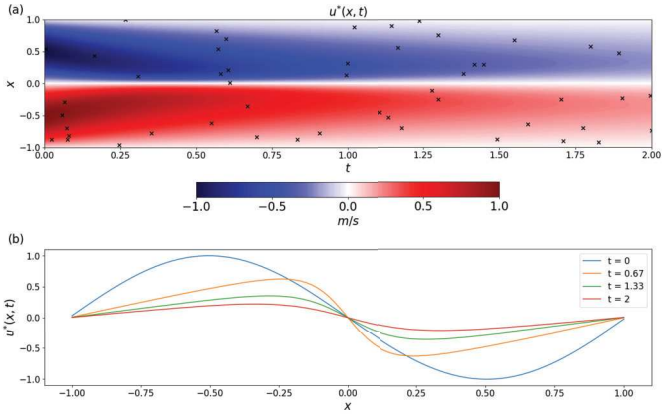

Fig. 1. Solution $u^*(x,t)$ to Burgers' equation in (12), with $\nu = 0.05$. (a) $u^*(x,t)$ throughout the domain (11) and the n spatiotemporal locations chosen for training the model. (b) $u^*(x,t)$ at four cross sections of $t \in \{0, 0.67, 1.33, 2\}$. As time increases, the initial sinusoidal shock dissipates and the fluid reverts back to its resting state.

$$u(x_i, t_i) = u^*(x_i, t_i) + \epsilon_i, \quad \epsilon_i \overset{\text{i.i.d.}}{\sim} N\left(0, \sigma_D^2\right) \tag{13}$$

so that $u(x_i, t_i)$ are N noisy observations from Burgers' equation model. We subsample $n = 50$ data points uniformly in $\{u(x_1, t_1), \ldots, u(x_N, t_N)\}$, shown in black in Fig. 1, and we assume that we observe $\boldsymbol{u} = \{u(x_1, t_1), \ldots, u(x_n, t_n)\}$ to predict the spatiotemporal process $u(x,t)$ across the entire domain of Burgers' equation using our deep BNN model.

3) Physics-Informed Priors: As outlined in Section II, the first step to implement our approach is to calibrate the prior distribution to a BNN using the physical model (12). We consider a BNN comprised of $L = 4$ hidden layers, $k = 10$ nodes in each layer, and a hyperbolic tangent activation function $g(\cdot) = \tanh(\cdot)$. We employ the same $N_s \times N_t$ discretization grid of the domain (11) to assign prior means for $p(\boldsymbol{\theta})$. Denote with $N_{\text{PDE}} = N_s \times N_t$ the grid for the domain of (12a), $N_{\text{IC}} = N_s$ the grid for the domain of (12b), and $N_{\text{BC}} = 2N_t$ the grid for the domain of (12c), for a total of $N_{\text{PDE}} + N_{\text{IC}} + N_{\text{BC}}$ calibration points. We find $\hat{\boldsymbol{\mu}}_{\text{physics}}$ by minimizing

$$\begin{aligned}
\text{mse}(\boldsymbol{\mu}) \approx & \frac{1}{N_{\text{PDE}}} \sum_{i=1}^{N_{\text{PDE}}} \left|\hat{f}_{\boldsymbol{\mu}}(x_i, t_i)\right|^2 \\
& + \frac{1}{N_{\text{IC}}} \sum_{i=1}^{N_{\text{IC}}} \left|\sin(\pi x_i) + \hat{u}_{\boldsymbol{\mu}}(x_i, 0)\right|^2 \\
& + \frac{1}{N_{\text{BC}}} \sum_{i=1}^{N_{\text{BC}}} \left|\hat{u}_{\boldsymbol{\mu}}(-1, t_i)\right|^2 \\
& + \frac{1}{N_{\text{BC}}} \sum_{i=1}^{N_{\text{BC}}} \left|\hat{u}_{\boldsymbol{\mu}}(1, t_i)\right|^2 \\
\hat{f}_{\boldsymbol{\mu}}(x_i, t_i) \; := & \; \frac{\partial \hat{u}_{\boldsymbol{\mu}}(x_i, t_i)}{\partial t} + \hat{u}_{\boldsymbol{\mu}}(x_i, t_i) \frac{\partial \hat{u}_{\boldsymbol{\mu}}(x_i, t_i)}{\partial x} \\
& - \nu \frac{\partial^2 \hat{u}_{\boldsymbol{\mu}}(x_i, t_i)}{\partial x^2}.
\end{aligned}$$

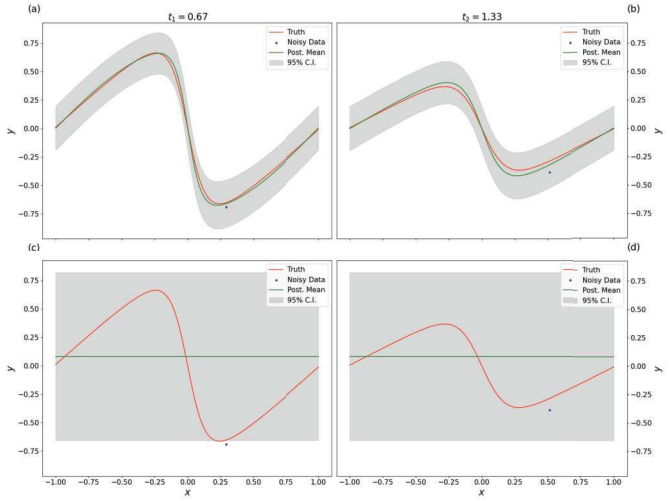

Fig. 2. (a) and (b) Comparison between physics-informed and (c) and (d) nonphysics-informed predictions using Burgers' equation at $t_1 = 0.67$ and $t_2 = 1.33$ (two of the same times from Fig. 1). (a) and (b) Simulation results using priors informed by Burgers' equation (12). The domain-wide median squared bias and variance across the $N_{\text{sim}} = 100$ simulations are 0.15×10^{-2} and 1.08×10^{-2}, respectively. (c) and (d) Results using a noninformed prior, now with $\boldsymbol{\mu} = \mathbf{0}$. In this case, the median squared bias and variance are 15.69×10^{-2} and 16.73×10^{-2}, respectively.

For simplicity, we set the prior variance as $\boldsymbol{\Sigma}_l = \sigma^2_{\text{physics}} \boldsymbol{I}$, with $\sigma^2_{\text{physics}} = 0.0025^2$, for every l. The sensitivity of our approach's performance with respect to n, $\sigma^2_{\text{physics}}$, σ^2_D, the initial and boundary conditions, and the viscosity ν is addressed in Section III-C.

4) Bayesian Inference: Using the $n = 50$ observations in the vector $\boldsymbol{u}$, we performed the BBB optimization (Algorithm 1) with $N_{\text{BBB}} = 10\,000$ iterations to approximate the true posterior. The choice to rely on a small number of observations n is intentional, as our approach is most effective for applications where data may be scarce. As such, using too many observations may defeat the purpose of relying on physics-informed priors. We performed $N_{\text{sim}} = 100$ different simulations, each corresponding to a different realization of (13).

