Proceedings of Machine Learning Research vol XX:1–13, 2023

# Physics-Guided Active Learning of Environmental Flow Fields

**Reza Khodayi-mehr** REZA.KHODAYI.MEHR@DUKE.EDU
*Duke University, Durham, NC, USA*

**Pingcheng Jian** PINGCHENG.JIAN@DUKE.EDU
*Duke University, Durham, NC, USA*

**Michael M. Zavlanos** MICHAEL.ZAVLANOS@DUKE.EDU
*Duke University, Durham, NC, USA*

## Abstract

We propose a physics-based method to learn environmental fields (EFs) using a mobile robot. Common data-driven methods require prohibitively many measurements to accurately learn such complex EFs. On the other hand, while physics-based models provide global knowledge of EFs, they require experimental validation, depend on uncertain parameters, and are intractable to solve onboard mobile robots. To address these challenges, we propose a Bayesian framework to select and improve upon the most likely physics-based models of EFs in real-time, from a pool of numerical solutions generated offline as a function of the uncertain parameters. Specifically, we use Gaussian Processes (GPs) to construct statistical models of EFs, and rely on the pool of numerical solutions to inform their prior mean. To incorporate flow measurements into these GPs, we control a custom-built mobile robot through a sequence of waypoints that maximize the information content of the measurements. We experimentally demonstrate that our proposed framework constructs a posterior distribution of the flow field that better approximates the real flow compared to the prior numerical solutions and purely data-driven methods.

**Keywords:** Environmental flow fields, physics-based learning, active learning, mobile robots, Gaussian processes.

## 1. Introduction

Mobile robots have been widely used in environmental sensing applications to collect informative measurements in a cost effective manner Dunbabin and Marques (2012); Yuh (2000). Knowledge of the underlying flow field is often essential in these applications both for estimation and navigation. Our goal in this paper is to develop a new physics-based method to learn high-fidelity statistical models of flow fields using only sparse flow measurements, that is also tractable so that these measurements can be collected online by a mobile robot.

A widely used method for estimating flow properties is numerical simulation based on the Reynolds-averaged Navier Stokes (RANS) equations, a system of Partial Differential Equations (PDEs). This approach is cost-effective and provides global estimation of the flow over a domain of interest. Nevertheless, solutions provided by RANS models are generally incompatible with each other and with empirical data and require experimental validation Ling and Templeton (2015). Furthermore, precise knowledge of boundary conditions (BCs) and domain geometry is often unavailable, which results in even larger inaccuracies in the predicted flow properties. Finally, solving RANS models onboard mobile robots with limited computational resources is still intractable. Due to such challenges, the authors in Lee et al. (2019); Xu et al. (2013); Duecker et al. (2019); Nguyen

et al. (2020) forgo the use of physics-based models and instead advocate the use of purely data-driven statistical methods that rely on Gaussian processes (GPs) to estimate spatiotemporal fields starting from non-informative constant priors. Such purely data-driven methods, however, can require a prohibitively large number of measurements to accurately estimate complex flow fields, making them intractable in practice.

To address the computational and sample complexity of purely physics-based and data-driven approaches, respectively, in this paper, we propose a Bayesian framework that combines empirical data with physics-based models to learn accurate representations of unknown flow fields in *real-time*. Specifically, given a distribution of the uncertain parameters like BCs, we generate offline a pool of numerical models of the flow field by solving the RANS equations for different choices of uncertain parameters selected according to this distribution. These numerical models may be inconsistent with each other or the true flow and thus, are only used to inform the prior mean of a corresponding pool of statistical GP models of the flow properties, specifically the mean velocity components and turbulent intensity field. Then, the proposed Bayesian framework allows to incorporate empirical data collected by a mobile robot sensor, to select the most likely flow models from the pool and to obtain the posterior distribution of the flow properties given each model. As such, our approach takes advantage of the global information provided by physics-based models without needing to solve for them online, and produces high fidelity estimations of the flow properties using only sparse data, which is not possible with purely data-driven methods. A major contribution of this paper is also the experimental validation of the proposed framework, demonstrating that it is robust to the significant uncertainties that are present in the real-world; see Experiment (b,a) for a visualization of such uncertainties. This is in contrast to the relevant literature that typically relies on numerical simulations to showcase the proposed methods Xu et al. (2013); Duecker et al. (2019); Nguyen et al. (2020).