5) Performance Metrics: We measure each simulation's performance by reporting the mean squared error (MSE), defined as the sum of the squared bias between the posterior mean and ground truth and the average posterior variance

$$\text{MSE} = \frac{1}{N_s \times N_t} \sum_{i=1}^{N_s} \sum_{j=1}^{N_t} \left\{ \underbrace{\left[\mathbb{E}\left(\mu_u\left(x_i, t_j\right) \mid \boldsymbol{u}\right) - u^*\left(x_i, t_j\right)\right]^2}_{\text{Bias}^2} + \underbrace{\mathbb{E}\left[\sigma_u^2\left(x_i, t_j\right) \mid \boldsymbol{u}\right]}_{\text{Variance}} \right\} \tag{14}$$

where $\mathbb{E}(\mu_u(x_i, t_j) \mid \boldsymbol{u})$ and $\mathbb{E}(\sigma_u^2(x_i, t_j) \mid \boldsymbol{u})$ are obtained as in (10).

B. Comparing Physics-Informed and Noninformed Priors

Table I shows summary statistics across the N_{sim} simulations, reporting the median metrics from the end of Section III-A5 and their interquantile ranges. We just report the median computational demand, since it is practically the same across the N_{sim} simulations. In the physics-informed case, we let the prior means be informed by Burgers' equation, $\mathbb{E}(\boldsymbol{\theta}) = \hat{\boldsymbol{\mu}}_{\text{physics}}$ from (5) and $\boldsymbol{\Sigma} = \sigma^2_{\text{physics}} \boldsymbol{I}$, while the noninformed priors assume $\mathbb{E}(\boldsymbol{\theta}) = \mathbf{0}$ and $\boldsymbol{\Sigma} = \boldsymbol{I}$. A visual illustration of the difference in performance is shown in Fig. 2, where we show forecasts at two time points for one of the simulations. (a) and (b) show simulation results using the physics-informed priors (PIP-BNN), while (c) and (d) show the same results using noninformative priors.

TABLE I

PHYSICS-INFORMED VERSUS NONINFORMED PRIORS SIMULATION USING $n = 50$ OBSERVATIONS OF BURGERS' EQUATION (12) WITH VISCOSITY $\nu = 0.05$. DISPLAYED ARE THE MEDIAN AND INTERQUARTILE RANGES ACROSS $N_{\text{SIM}} = 100$ DIFFERENT REALIZATIONS FOR THE BIAS, VARIANCE, AND MSE AS DEFINED IN (14). THE COMPUTATIONAL DEMAND DOES NOT MATERIALLY CHANGE ACROSS THE SIMULATIONS, SO WE JUST REPORT THE MEDIAN VALUE

	Physics-Informed Priors	Non-Informed Priors
Bias^2 (10^{-2})	0.15 (0.12, 0.19)	15.69 (15.54, 15.94)
Variance (10^{-2})	1.08 (0.91, 1.26)	16.73 (14.11, 18.47)
MSE (10^{-2})	1.25 (1.06, 1.42)	32.36 (29.81, 34.34)
Runtime (mm:ss)	02 : 54	02 : 58

The physics-informed model is able to retrieve a good approximation for the true solution (as shown in red in Fig. 2) with a median squared bias of 0.15×10^{-2}, compared with an increase of almost two orders of magnitude across the noninformed simulations (15.69×10^{-2}) for the noninformed case. This model-based approach provides informed prior beliefs, while the scattered data points steer the posterior in the direction of observed behavior. As shown in Fig. 2(a) and (b), this knowledge propagates to the posterior yielding an improved forecast. Fig. 2(c) and (d) instead illustrates how the noninformed forecast is essentially a flat line.

In addition, even though only $n = 50$ observations were employed in the Bayesian update, the credibility intervals around the posterior mean (in gray in Fig. 2) are small relative to ground truth, as the median posterior variance across the domain is 1.08×10^{-2}, whereas in the noninformed case it rises by an order of magnitude (16.73×10^{-2}). Notably, the median posterior variance in the physics-informed simulations resembles the amount of Gaussian noise we added, as 1.08×10^{-2} is approximately $\sigma^2_D = 0.1^2$. The posterior variance is a function of the confidence in the physical model, in this case, $\sigma^2_{\text{physics}} = 0.0025^2$, and the noise in the data, which we set at $\sigma^2_D = 0.1^2$, and in Section III-C, we perform a sensitivity analysis with respect to these quantities.

C. Sensitivity Analysis

We conduct a sensitivity analysis to understand our approach's performance as we change data availability or the viscosity in Burgers' equation. For each scenario, we perform $N_{\text{sim}} = 20$ different simulations and report the median MSE as defined in (14) as well as the interquantile ranges. We report the median computational demand, since it does not change

TABLE II

SENSITIVITY ANALYSES OF THE PHYSICS-INFORMED PRIOR WITH RESPECT TO (A) NUMBER OF AVAILABLE OBSERVATIONS, (B) TEMPORAL SAMPLING, (C) SPATIAL SAMPLING, (D) AMOUNT OF GAUSSIAN NOISE ADDED, (E) CONFIDENCE IN THE PHYSICAL MODEL, AND (F) INITIAL AND BOUNDARY CONDITIONS AND THE VISCOSITY IN THE PDE. DISPLAYED ARE THE MEDIAN AND INTERQUARTILE RANGES ACROSS $N_{\text{SIM}} = 20$ DIFFERENT REALIZATIONS FOR THE BIAS, VARIANCE, AND MSE AS DEFINED IN (14). THE COMPUTATIONAL DEMAND DOES NOT MATERIALLY CHANGE ACROSS THE SIMULATIONS, SO WE JUST REPORT THE MEDIAN VALUE

(a)	$n = 2$	$n = 250$
Bias^2 (10^{-2})	3.05 (0.98, 10.88)	0.12 (0.11, 0.13)
Variance (10^{-2})	0.03 (0.02, 0.10)	1.16 (1.07, 1.22)
MSE (10^{-2})	3.18 (1.01, 10.91)	1.28 (1.19, 1.33)
Runtime (mm:ss)	00 : 53	12 : 13

(b)	$t \in [0, 0.25]$	$t \in [0.6, 0.90]$
Bias^2 (10^{-2})	0.22 (0.19, 0.25)	0.17 (0.14, 0.20)
Variance (10^{-2})	1.00 (0.87, 1.27)	1.02 (0.80, 1.12)
MSE (10^{-2})	1.22 (1.11, 1.46)	1.20 (1.02, 1.30)
Runtime (mm:ss)	04 : 16	04 : 10

(c)	$x \in \{0, 0.5, 1\}$	$x \in \{-1, -0.5, 0, 0.5, 1\}$
Bias^2 (10^{-2})	3.28 (1.30, 7.66)	0.18 (0.13, 0.23)
Variance (10^{-2})	2.01 (1.24, 10.49)	0.99 (0.80, 1.17)
MSE (10^{-2})	11.00 (4.58, 16.81)	1.18 (1.00, 1.39)
Runtime (mm:ss)	03 : 24	03 : 58

(d)	$\sigma_D^2 = 0.5^2$	$\sigma_D^2 = 0.001^2$
Bias^2 (10^{-2})	1.06 (0.59, 1.55)	0.12 (0.11, 0.13)
Variance (10^{-2})	23.95 (21.48, 28.87)	0.11 (0.09, 0.14)
MSE (10^{-2})	26.49 (22.47, 30.60)	0.23 (0.21, 0.26)
Runtime (mm:ss)	04 : 10	04 : 10

(e)	$\sigma^2_{\text{physics}} = 0.01^2$	$\sigma^2_{\text{physics}} = 0.0001^2$
Bias^2 (10^{-2})	0.15 (0.13, 0.18)	0.14 (0.13, 0.17)
Variance (10^{-2})	1.34 (1.24, 1.51)	1.06 (0.98, 1.23)
MSE (10^{-2})	1.47 (1.43, 1.71)	1.19 (1.13, 1.40)
Runtime (mm:ss)	04 : 11	04 : 09

(f)	$u(x, 0) = \cos(\pi x)$ $u(-1, t) = u(1, t)$	$\nu = 0.01$
Bias^2 (10^{-2})	0.16 (0.14, 0.19)	0.19 (0.18, 0.26)
Variance (10^{-2})	1.10 (0.93, 1.22)	1.15 (1.05, 1.35)
MSE (10^{-2})	1.29 (1.10, 1.50)	1.38 (1.24, 1.65)
Runtime (mm:ss)	03 : 54	03 : 50

materially across the N_{sim} simulations. All of the simulations in the sensitivity study assume physics-informed priors.

1) Sample Size, Spatial and Temporal Sampling, and Noise: We perform the same experiment with less and more observations ($n = 2$ and 250, as opposed to $n = 50$). We also study the effects of spatial and temporal subsampling. Finally, we test our approach for different amounts of Gaussian noise, $\sigma_D^2 = 0.5^2$ and $\sigma_D^2 = 0.001^2$, added to the solutions to Burgers' equation. Beyond the sensitivity of our model performance, these simulations could replicate several scenarios of data availability such as the volume of collectable observations, the locations where sensors may be placed, or the reliability of those sensors. In Table II, we consider differing numbers of observations n for the Bayesian update [see (a)], restrictions on the temporal domain in which observations are available [see (b)], restrictions on the spatial domain in which observations are available [see (c)], and magnitude of the Gaussian noise σ_D^2 [see (d)].

As apparent from Table II (a), if $n = 250$ observations are available, we observe a very low median bias but a larger median posterior variance compared to the scenario with $n = 2$. When the number of observations is low, the Gaussian noise's influence on the posterior variance is minimal and the prior confidence in the PDE, $\sigma^2_{\text{physics}} = 0.0025^2$, is the main determinant of the posterior variance. When instead $n = 250$, the posterior variance mimics the magnitude of the Gaussian noise σ_D^2, as in the physics-informed results of Table I. The simulations in Table II (b) assume observations over a selected subset of the temporal domain, while the simulations in Table II (c) assume that one can collect $n = 50$ observations at just three and five locations across the spatial domain. The bias and variance in (b) are comparable to those of Table I, which suggests that temporal sampling does not affect our approach's forecasting ability. (c) instead shows a notable increase in bias when we restrict the spatial domain to just three locations compared with five, suggesting that collecting many observations at scattered times may be preferable to collecting many observations in the same locations over time (visual evidence is in Figs. S2 and S3 of the supplementary material). In Table II(d), we also see how increasing the Gaussian noise added to the solutions to Burgers' equation by a factor of 5 still results in a good performance for the physics-informed prior, as it allows the posteriors to overcome the comparatively low signal-to-noise ratio in the data, yielding forecasts with low bias. At the same time, the amount of noise in the data propagates through the model, yielding a large posterior variance, as is apparent in Figs. S4 and S5 of the supplementary material. Overall, Table II(a)–(d) shows that, if the Bayesian priors are physics-informed and data are available at several spatial locations, our approach does not require a large number of observations to produce physically consistent forecasts.

2) Prior Confidence in the PDE, Initial and Boundary Conditions, and Viscosity: Table II(e) and (f) shows results for different levels of prior confidence in the PDE as well as different initial and boundary conditions and viscosity. We test our approach under different levels of confidence in the PDE, $\sigma^2_{\text{physics}} = 0.01^2$ and $\sigma^2_{\text{physics}} = 0.0001^2$, a greater and smaller prior variance than in the simulations discussed in Section III-B in Table II(e). We also consider a less viscous fluid, $\nu = 0.01$, and a specification of Burgers' equation with initial condition $u(x, 0) = \cos(\pi x)$ and a periodic boundary condition, as shown in Table II(f).

Higher levels of confidence in the physical model, i.e., lower prior variance, translate to more accurate forecasts and lower levels of uncertainty, and vice versa, as we show in Table II(e). When the prior variance is very low, our approach yields the smallest posterior bias, with an interquartile range of (0.13, 0.17), visually illustrated in Fig. S6 of the supplementary material. On the other hand, lesser confidence in the physical model, in this case by a factor of 16, implies that a greater

level of uncertainty propagates through the model, yielding greater posterior bias and variance. Table II(f) shows the results from simulations whose priors are calibrated with respect to different initial and boundary conditions or viscosity, yielding a similar performance to the main simulation results we report in Table I, as also shown in Figs. S7 and S8 of the supplementary material.

D. Additional Sensitivity Analysis

We also tested our PIP-BNN model for different NN architectures and sample sizes. Our results suggest that model architecture does not materially affect predictive ability, while runtime, expectedly, increases as the amount of parameters increases. Similarly, when we tested PIP-BNN for different sample sizes, we observed that the predictive ability remained relatively stable, while runtime increased significantly. Further details on these simulations are available in Section S3 of the supplementary material.

IV. Application to Boundary Layer Flow

We apply our approach to model and forecast boundary layer velocity from a high-resolution experimental dataset from a water tunnel [37], [38]. Boundary layer velocity presents highly nonlinear patterns, in time and space, described by the 3-D NS equations. Forecasting boundary layers presents considerable challenges, as the velocity field interacts with surfaces, and is one of the most active areas of research in fluid mechanics [39], [40], with applications in drag optimization along a moving object in fields such as aerospace, civil, or mechanical engineering. The section proceeds as follows. Section IV-A presents the data. Section IV-B describes the physical model used to calibrate the physics-informed priors and the associated calibration of the prior mean. Section IV-C presents the results.

A. Data

The dataset used in this work is part of an experimental campaign to study boundary layer velocity focused near the bottom wall of a water tunnel. Data were collected in a cross section, wall-normal plane parallel to the direction of the flow. The region of interest, in the middle of the 8-m base, has a length of about 20 cm along the wall, a height of about 5 cm, and a temporal resolution of 0.001 s over a total period of 5.004 s [37], [38]. The velocity data was obtained via particle image velocimetry (PIV), the noninvasive optical measurement technique of choice when fluid mechanics experiments are performed with an unobstructed view of the flow [41]. PIV takes multiple laser-fluoresced images of a 2-D area of interest, spaced a known time interval apart, and calculates the fluid velocity on a Cartesian grid via statistical correlation of the seed particle locations between image pairs [42], [43], [44], [45].