Compared to path planning methods in marine robotics Hollinger et al. (2016); Ma et al. (2016); Jones and Hollinger (2017); Kularatne et al. (2016); Edwards et al. (2017); Lee et al. (2019), where flow fields are used for *persistent* monitoring of aquatic phenomena with an emphasis on designing optimal paths subject to time and energy budget constraints, here the purpose of planning is to maximize the information collected about the flow field itself. While those planning methods might be optimal in theory, in practice they often require approximations to mitigate their computational cost Jones and Hollinger (2017). Moreover, sophisticated algorithms like the stochastic optimal control approach proposed in Duecker et al. (2019) for data-driven environmental sensing, have only been demonstrated on simple uncertainty fields and it is unclear if they can be applied to the highly nonlinear uncertainty fields considered here. In general, active learning of GPs has been extensively investigated in the robotics literature with applications ranging from estimation of nonlinear dynamics to spatiotemporal fields Berkenkamp et al. (2016); Berkenkamp and Schoellig (2015); Ostafew et al. (2014); Wei et al. (2014); Lan and Schwager (2017, 2013). Closely related are also methods for robotic state estimation and planning with Gaussian noise; see Freundlich et al. (2017, 2015); Khodayi-mehr et al. (2019). This literature typically employs simple, explicit models of small dimensions and does not consider model ambiguity or parameter uncertainty. Instead, here we focus on complex models of continuous flow fields that are implicit solutions of RANS models, a system of PDEs. Although our planning method is not theoretically optimal, we show that it is effective in solving complex real-world environmental sensing problems.

## 2. Statistical Model of Flow Fields

Consider a turbulent flow field over a domain of interest $\bar{\Omega} \subset \mathbb{R}^3$ and let $\mathbf{q}(x,t) : \bar{\Omega} \times [0,T] \to \mathbb{R}^3$ denote the corresponding flow velocity vector, where $x \in \bar{\Omega}$ and $t \in [0,T]$. Due to turbulence, $\mathbf{q}(x,t)$ is a random variable (RV) subject to high variation. In what follows, we assume that the turbulent flow $\mathbf{q}(x,t)$ is stationary and ergodic. In this case, the turbulent velocity field $\mathbf{q}(x,t)$ and the turbulent intensity field $i(x,t)$, which is a measure of turbulent variations, can be approximated by their time averaged values $\mathbf{q}(x)$ and $i(x)$, respectively. Consider also a mobile robot sensor that can obtain instantaneous measurements of the random velocity field $\mathbf{q}(x,t)$ for a period of time at a set of locations and let the vector $\hat{\mathbf{y}}$ denote the collection of these measurements.[1] Then, the problem that we address in the paper can be defined as follows.

**Problem 1 (Environmental Sensing)** *Given the vector of measurements $\hat{\mathbf{y}}$ of the instantaneous velocity field $\mathbf{q}(x,t)$, collected by a mobile robot, obtain the posterior distributions $\bar{\pi}(\hat{\mathbf{q}}(x,t)|\hat{\mathbf{y}})$ and $\bar{\pi}(i(x,t)|\hat{\mathbf{y}})$ of the mean velocity field and turbulent intensity field.*

### 2.1. Gaussian Process Models of Flow Properties

Let $\xi \in \mathbb{R}^{n_\xi}$ encode the parameters needed to specify the domain $\bar{\Omega}$ and flow conditions imposed on its boundaries. Given $\xi$, we employ Reynolds-Averaged Navier Stokes (RANS) models to predict the flow properties like the mean velocity components and turbulent intensity, globally over $\bar{\Omega}$. Various RANS models exist that are broadly categorized into eddy viscosity models (EVMs) and Reynolds stress models (RSMs) Wilcox (1993). The solution returned by these models can be incompatible with each other. Moreover, there is often high uncertainty in the parameters $\xi$, i.e., the domain geometry and boundary conditions (BCs). As a result, numerical solutions generally require experimental validation. Finally, solving RANS models onboard mobile robots in real-time is still intractable. In what follows, we propose a statistical framework to address these challenges and obtain a physics-based solution to Problem 1.

To do so, we first define Gaussian Process (GP) models for the mean velocity and turbulent intensity fields. Specifically, given a value for parameters $\xi$, we solve a RANS model, e.g., RSM, to obtain a prediction of the flow properties. Let $\mu_u(x)$ denote the prediction for the first mean velocity component $u(x) = \mathbf{q}_1(x)$. Then, we model the prior distribution of $u(x)$, before collecting any measurements, using the following GP

$$u(x) \sim \mathcal{GP}\left(\mu_u(x), \bar{\kappa}_u\left(x, x'\right)\right), \tag{1}$$

where, the kernel function $\bar{\kappa}_u\left(x, x'\right)$ is defined as

$$\bar{\kappa}_u\left(x, x'\right) = \bar{\sigma}_u^2\left(x, x'\right) \rho\left(x, x'\right). \tag{2}$$