The data are comprised of the instantaneous velocity in the streamwise and wall-normal directions (not in the third, spanwise dimension), $u(\boldsymbol{s}_i, t)$ and $v(\boldsymbol{s}_i, t)$, respectively, in an $N = N_x \times N_y$ rectangular grid, $N_x = 503$ and $N_y = 124$, where $\boldsymbol{s}_i = (x_i, y_i)$ denotes the spatial coordinates $\{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_N\}$ and

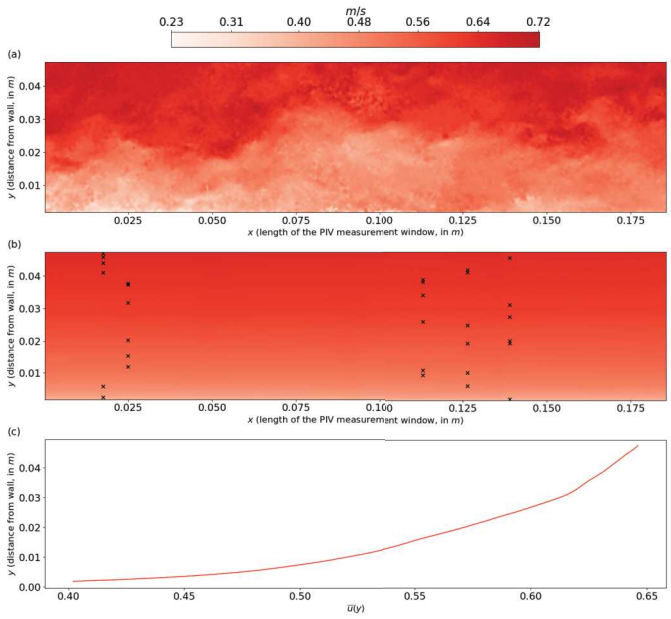

Fig. 3. Boundary layer data collected in a rectangular window inside an experimental water tunnel. (a) Single snapshot of the instantaneous streamwise velocity (going from left to right), namely, $t = 1000$, $\boldsymbol{s}_i \in \{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_N\}$. (b) Streamwise velocity $\overline{u}(\boldsymbol{s}_i)$, after averaging across the $T = 5004$ time steps. The pattern is stratified from the wall ($x \approx 0$) to the top of the domain, where $\overline{u}$ approaches the free-stream, i.e., undisturbed by the wall, velocity which the experimental set at 0.67 m/s. The points are a subsample of size $n = 30$ used to predict the entire domain. (c) Average vertical profile of $\overline{u}$, i.e., after averaging out x, showing the behavior of the streamwise velocity in a fully developed turbulent boundary layer.

$t = 1, \ldots, T = 5004$. We decompose the instantaneous velocity at spatial location $\boldsymbol{s}_i$ and time t into its expectation (time trend) and fluctuation and average it over the T time steps. In the case of the streamwise velocity, we have

$$\frac{1}{T}\sum_{t=1}^{T} u(\boldsymbol{s}_i, t) = \frac{1}{T}\sum_{t=1}^{T} \left\{ \underbrace{\overline{u}(\boldsymbol{s}_i)}_{\text{time trend}} + \underbrace{u'(\boldsymbol{s}_i, t)}_{\text{fluctuation}} \right\} = \overline{u}(\boldsymbol{s}_i)$$

since the average of the fluctuations over time is 0 by definition. The manipulation for the instantaneous wall-normal velocity $v(\boldsymbol{s}, t)$ is similar. Since the data collection region is sufficiently downstream in the water tunnel, the flow can be considered fully developed—i.e., streamwise gradients in the mean velocities are approximately 0. We also assume that there are no gradients in the mean velocities in the spanwise direction. With these assumptions, mass conservation, which requires a divergence-free mean velocity for an incompressible flow, implies $\partial\overline{v}/\partial y \approx 0$. At the lower wall, $\overline{v}(0, y) = 0$ m/s, so that $\overline{v}(\boldsymbol{s}_i) = 0$, $i \in \{1, \ldots, N\}$. Thus, we expect that the only nonzero mean velocity component $\overline{u}$ is approximately only a function of the wall-normal coordinate y. Note, however, that for generality, we retain an explicit dependence of $\overline{u}$ on both x and y [i.e., $\overline{u}(x, y) = \overline{u}(\boldsymbol{s}_i)$]. Diagnostics are available in Figs. S14 and S15 of the supplementary material.

In summary, we consider $\overline{u}(\boldsymbol{s}_i)$, the average velocity in the streamwise direction across the 5004 time steps. We subsample observations from five vertical cross sections and only keep 5% of them from each section, as shown in black in Fig. 3(b),

so the data used to predict the entire spatial domain are now $\boldsymbol{u} := \overline{\boldsymbol{u}} = \{\overline{u}(\boldsymbol{s}_1), \ldots, \overline{u}(\boldsymbol{s}_n)\}$, $n = 30$, which amounts to less than 0.05% of the N total observations.

Fig. 3(a) shows one of the $T = 5004$ snapshots of the instantaneous streamwise velocity going from left to right, in this case $u(\boldsymbol{s}, 1000)$. The turbulent nature of this boundary layer is clearly shown in the large variance of the gradients in both the streamwise and wall-normal directions. Fig. 3(b) shows the average streamwise velocity $\overline{u}(\boldsymbol{s})$, where the spatial coordinates x and y, both measured in meters, represent the base and height of the PIV measurement window, respectively. (c) shows the average (in both time and x) vertical profile of $\overline{u}$ which varies from 0.4 m/s near the wall to 0.65 m/s at the top of the observable domain. As we move away from the surface, the flow's velocity approaches its free-stream profile, i.e., undisturbed by the wall, which the experiment set at 0.67 m/s. The lowest recorded velocity being 0.4 m/s (close to but not at the wall) is consistent with boundary layer theory, as the gradient of $\overline{u}$ with respect to the wall-normal direction is largest near the wall, and we would expect measurements of $\overline{u}$ at $y = 0$, if recordable, to be 0 m/s [46]. Since the flow moves from left to right, time-averaging smooths out most of the vertical perturbations and the wall-normal component becomes negligible [see Fig. 3(a) and (b)].

B. Governing PDE

The (3-D, instantaneous) velocity in turbulent boundary layers is governed by the NS equations, which encode momentum and mass conservation. The lack of observations in the third dimension compels us to consider the Reynolds-averaged NS (RANS) (i.e., time-averaged) equations, a static version of NS that governs the behavior of the stream-wise velocity in space. Under appropriate assumptions (verified for this data in the supplementary material), the governing model simplifies to a 1-D PDE (the full derivation is also in the supplementary material)

$$\frac{u_\tau^2}{\delta} + \frac{\delta u_\tau}{\mathrm{Re}} \frac{\partial^2 \overline{u}(x,y)}{\partial y^2} - \frac{\partial}{\partial y} \overline{u'v'} = 0 \tag{15a}$$

$$-\overline{u'v'} = \underbrace{(\kappa y) \frac{\overline{u}(x,y)}{u_\tau} \frac{\partial \overline{u}(x,y)}{\partial y}}_{\text{Reynolds stress approximation}} \tag{15b}$$

where (15b) introduces a standard model for the Reynolds stress term $\overline{u'v'}$. In this case, the derivative of $\overline{u}$ with respect to time is 0 and the nonlinear differential operator as defined in the PDE (3a) is

$$N\left[\overline{u}(x,y)\right] = \frac{u_\tau^2}{\delta} + \frac{\delta u_\tau}{\mathrm{Re}} \frac{\partial^2 \overline{u}(x,y)}{\partial y^2} + \frac{\partial}{\partial y}\left((\kappa y) \frac{\overline{u}(x,y)}{u_\tau} \frac{\partial \overline{u}(x,y)}{\partial y}\right)$$

where from the experimental data, $\mathrm{Re} = 2700$ is the Reynolds number, $\delta = 0.1$ m is the boundary layer thickness, $u_\tau = 0.027$ m/s is the friction velocity, and $\kappa = 0.01107$. The boundary conditions are $\overline{u}(x,0) = 0$ m/s and $\overline{u}(x,\delta) = U_\infty$, where $U_\infty = 0.67$ m/s is the free-stream velocity.

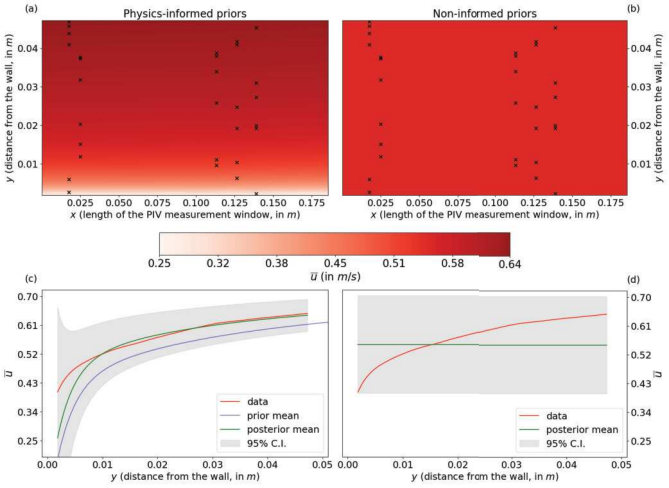

Fig. 4. Comparing boundary layer predictions with (a) and (c) physics-informed priors and (b) and (d) noninformed priors. The spatial forecast for each is shown in (a) and (b) along with the $n = 30$ observations. Average vertical profile of the predicted and observed velocity, along with the 95% credibility intervals is shown in (c) and (d). In (c), we also show the velocity forecast induced by the prior means. We do not show its counterpart in (d) since it is a flat line at 0 m/s.

1) BNN Assumptions and Prior Calibration: We employ a deep BNN comprised of $L = 2$ hidden layers, $k = 10$ nodes in each layer, a hyperbolic tangent activation function $g(\cdot) = \tanh(\cdot)$, and the likelihood is now expressed as $\overline{u}(x,y) \sim \mathcal{N}(\mu_{\overline{u}}(x,y), \sigma^2_{\overline{u}}(x,y))$. We can afford a shallower BNN here than the one in Section III because the process is static in time and the spatial relationships are 1-D, as shown in Fig. 3.

Since the boundary conditions on (15) are $\overline{u}(x,0) = 0$ and $\overline{u}(x,\delta) = 0.67$ and the data only cover the wall-normal direction up to $y = 0.0473$ m, we discretize the interval $[0,\delta]$ with N_y points and use a $N_{\mathrm{PDE}} = N_x \times N_y$ grid (N_x and N_y are defined as in Section IV-A) to obtain $\hat{\boldsymbol{\mu}}_{\mathrm{physics}}$ by minimizing the MSE with respect to (15)

$$\begin{aligned}\mathrm{mse}(\boldsymbol{\mu}) \approx{}& \frac{1}{N_{\mathrm{PDE}}} \sum_{i=1}^{N_{\mathrm{PDE}}} \left|\hat{f}_{\boldsymbol{\mu}}(x_i, y_i)\right|^2 \\ &+ \frac{1}{N_x} \sum_{i=1}^{N_x} \left\{ \left|\hat{\overline{u}}_{\boldsymbol{\mu}}(x_i, 0)\right|^2 + \left|\hat{\overline{u}}_{\boldsymbol{\mu}}(x_i, \delta) - U_\infty\right|^2 \right\} \\ \hat{f}_{\boldsymbol{\mu}}(x_i, y_i) :={}& \frac{u_\tau^2}{\delta} + \frac{\delta u_\tau}{\mathrm{Re}} \frac{\partial^2 \hat{\overline{u}}_{\boldsymbol{\mu}}(x_i, y_i)}{\partial y^2} \\ &+ \frac{\partial}{\partial y}\left\{ (\kappa y_i) \frac{\hat{\overline{u}}_{\boldsymbol{\mu}}(x_i, y_i)}{u_\tau} \frac{\partial \hat{\overline{u}}_{\boldsymbol{\mu}}(x_i, y_i)}{\partial y} \right\}. \end{aligned} \tag{16}$$

Fig. 4(c) in Section IV-C includes the results of the prior calibration (in blue), showing a clear departure from the observations (in red). Even with (16) on the order of magnitude of 10^{-5}, the inferred approximation $\hat{\overline{u}}_{\hat{\boldsymbol{\mu}}_{\mathrm{physics}}}$ shows a clear discrepancy with $\overline{u}$, due to the Reynolds stress approximation in (15b). Our prior confidence in $\hat{\boldsymbol{\mu}}_{\mathrm{physics}}$ is, therefore, smaller than in the simulation study in Section III and we set $C = 0.03$, where C is as defined in Section II-B. By comparison, the average entry of $\boldsymbol{\Sigma}$, from (2), is over four times larger than it

TABLE III

PHYSICS-INFORMED APPLICATION RESULTS OF $n = 30$ OBSERVATIONS FOR DIFFERENT LEVELS OF PRIOR CONFIDENCE, AS DEFINED IN SECTION II-B, $C \in \{0.03, 0.01/100\}$, IN (A) AND (B). (C) RESULTS FOR THE NONPHYSICS INFORMED APPLICATION, I.E., WITH A VAGUE PRIOR. DISPLAYED ARE THE BIAS, VARIANCE, AND MSE AS DEFINED IN SECTION III-A5 FOR THE PREDICTED STREAM-WISE VELOCITY $\bar{u}$. WE ALSO REPORT THE COMPUTATIONAL DEMAND

	Physics-Informed		Non-Informed
	$C = 0.03$	$C = 0.01/100$	$\boldsymbol{\Sigma}_l = (1/N_l)\boldsymbol{I}$
	(a)	(b)	(c)
Bias^2 (10^{-3})	0.77	1.31	4.20
Variance (10^{-3})	43.38	12.96	77.66
MSE (10^{-3})	44.15	14.27	81.86
Runtime (mm:ss)	05 : 54	05 : 58	06 : 10

is in the simulation study, where each diagonal entry of $\boldsymbol{\Sigma}$ is equal to $\sigma^2_{\text{physics}} = 0.0025^2$. In Section IV-C, we also consider $C = 0.01/100$ to discuss the role of an excessively high prior confidence in (15).

Using the $n = 30$ observations in the vector $\bar{\boldsymbol{u}}$, we perform the Bayesian update with variational inference with $N_{\text{BBB}} = 10000$ iterations of Algorithm 1 to approximate the true posterior. This particular choice of observations is meant to replicate an instrument that can record 2-D velocity at scattered locations along a vertical line with some amount of measurement error.