In (2), the standard deviation $\bar{\sigma}_u\left(x, x'\right) \in \mathbb{R}_+$ encapsulates the prior uncertainty in $u(x)$ and $\rho(x, x')$ is the correlation function. Explicitly, we define the standard deviation as

$$\bar{\sigma}_u^2\left(x, x'\right) = \bar{\sigma}_{u,0}^2 + \frac{1}{n_0} q_{\text{ref}}^2 i(x) i\left(x'\right), \tag{3}$$

1. See Section II-A in Khodayi-mehr and Zavlanos (2021) for details on computing $\mathbf{q}(x)$ and $i(x)$ from measurements of $\mathbf{q}(x,t)$.

where the constant $\bar{\sigma}_{u,0} \in \mathbb{R}_+$ is a measure of confidence in the numerical solution $\mu_u(x)$ and is selected depending on the convergence metrics provided by the RANS solver. The second term in (3), which is unavailable in purely data-driven approaches, captures the local variability due to turbulence, by relying on the estimation of turbulent intensity $i(x)$ provided by the RANS model. $n_0 \in \mathbb{N}_+$ is a nominal number of samples that scales this variability for the *averaged* velocity component. We define the correlation function $\rho(x, x')$ in (2) as a compactly supported polynomial

$$\rho\left(x, x'\right) = \left[\left(1 - \frac{\|x - x'\|}{\ell}\right)_+\right]^2, \tag{4}$$

where $\ell \in \mathbb{N}_+$ is the correlation characteristic length and the operator $(\alpha)_+ = \max(0, \alpha)$. The correlation function (4) implies that two points with distance larger than $\ell$ are uncorrelated, which results in sparse covariance matrices Rasmussen and Williams (2006).

In practice, it is impossible to obtain noiseless samples of $u(x)$. Thus, we consider a measurement model for $u(x)$ with additive Gaussian noise $\epsilon_u \sim \mathcal{N}(0, \sigma_u^2(x))$; see Section II-D in Khodayi-mehr and Zavlanos (2021) for more details. Specifically, let $y_u(x) \in \mathbb{R}$ denote a measurement of the first mean velocity component at a location $x$ given by

$$y_u(x) = u(x) + \epsilon_u(x). \tag{5}$$

Then $y_u(x)$ is also a GP

$$y_u(x) \sim \mathcal{GP}(\mu_u(x), \kappa_u(x, x')) \tag{6}$$

with the following kernel function

$$\kappa_u(x, x') = \bar{\kappa}_u(x, x') + \sigma_u^2(x)\, \delta(x - x'), \tag{7}$$

where $\delta(x - x')$ is the Kronecker delta function.

Given a vector of measurements $\mathbf{y}_{u,k}$ collected by the mobile robot at a set of $k$ locations $\mathcal{X}_k$, the predictive distribution of $u(x)$ at a point $x$, conditioned on measurements $\mathbf{y}_{u,k}$, is a Gaussian distribution whose mean and variance are given in closed-form by

$$\mu_u\left(x \mid \mathcal{X}_k\right) = \mu_u(x) + \overline{\mathbf{\Sigma}}_{x\mathcal{X}} \mathbf{\Sigma}_{\mathcal{X}\mathcal{X}}^{-1} \left(\mathbf{y}_{u,k} - \boldsymbol{\mu}_{\mathcal{X}}\right), \tag{8a}$$

$$\gamma_u^2\left(x \mid \mathcal{X}_k\right) = \bar{\kappa}_u(x, x) - \overline{\mathbf{\Sigma}}_{x\mathcal{X}} \mathbf{\Sigma}_{\mathcal{X}\mathcal{X}}^{-1} \overline{\mathbf{\Sigma}}_{\mathcal{X}_x}, \tag{8b}$$

where $\mu_{\mathcal{X}}$ denotes the mean function evaluated at measurement locations $\mathcal{X}_k$ and the entries of the covariance matrices $\bar{\Sigma}_{x\mathcal{X}}$ and $\mathbf{\Sigma}_{\mathcal{X}\mathcal{X}}$ are computed using (2) and (7), respectively. Note that for simplicity, we have dropped the subscripts $u$ and $k$ from the matrices in (8). It is easy to show that the matrix $\mathbf{\Sigma}_{\mathcal{X}\mathcal{X}}$ is positive-definite and invertible; see Proposition II.4 in Khodayi-mehr and Zavlanos (2021).

From the Navier-Stokes PDEs, it follows that the mean velocity components are correlated Wilcox (1993). However, since the prior fields, obtained from RANS models, already capture this correlation, to simplify the pursuant development, we assume that turbulent flow properties are independent and thus, uncorrelated. Then, we can independently define GPs for the second mean velocity component $v$ and the turbulent intensity field with appropriate subscripts.[2]

2. We only consider in-plane velocity components throughout the paper. The extension of the theoretical developments for the third component is trivial.