C. Results

We applied our approach with a physics-informed prior, $\mathbb{E}(\boldsymbol{\theta}) = \hat{\boldsymbol{\mu}}_{\text{physics}}$ and $\boldsymbol{\Sigma}$ determined by $C \in \{0.03,\ 0.01/100\}$, as well as a noninformed prior, $\mathbb{E}(\boldsymbol{\theta}) = \boldsymbol{0}$ and $\boldsymbol{\Sigma}_l = (1/N_l)\boldsymbol{I}$, where N_l is the number of parameters in the lth layer. Fig. 4(a) and (b) shows the forecast throughout the spatial domain and Fig. 4(c) and (d) shows the average vertical profiles of the streamwise velocity and 95% credibility intervals. A similar figure for the case with $C = 0.01/100$ is available in the supplementary material. Table III(a) and (c) presents the MSE as introduced in (14).

Our results show a clear improvement in performance both in terms of bias and variance when the priors are informed by the PDE in (15). The bias and variance in the physics-informed case are 0.77×10^{-3} and 43.38×10^{-3}, respectively, as the spatial map in Fig. 4(a) shows the forecast velocity in a similar stratified pattern to the data in Fig. 3. The noninformed model forecasts a flat streamwise velocity, approximately equal to the wall-normal average for all y, but is unable to capture the nonlinearity in the profile, as is apparent in Fig. 4(c) and (d), with a bias and variance equal to 4.20×10^{-3} and 77.66×10^{-3}, respectively. Most importantly, boundary layers such as the one outlined in Section IV-A always exhibit increasing gradients in the wall-normal direction [46] and a forecast such as the one shown in Fig. 4(b) is not only poor but also unphysical. On the other hand, the physics-informed forecast in Fig. 4(a) shows a physically consistent boundary layer due to the spatial relationship established a priori.

The additional evidence for the improvement in performance in the physics-informed case is in diagnostic analysis of the Reynolds stress $\overline{u'v'}$, which, physically, represents momentum transport due to covariances in turbulent velocity fluctuations. From the instantaneous velocity fields $u(\boldsymbol{s}, t)$ and $v(\boldsymbol{s}, t)$, we computed $\overline{u'v'}$ for our data as folows:

$$\overline{u'v'}(\boldsymbol{s}) = \sum_{t=1}^{T} u(\boldsymbol{s}, t)v(\boldsymbol{s}, t) - \left[\sum_{t=1}^{T} u(\boldsymbol{s}, t)\right]\left[\sum_{t=1}^{T} v(\boldsymbol{s}, t)\right]$$

and compared it with its estimated value [derived from (15b)] by computing the MSE across the spatial domain, where, recall, the points are on an evenly spaced grid. In the physics-informed case, the MSE is 4.16×10^{-7}, one order of magnitude smaller than the MSE in the noninformed case, 1.53×10^{-6}, and visual evidence for this difference is in Fig. S17 of the supplementary material.

The results with high prior confidence, $C = 0.01/100$, show the role of excessive confidence in the prior. As can be seen from Table III(b), setting a high prior confidence in the PDE prevents the data from properly augmenting the physics-informed prior. In this case, the bias increases by over 70%, 1.31×10^{-3}, even though the variance is lower, 12.96×10^{-3}. The governing PDE (15) does not faithfully describe observed behavior in the way that Burgers' equation (12) did in Section III, since the Reynolds stress approximation in (15b) is only approximately correct. As such, (15) constitutes prior knowledge on the behavior of $\bar{u}$, but only loosely so. The physics-informed model with very high prior confidence unbalances the posterior in the direction of the PDE and, while outperforming the noninformed model, falls short of producing a reliable forecast.

These results have significant practical implications. Fluid dynamics datasets are notoriously difficult to obtain and experimental settings are often required. Since we consider the time-averaged streamwise velocity, the vast majority of the variability in the data is in the y-direction, as shown in Fig. 3. In spite of this simplification and the $n = 30$ observations being collected along five vertical cross sections of the domain, the noninformed models were unable to identify the spatial relationships in the data. The physics-informed approach instead established the spatial dependence a priori and did not require observations evenly distributed in space to produce physically consistent forecasts. In addition, this approach also has application in computational fluid dynamics, where models for unclosed quantities (such as the Reynolds stress) can be developed with knowledge of a governing PDE. For example, in the family of Reynolds stress models known as algebraic stress models, an equation for $\overline{u'v'}$ is solved in space and time (this equation is itself subject to a closure model). If this PDE were to be used as the prior, with appropriate prior variance, sparse measurements could potentially yield a PIP-BNN model for the Reynolds stress which covers space and time, and which is informed by a governing PDE. Furthermore, the use of the PIP-BNN approach with varying degrees of modeling in the governing PDE can be used to understand more deeply the implications of model accuracy.

V. Conclusion

This work introduced a novel model-based Bayesian approach, PIP-BNN, to incorporate contextual knowledge in the form of a PDE in a BNN to forecast spatiotemporal processes. The proposed model produces improved forecasts relative to a BNN with vague priors, especially with a small amount of data, as we show in both a simulation study and an application to boundary layer velocity. We propose a different paradigm of PINNs by incorporating physical knowledge in the prior without relying on observations, rather than expressing it as a constraint, so that true a priori knowledge is encoded. Given a level of confidence in the priors, expressed as a prior variance, the result is a physics-informed forecast. Incorporating in the prior the spatiotemporal relationships described by a PDE constitutes a modeling strategy to borrow strength from a governing model and data to produce physical consistent forecasts.

In a simulated environment governed by the viscous Burgers' equation, our Bayesian framework shows strong predictive skills in terms of posterior mean squared error despite using a small number of observations. We also performed a sensitivity study and assessed how a larger number of observations, lower amounts of noise, and lower prior variance all lead to improved posteriors.

We apply our approach to average boundary layer velocity as it interacts with a surface, governed by a simplified NS that only approximately describes observed behavior due to additional modeling assumptions. The physics-informed prior again shows superior forecasting abilities to the vague prior alternative. Our results also show that excessively high confidence in the PDE yields forecasts that resemble theoretical behavior too closely, as the posterior skews in the direction of the physics-informed prior means. Overall, the application results show that a governing model need not perfectly describe observed behavior in order for our approach to perform better than the vague prior alternative.

Our approach represents the first step toward a Bayesian-driven model-based methodology that yields physically consistent predictions. In NNs, the parameters, even if physics-informed a priori, lack interpretability, and measuring our prior confidence in them without employing observations is challenging. We introduce a metric inspired by the coefficient of variation so that each parameter's prior variance is relative to its mean, but it still does not easily translate to confidence. As such, our results may be interpreted as conditional on a level of prior confidence, as our approach shows promising results in its ability to propagate uncertainty through the model. For a given level of noise in the data, the posterior variance in both the simulation study and the application follows the prior confidence, as is expected in any Bayesian model.

The difference in results of PIP-BNN applied to a simulated setting versus a real-world case highlights the pivotal role of the prior variance. The case of Burgers' equation shows that when the PDE is the data-generating process, a comparably large prior variance has a negligible effect on the posterior mean. In contrast, the boundary layer results suggest the presence of a lower bound below which the data, especially when scarce, is unable to yield adequate predictions. As such, it is sensible to conclude that the optimal choice for the prior variance, addressing the demonstrably subjective issue of confidence in a PDE, belongs to a range of values, which may be found using cross validation.