### 2.2. Hierarchical Bayesian Model Selection

In Section 2.1, we constructed GP models of the flow properties given known parameters $\xi$ and a specific RANS model. However, as discussed before, models constructed in this way, can be inaccurate due to the high uncertainty in the parameters $\xi$ and the assumptions that need to be made to derive the RANS model. Next, we outline a model selection method that takes into account this parameter uncertainty to generate *offline*, a pool of likely models that later, the robot can select from and improve upon based on empirical data that it collects *online*.

Let $\mathcal{N}$ denote a RANS model and consider a possibly uniform discrete prior distribution $\tilde{\pi}(\mathcal{N})$ over the available models. Furthermore, let $\tilde{\pi}(\xi)$ denote the discrete prior distribution on the domain geometry and BC parameters $\xi$. If the prior on the parameters is continuous, we can construct a discrete approximation $\tilde{\pi}(\xi)$ using stochastic reduced order models (SROMs); see Calkins et al. (2017) for details. Let $\mathcal{M} = (\mathcal{N}, \xi)$ denote the numerical solution obtained using the RANS model $\mathcal{N}$ given the parameters $\xi$. Noting that $\mathcal{N}$ and $\xi$ are independent, $\tilde{\pi}(\mathcal{M}) = \tilde{\pi}(\mathcal{N})\tilde{\pi}(\xi)$. Let $\mathcal{M}_j$ denote the $j$-th numerical model obtained for one combination of discrete $\mathcal{N}$ and $\xi$ values and let the collection $\tilde{\pi}(\mathcal{M}) = \{p_{j,0}, \mathcal{M}_j\}_{j=1}^{\bar{n}}$ denote the pool of numerical models, where $p_{j,0}$ denotes the prior probability of model $\mathcal{M}_j$ and $\bar{n}$ is the number of models.

Given measurements $\mathbf{y}_{u,k}$, $\mathbf{y}_{v,k}$, and $\mathbf{y}_{i,k}$ of the mean velocity components and turbulent intensity at a set of $k$ locations $\mathcal{X}_k$, the posterior distribution over models can be obtained using Bayes' rule as

$$\begin{aligned} \tilde{\pi}\left(\mathcal{M}_j \mid \mathcal{X}_k\right) &= \alpha \bar{\pi}\left(\mathbf{y}_{u,k}, \mathbf{y}_{v,k}, \mathbf{y}_{i,k} \mid \mathcal{M}_j\right) \tilde{\pi}\left(\mathcal{M}_j\right) \\ &= \alpha \bar{\pi}\left(\mathbf{y}_{u,k} \mid \mathcal{M}_j\right) \bar{\pi}\left(\mathbf{y}_{v,k} \mid \mathcal{M}_j\right) \bar{\pi}\left(\mathbf{y}_{i,k} \mid \mathcal{M}_j\right) \tilde{\pi}\left(\mathcal{M}_j\right), \end{aligned} \tag{9}$$

where $\alpha$ is the normalizing constant in Bayes' rule and $\bar{\pi}(\mathbf{y}_{u,k}|\mathcal{M}_j)$ is the likelihood of the measurements $\mathbf{y}_{u,k}$ given model $\mathcal{M}_j$ and similarly for $\mathbf{y}_{v,k}$ and $\mathbf{y}_{i,k}$. Note that the joint likelihood of the measurements in (9) is equivalent to the product of the individual likelihoods since the flow properties are independent. From the definition of the GPs constructed for model $\mathcal{M}_j$ in Section 2.1, we can obtain these likelihoods in closed-form. For instance

$$\bar{\pi}\left(\mathbf{y}_{u,k} \mid \mathcal{M}_j\right) = \det\left(2\pi \boldsymbol{\Sigma}_{u,j}\right)^{-0.5} \exp\Big( -\frac{1}{2}\left(\mathbf{y}_{u,k} - \boldsymbol{\mu}_{u,j}\right)^T \boldsymbol{\Sigma}_{u,j}^{-1}\left(\mathbf{y}_{u,k} - \boldsymbol{\mu}_{u,j}\right)\Big), \tag{10}$$

where $\mu_{u,j}$ and $\boldsymbol{\Sigma}_{u,j}$ are short-hand notation for the mean and covariance of the GP corresponding to $y_u(x)$ and $\mathcal{M}_j$ at locations $\mathcal{X}_k$; see the discussion after equation (8). Since the sum of the discrete posterior model probabilities equals one, i.e., $\sum_{k=1}^{\bar{n}} \tilde{\pi}(\mathcal{M}_j|\mathcal{X}_k) = 1$, we can compute the normalizing constant $\alpha$ from (9) as

$$\alpha = \Big( \sum\nolimits_{j=1}^{\bar{n}} \bar{\pi}\left(\mathbf{y}_{u,k}, \mathbf{y}_{v,k}, \mathbf{y}_{i,k} \mid \mathcal{M}_j\right) \tilde{\pi}\left(\mathcal{M}_j\right) \Big)^{-1}. \tag{11}$$

Given $\alpha$, we can finally compute the posterior distribution $\tilde{\pi}(\mathcal{M}_j|\mathcal{X}_k)$ over the pool of models using (9). This amounts to Bayesian *model selection* and enables the mobile robot to assign probabilities to models constructed for likely parameter values, given the latest empirical data. Note that although hierarchical Bayesian models are extensively used in the literature Xu et al. (2013); Duecker et al. (2019); Krause and Guestrin (2007), to the best of our knowledge, this work is the first to utilize them for physics-based learning.

Given $\tilde{\pi}(\mathcal{M}_j|\mathcal{X}_k)$, the desired posterior distributions in Problem 1 are the marginal distributions $\pi(u|\mathcal{X}_k), \pi(v|\mathcal{X}_k)$, and $\pi(i|\mathcal{X}_k)$ after integrating over the models. These marginal distributions are

GP mixtures (GMs) with their mean and variance given by

$$\mu_u(x \mid \mathcal{X}_k) = \sum_{j=1}^{\bar{n}} p_{j,k}\mu_u(x \mid \mathcal{X}_k, \mathcal{M}_j), \tag{12a}$$

$$\gamma_u^2(x \mid \mathcal{X}_k) = \sum_{j=1}^{\bar{n}} p_{j,k}\gamma_u^2(x \mid \mathcal{X}_k, \mathcal{M}_j) + \sum_{j=1}^{\bar{n}} p_{j,k}\left[\mu_u(x \mid \mathcal{X}_k, \mathcal{M}_j) - \mu_u(x \mid \mathcal{X}_k)\right]^2, \tag{12b}$$

where $p_{j,k} = \tilde{\pi}(\mathcal{M}_j|\mathcal{X}_k)$ denotes the posterior model probabilities obtained from (9). Equation (12b) follows from the fact that the variance of a RV is the mean of conditional variances plus the variance of the conditional means. The expressions for $v(x)$ and $i(x)$ are identical.

## 3. Learning Flow Fields using a Mobile Robot

In this section, we formulate a path planning problem for a mobile robot sensor to collect measurements with maximum information content.[3] Specifically, let $\Omega$ denote a discretization of the environment where the robot sensor operates, excluding the points occupied by obstacles, and let $\mathcal{R} > 0$ denote a constraint on maximum travel distance. Furthermore, given the current measurement location $x_k$ and set of currently collected measurements $\mathcal{X}_k$, let $\mathcal{S}_{k+1} = \{x \in \Omega \backslash \mathcal{X}_k \mid \|x - x_k\| \leq \mathcal{R}\}$ denote the feasible subset of candidate measurement locations at step $k+1$. Then, our goal is to select the next measurement location $x_{k+1}$ from $\mathcal{S}_{k+1}$ so that the joint entropy of mean velocity components $u(x)$ and $v(x)$ at unobserved locations $\Omega \backslash \mathcal{X}_{k+1}$, given the measurements in $\mathcal{X}_{k+1}$, is minimized. With a slight abuse of notation, let $H(\Omega \backslash \mathcal{X}_{k+1} \mid \mathcal{X}_{k+1})$ denote this entropy.

Noting that $H(\Omega \backslash \mathcal{X}_{k+1} \mid \mathcal{X}_{k+1}) = H(\Omega) - H(\mathcal{X}_{k+1})$, minimizing $H(\Omega \backslash \mathcal{X}_{k+1} \mid \mathcal{X}_{k+1})$ is equivalent to maximizing $H(\mathcal{X}_{k+1})$. Furthermore, by the chain rule of entropy

$$H(\mathcal{X}_{k+1}) = H(x_{k+1} \mid \mathcal{X}_k) + \cdots + H(x_2 \mid \mathcal{X}_1) + H(x_1). \tag{13}$$

Thus, we can find the next best measurement location $x_{k+1}$ by solving the following optimization problem

$$x_{k+1}^* = \underset{x \in \mathcal{S}_{k+1}}{\operatorname{argmax}} H(x \mid \mathcal{X}_k). \tag{14}$$

Next we derive an expression for the objective $H(x \mid \mathcal{X}_k) = H(u(x), v(x) \mid \mathcal{X}_k)$ in (14). Since we assume that $u(x)$ and $v(x)$ are independent, we have

$$H(u, v) = H(u \mid v) + H(v) = H(u) + H(v), \tag{15}$$

where we have dropped dependence on $x$ and $\mathcal{X}_k$ for simplicity.