Since PIP-BNN is not a Bayesian PDE solver, we refrained from comparing its performance with numerical methods, a critic often levied in "ML-for-PDE" papers [47], as we also highlight scenarios for which our proposed approach yields suboptimal predictions. In fact, in the simulation study results we report poor model performance even in a highly idealized setting such as Burgers' equation. Similarly, in the application, we show the potentially detrimental role of an excessively low prior variance, which leads to inadequate posteriors and therefore suboptimal predictions.

An important issue to be addressed in future work concerns the physics with respect to which we calibrate the prior. The governing models employed in Sections III and IV assume no uncertainty in the PDE itself, as our results may also be interpreted as conditional on the physical parameters. Large portions of the PINN literature focus on finding the best physical parameters for a given system [5], [17], [48] and an extension of this work would impose a prior on them as well. Such a framework would provide, in principle, a more rigorous quantification of our level of prior confidence in the governing model while producing a posterior on the physical parameters. Finally, the flexibility of our approach, along with its modest computational demand, shows promise for a future application to the 3-D NS equations, which is more realistic than the average boundary layer velocity studied in Section IV. Researchers typically tackle these problems using computational fluid dynamics, which solve NS in every point of a fine mesh and whose computational demand is especially high for complex flows [49]. As the classical PINN approach has already shown promise toward easing the computational burden [50], extensions of our work may provide an alternative model-based strategy with significant implications in several areas of science and engineering.

Code and Data Availability

This code and data can be found in the following GitHub repository: github.com/Env-an-Stat-group/25.Menicali.IEEETNNLS. The application data that support the findings of this study are openly available upon request from [51].

References

[1] Y.-G. Ham, J.-H. Kim, and J.-J. Luo, "Deep learning for multi-year ENSO forecasts," *Nature*, vol. 573, no. 7775, pp. 568–572, Sep. 2019.

[2] J. Zhang and X. Zhao, "Spatiotemporal wind field prediction based on physics-informed deep learning and LiDAR measurements," *Appl. Energy*, vol. 288, Apr. 2021, Art. no. 116641.

[3] J. N. Kutz, "Deep learning in fluid dynamics," *J. Fluid Mech.*, vol. 814, pp. 1–4, Mar. 2017.

[4] D. Greenwood, "An overview of neural networks," *Behav. Sci.*, vol. 36, no. 1, pp. 1–33, Jan. 1991.

[5] M. Raissi, P. Perdikaris, and G. E. Karniadakis, "Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations," *J. Comput. Phys.*, vol. 378, pp. 686–707, Feb. 2019.

[6] A. Arzani, J.-X. Wang, and R. M. D'Souza, "Uncovering near-wall blood flow from sparse data with physics-informed neural networks," *Phys. Fluids*, vol. 33, no. 7, p. 071905, Jul. 2021.

[7] M. Chu et al., "Physics informed neural fields for smoke reconstruction with sparse data," *ACM Trans. Graph.*, vol. 41, no. 4, pp. 1–14, Jul. 2022, doi: 10.1145/3528223.3530169.

[8] S. Greydanus, M. Dzamba, and J. Yosinski, "Hamiltonian neural networks," 2019, *arXiv:1906.01563*.

[9] M. Cranmer, S. Greydanus, S. Hoyer, P. Battaglia, D. Spergel, and S. Ho, "Lagrangian neural networks," 2020, *arXiv:2003.04630*.

[10] A. Dutta, K. Lakshmanan, S. Kumar, and A. Ramamoorthy, "Newtonian physics informed neural network (NwPiNN) for spatio-temporal forecast of visual data," *Hum.-Centric Intell. Syst.*, vol. 4, no. 3, pp. 335–343, May 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:269875859

[11] O. Graf, P. Flores, P. Protopapas, and K. Pichara, "Error-aware B-PINNs: Improving uncertainty quantification in Bayesian physics-informed neural networks," 2022, *arXiv:2212.06965*.

[12] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.

[13] A. Labach, H. Salehinejad, and S. Valaee, "Survey of dropout methods for deep neural networks," 2019, *arXiv:1904.13310*.

[14] L. V. Jospin, H. Laga, F. Boussaid, W. Buntine, and M. Bennamoun, "Hands-on Bayesian neural networks—A tutorial for deep learning users," *IEEE Comput. Intell. Mag.*, vol. 17, no. 2, pp. 29–48, May 2022.

[15] J. Stiasny, G. S. Misyris, and S. Chatzivasileiadis, "Physics-informed neural networks for non-linear system identification for power system dynamics," in *Proc. IEEE Madrid PowerTech*, Jun. 2021, pp. 1–6.

[16] M. Izzatullah, I. E. Yildirim, U. B. Waheed, and T. Alkhalifah, "Predictive uncertainty quantification for Bayesian physics-informed neural network (Pinn) in hypocentre estimation problem," in *Proc. 83rd EAGE Annu. Conf. Exhib.*, 2022, pp. 1–5.

[17] L. Yang, X. Meng, and G. E. Karniadakis, "B-PINNs: Bayesian physics-informed neural networks for forward and inverse PDE problems with noisy data," *J. Comput. Phys.*, vol. 425, Jan. 2021, Art. no. 109913.

[18] R. B. Litterman, "Forecasting with Bayesian vector autoregressions: Five years of experience," *J. Bus. Econ. Statist.*, vol. 4, no. 1, pp. 25–38, Jan. 1986. [Online]. Available: http://www.jstor.org/stable/1391384

[19] F. Huber, G. Kastner, and M. Feldkircher, "Should I stay or should I go? A latent threshold approach to large-scale mixture innovation models," *J. Appl. Econometrics*, vol. 34, no. 5, pp. 621–640, Aug. 2019.

[20] J. C. C. Chan, "Minnesota-type adaptive hierarchical priors for large Bayesian VARs," *Int. J. Forecasting*, vol. 37, no. 3, pp. 1212–1226, Jul. 2021.

[21] F. Lindgren, H. Rue, and J. Lindström, "An explicit link between Gaussian fields and Gaussian Markov random fields: The stochastic partial differential equation approach," *J. Roy. Stat. Soc. B, Stat. Methodol.*, vol. 73, no. 4, pp. 423–498, Sep. 2011.

[22] T. Bui-Thanh and O. Ghattas, "A construction of inverse problems with PDE constraints for Bayesian inference," *SIAM J. Sci. Comput.*, vol. 35, no. 4, pp. A2493–A2523, 2013.

[23] V. Fortuin, D. Baranchuk, G. Rätsch, and S. Mandt, "Posterior collapse and latent variable non-identifiability," in *Proc. Adv. Neural Inf. Process. Syst.*, Nov. 2020, pp. 1–13.

[24] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, Jan. 2020, pp. 37799–37812.

[25] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning* (Adaptive Computation and Machine Learning series). Cambridge, MA, USA: MIT Press, 2016. [Online]. Available: https://books.google.com/books?id=-s2MEAAAQBAJ

[26] T. Park and G. Casella, "The Bayesian Lasso," *J. Amer. Stat. Assoc.*, vol. 103, no. 482, pp. 681–686, 2008.