Recall from Section 2.2 that the posterior distributions of $u(x)$ and $v(x)$ are GMs, for which closed-form expressions for $H(u)$ and $H(v)$ are unavailable Huber et al. (2008). Instead, we optimize the expected entropy over the $\bar{n}$ models. Particularly, given a model $\mathcal{M}_j$ and measurements $\mathcal{X}_k$, $u(x)$ is normally distributed according to (8). Then, the value of (differential) entropy is independent of the mean and is given in closed-form as

$$H(u(x \mid \mathcal{X}_k, \mathcal{M}_j)) = \log(c\,\gamma_u(x \mid \mathcal{X}_k, \mathcal{M}_j)), \tag{16}$$

3. See Sections III-A and III-B in Khodayi-mehr and Zavlanos (2021) for details on the mobile robot design and measurement noise, respectively.

**noend 1** Next Best Measurement Location

**Require:** Covariance matrix $\bar{\Sigma}_{\Omega\Omega}$, the sets $\mathcal{X}_k$ and $\mathcal{S}_{k+1}$, and current model probabilities $p_{j,k} = \tilde{\pi}(\mathcal{M}_j \,|\, \mathcal{X}_k)$;

1: **for** $x \in \mathcal{S}_{k+1}$ **do**
2: **for** $j = 1 : \bar{n}$ **do**
3: Compute $\delta I_H(x \,|\, \mathcal{X}_k, \mathcal{M}_j)$ using (20);
4: Set $x^*_{k+1} = \arg\max_{x \in \mathcal{S}_{k+1}} \sum_{j=1}^{\bar{n}} p_{j,k}\, \delta I_H(x \,|\, \mathcal{X}_k, \mathcal{M}_j)$;

where $c = \sqrt{2\pi e}$ and $\gamma_u^2(x \,|\, \mathcal{X}_k, \mathcal{M}_j)$ is defined in (8b). Given (16), the expected entropy for $u(x)$ is given by

$$H(u(x \,|\, \mathcal{X}_k)) = \sum\nolimits_{j=1}^{\bar{n}} p_{j,k}\, H(u(x \,|\, \mathcal{X}_k, \mathcal{M}_j)), \tag{17}$$

where $p_{j,k} = \tilde{\pi}(\mathcal{M}_j \,|\, \mathcal{X}_k)$ denotes the probability of model $\mathcal{M}_j$, given the current measurements $\mathcal{X}_k$, obtained from (9). A similar expression holds true for $v(x)$. Then, from (15)

$$H(x \,|\, \mathcal{X}_k) = \sum\nolimits_{j=1}^{\bar{n}} p_{j,k}\, \log\left[c^2\, \gamma_u(x \,|\, \mathcal{X}_k, \mathcal{M}_j)\, \gamma_v(x \,|\, \mathcal{X}_k, \mathcal{M}_j)\right] \tag{18}$$

and we can rewrite the planning problem (14) explicitly as

$$x^*_{k+1} = \arg \max_{x \in \mathcal{S}_{k+1}} \sum\nolimits_{j=1}^{\bar{n}} p_{j,k}\, \delta I_H(x \,|\, \mathcal{X}_k, \mathcal{M}_j), \tag{19}$$

where

$$\delta I_H(x \,|\, \mathcal{X}_k, \mathcal{M}_j) \propto \gamma_u^2(x \,|\, \mathcal{X}_k, \mathcal{M}_j)\, \gamma_v^2(x \,|\, \mathcal{X}_k, \mathcal{M}_j) \tag{20}$$

measures the information added by a potential measurement at $x \in \mathcal{S}_{k+1}$, given current measurements $\mathcal{X}_k$ and model $\mathcal{M}_j$.

Algorithm 1 summarizes our proposed solution to the planning problem at step $k+1$. Note that Algorithm 1 is suboptimal in that it only maximizes the information content of the next immediate measurement. It is possible to consider a longer horizon although at the expense of exponentially increased computational cost. Note also that Algorithm 1 exhaustively evaluates the objective in (20) for all candidate measurement locations in $\mathcal{S}_{k+1}$. Given that the domain $\bar{\Omega}$ is generally non-convex, the prior uncertainty fields (3) are highly nonlinear, and the planning objective (20) can be computed efficiently, this exhaustive approach is effective in practice and there is no need to formulate and solve sophisticated optimization problems. In the relevant literature, more sophisticated planning algorithms are often only studied for simple convex domains and constant prior uncertainty fields for which the planning problem reduces to a simple exploration; see e.g. Xu et al. (2013); Duecker et al. (2019).

## 4. Experimental Results

In this section, we demonstrate the robustness of our proposed framework to significant uncertainties present in the real-world by considering an experiment in a $2.2 \times 2.2 \times 0.4\,\mathrm{m}^3$ domain with an inlet, an outlet, and an obstacle inside as shown in Figure 1; the origin of the coordinate system is located at the bottom left corner of the domain.[4]

4. Extensive numerical and experimental results illustrating scalability of the proposed method and comparing its performance to purely data-driven approaches are presented in Section IV in Khodayi-mehr and Zavlanos (2021). They are omitted from this manuscript due to space limitations.

| No. | model | $q_{\text{in}}$ (m/s) | $i_{\text{in}}$ | $\bar{\sigma}_{u,0}$ (m/s) | $\bar{\sigma}_{i,0}$ |
|---|---|---|---|---|---|
| 1 | $k-\epsilon$ | 0.78 | 0.02 | 0.10 | 0.05 |
| 2 | RSM | 0.78 | 0.02 | 0.20 | 0.10 |
| 3 | $k-\omega$ | 0.78 | 0.02 | 0.20 | 0.10 |
| 4 | $k-\epsilon$ | profile | 0.02 | 0.14 | 0.07 |
| 5 | RSM | profile | 0.02 | 0.20 | 0.10 |
| 6 | $k-\omega$ | profile | 0.02 | 0.20 | 0.10 |
| 7 | RSM | profile | 0.05 | 0.14 | 0.07 |
| 8 | RSM | profile | 0.03 | 0.20 | 0.10 |
| 9 | RSM | profile | 0.01 | 0.20 | 0.10 |
| 10 | RSM | profile | 0.04 | 0.20 | 0.10 |
| 11 | RSM | 0.76 | 0.03 | 0.20 | 0.10 |
| 12 | RSM | 0.80 | 0.03 | 0.14 | 0.07 |

**Table 1:** BCs and prior uncertainty values for the pool of numerical solutions obtained for combinations of solvers and uncertain parameters.

We use a fan to generate a flow at the inlet with average velocity $q_{\text{in}} = 0.78\,\text{m/s}$ and utilize a custom-built mobile robot to conduct the experiment; see Sections III-A nd III-B in Khodayi-mehr and Zavlanos (2021) for details on the design of the robot and measurement model. We assume uncertainty in the inlet velocity and turbulent intensity values. This results in $\bar{n} = 12$ different combinations of BCs and RANS models; see the first four columns in Table 1 for details. In the third column, 'profile' refers to cases where the inlet velocity is modeled by an interpolated function instead of the constant value $q_{\text{in}} = 0.78\,\text{m/s}$. Columns 5 and 6 show the prior uncertainty in the solutions of the first velocity component and turbulent intensity, where we set $\bar{\sigma}_{v,0} = \bar{\sigma}_{u,0}$; see equation (3). As discussed in Section 2.1, these values should be selected to reflect the uncertainty in the numerical solutions. Here, we use the residual values provided by ANSYS FLUENT as an indicator of the confidence in each numerical solution.

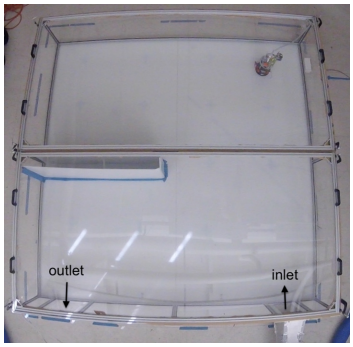
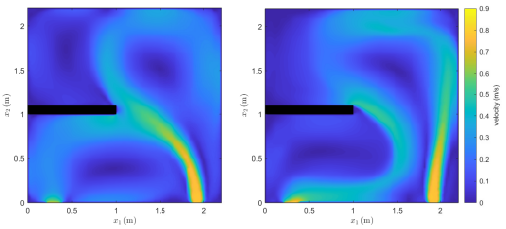



**Figure 1:** Domain of the experiment. A $2.2 \times 2.2 \times 0.4\text{m}^3$ box with velocity inlet at bottom right and outlet at bottom left. The origin of the coordinate system is located at the bottom left corner.

Figure 2 shows the velocity magnitude fields obtained using models 1 and 2. The former is obtained using the $k-\epsilon$ model whereas the latter is obtained using the RSM, as reported in Table 1. Note that these two solutions are inconsistent and require experimental validation to determine the correct flow pattern. We use $\bar{m} = 9$ initial exploration measurements and set the maximum number of measurements to be $m = 200$ and the maximum travel distance to $\mathcal{R} = 1\,\text{m}$. Figure 3 shows the sequence of first 30 waypoints selected by Algorithm 1. The green dots in Figure 3 show the 1206 candidate measurement locations, collected in the set $\Omega$, and the yellow stars show the $\bar{m} = 9$ initial exploration measurement locations selected over a lattice. The black dots show the sequence of waypoints returned by Algorithm 1. Figure 4 shows the added information using the entropy metric (20) after the addition of each of these measurements. It can be observed that the amount of added information generally decreases as the mobile robot keeps adding more measurements. This is expected by the submodularity of the entropy information metric (20). The oscillations

in this figure are due to the travel distance constraint that might prevent the selection of the most informative measurement location at every step $k$. Figure 5 shows the collected velocity vector measurements. Referring to Figure 2, observe that this vector field qualitatively agrees with model 2 that was obtained using the RSM.