[27] L. Gruber and G. Kastner, "Forecasting macroeconomic data with Bayesian VARs: Sparse or dense? It depends!," *Int. J. Forecasting*, vol. 41, pp. 1–20, Feb. 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0169207025000081

[28] C. Blundell, J. Cornebise, K. Kavukcuoglu, and D. Wierstra, "Weight uncertainty in neural network," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 1613–1622.

[29] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe, "Variational inference: A review for statisticians," *J. Amer. Stat. Assoc.*, vol. 112, no. 518, pp. 859–877, Apr. 2017, doi: 10.1080/01621459.2017.1285773.

[30] P. Izmailov, S. Vikram, M. D. Hoffman, and A. G. Wilson, "What are Bayesian neural network posteriors really like?," 2021, *arXiv:2104.14421*.

[31] M. Magris and A. Iosifidis, "Bayesian learning for neural networks: An algorithmic survey," *Artif. Intell. Rev.*, vol. 56, no. 10, pp. 11773–11823, Oct. 2023.

[32] T. Papamarkou, J. Hinkle, M. T. Young, and D. Womble, "Challenges in Markov chain Monte Carlo for Bayesian neural networks," *Stat. Sci.*, vol. 37, no. 3, pp. 425–442, Aug. 2022, doi: 10.1214/21-sts840.

[33] S. Kullback and R. A. Leibler, "On information and sufficiency," *Ann. Math. Statist.*, vol. 22, no. 1, pp. 79–86, 1951, doi: 10.1214/aoms/1177729694.

[34] D. P. Kingma, T. Salimans, and M. Welling, "Variational dropout and the local reparameterization trick," 2015, *arXiv:1506.02557*.

[35] J. Paisley, D. Blei, and M. Jordan, "Variational Bayesian inference with stochastic search," 2012, *arXiv:1206.6430*.

[36] J. M. Burgers, "A mathematical model illustrating the theory of turbulence," *Adv. Appl. Mech.*, vol. 1, pp. 171–199, Jan. 1948. [Online]. Available: https://api.semanticscholar.org/CorpusID:117400324

[37] R. de Kat and B. Ganapathisubramani, "Frequency–wavenumber mapping in turbulent shear flows," *J. Fluid Mech.*, vol. 783, pp. 166–190, Nov. 2015.

[38] D. Fiscaletti, R. de Kat, and B. Ganapathisubramani, "Spatial–spectral characteristics of momentum transport in a turbulent boundary layer," *J. Fluid Mech.*, vol. 836, pp. 599–634, Feb. 2018.

[39] J.-X. Wang, J.-L. Wu, and H. Xiao, "Physics-informed machine learning approach for reconstructing Reynolds stress modeling discrepancies based on DNS data," *Phys. Rev. Fluids*, vol. 2, no. 3, p. 034603, Mar. 2017.

[40] H. Eivazi, M. Tahani, P. Schlatter, and R. Vinuesa, "Physics-informed neural networks for solving Reynolds-averaged Navier–Stokes equations," *Phys. Fluids*, vol. 34, no. 7, Jul. 2022, Art. no. 075117.

[41] R. Kohli, *Developments in Imaging and Analysis Techniques for Micro- and Nanosize Particles and Surface Features*. Oxford, U.K.: William Andrew, 2012, pp. 215–306.

[42] C. E. Willert and M. Gharib, "Digital particle image velocimetry," *Experim. Fluids*, vol. 10, no. 4, pp. 181–193, Jan. 1991.

[43] J. Westerweel, *Digital Particle Image Velocimetry: Theory and Application*. Delft, The Netherlands: Delft University Press, 1993. [Online]. Available: https://books.google.com/books?id=8pH0AQAACAAJ

[44] R. D. Keane and R. J. Adrian, "Optimization of particle image velocimeters. I. Double pulsed systems," *Meas. Sci. Technol.*, vol. 1, no. 11, pp. 1202–1215, Nov. 1990.

[45] T. Utami, R. F. Blackwelder, and T. Ueno, "A cross-correlation technique for velocity field extraction from particulate visualization," *Experim. Fluids*, vol. 10, no. 4, pp. 213–223, Jan. 1991.

[46] P. K. Kundu, I. M. Cohen, and D. R. Dowling, "Boundary layers and related topics," in *Fluid Mechanics*, 6th ed., Boston, MA, USA: Academic, 2016, ch. 10, pp. 469–532. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780124059351000101

[47] N. McGreivy and A. Hakim, "Weak baselines and reporting biases lead to overoptimism in machine learning for fluid-related partial differential equations," *Nature Mach. Intell.*, vol. 6, no. 10, pp. 1256–1269, Sep. 2024, doi: 10.1038/s42256-024-00897-5.

[48] X. Jiang, X. Wanga, Z. Wena, E. Li, and H. Wang, "An E-PINN assisted practical uncertainty quantification for inverse problems," 2022, *arXiv:2209.10195*.

[49] F. Balduzzi, J. Drofelnik, A. Bianchini, G. Ferrara, L. Ferrari, and M. S. Campobasso, "Darrieus wind turbine blade unsteady aerodynamics: A three-dimensional Navier–Stokes CFD assessment," *Energy*, vol. 128, pp. 550–563, Jun. 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0360544217305820

[50] S. Cai, Z. Mao, Z. Wang, M. Yin, and G. E. Karniadakis, "Physics-informed neural networks (PINNs) for fluid mechanics: A review," 2021, *arXiv:2105.09506*.

[51] R. De Kat, L. Gan, D. Fiscaletti, J. Dawson, and B. Ganapathisubramani, "Wide-field time-resolved particle image velocimetry data of a turbulent boundary layer," Aerodyn. Flight Mech. Group, University of Southampton, Southampton, U.K., Tech. Rep. SOTON/D0297, 2017. [Online]. Available: https://eprints.soton.ac.uk/416120/

Luca Menicali's working manuscripts include a physics-informed neural model for predicting lake temperature and oxygen profiles, NSF-funded Bayesian and semi-supervised models for reconstructing presettlement midwestern ecosystems, and deep learning approaches for turbulent convection. His work specializes in statistical modeling and machine learning for environmental and climate systems. His research combines physical knowledge with modern AI methods to model spatiotemporal processes in lake and ecosystem dynamics.

David H. Richter holds a concurrent faculty position at the Department of Aerospace and Mechanical Engineering, University of Notre Dame, Notre Dame, IN, USA. His research group is interested in computational techniques, such as direct numerical simulation and large eddy simulation for studying turbulence in the atmosphere and ocean. His detailed knowledge of turbulent exchange of heat, moisture, and momentum at the air-sea interface is critical for accurate storm prediction. He is also interested in developing tools for simulating multiphase systems that can be applied to a wide variety of geophysical flows, including subsurface sediment transport and saltation.

Stefano Castruccio work focuses on the development of spatiotemporal models for fluid dynamics and machine learning methods for data in space and time.

Dr. Castruccio is elected as a fellow of American Statistical Association and a member of the International Statistical Institute. He was a recipient of the 2024 Gordon Bell Prize for Climate Modelling, as well as the 2023 TIES President's Invited Lecture, the 2021 TIES Abdel El-Shaarawi Early Investigator Award, and the 2020 Early Investigator Award from the Section on Statistics and the Environment, American Statistical Association.