The posterior probabilities $p_{j,k}$ converge shortly after the exploration measurements are collected and do not change afterwards. Particularly, the numerical solution from the RSM model 7 is the only solution to have nonzero probability, i.e., $p_{7,155} = 1$. This means that the most accurate model can be selected given a handful of measurements that determine the general flow pattern. It is important to note that all solutions provided by RSM share a similar pattern and the empirical data help to select the most accurate model. Note also that these posterior probabilities are computed given 'only' the available models listed in Table 1 and they should be interpreted with respect to these models and not as absolute probability values.



$(a)$ $k-\epsilon$ model $(b)$ RSM

**Figure 2:** Predictions of the velocity magnitude field according to models 1 and 2 in the plane of the mobile sensor located at the height of 0.27 cm.

In Figure 6, the prior velocity magnitude and turbulent intensity fields corresponding to the most likely model 7 and the posterior fields, computed using equations (12), are given. Comparing the prior and posterior velocity fields, we observe a general increase in velocity magnitude at the top-left part of the domain indicating that the flow sweeps the whole domain unlike the prior prediction from model 7; see also Figure 5. Furthermore, comparing the prior and posterior turbulent intensity fields, we observe a considerable increase in turbulent intensity throughout the domain. Referring to Table 1, note that among all RSM models, model 7 has the highest turbulent intensity BC. In Experiment (b), a visualization of the flow field is shown that validates the flow pattern depicted in Figures 5 and 6.

To evaluate the prediction performance of the posterior model, we collect $\hat{m} = 100$ new measurements at randomly selected locations. We define the total mean prediction error as

$$e_k = \frac{1}{3}(e_{u,k} + e_{v,k} + q_{\text{ref}}\, e_{i,k}), \tag{21}$$

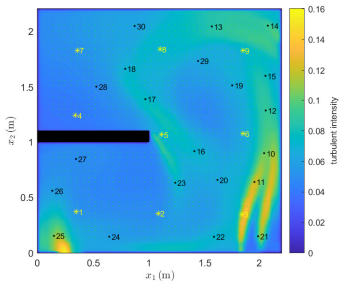


**Figure 3:** Path of the mobile sensor according to the entropy metric (20) overlaid on the turbulent intensity field from model 7.

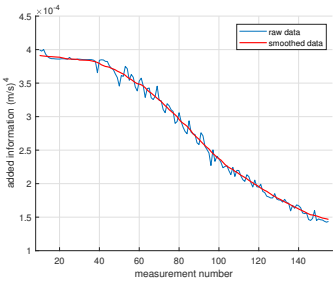



**Figure 4:** Added information vs measurement number for entropy metric (20).

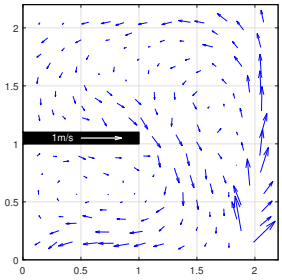
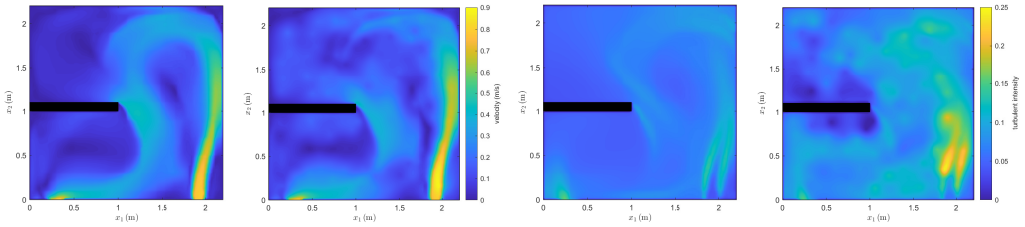



**Figure 5:** Velocity vector measurements for the experiment.



(a) Prior velocity magnitude (b) Posterior velocity magnitude (c) Prior turbulent intensity (d) Posterior turbulent intensity

**Figure 6:** The prior fields for the most likely model 7 and the posterior fields obtained using equations (12) after conditioning on empirical data.

where $e_{u,k} = 1/\hat{m} \sum_{l=1}^{\hat{m}} |\mu_u(x_l \mid \mathcal{X}_k) - \hat{y}_u(x_l)|$ and the expressions for $e_{v,k}$ and $e_{i,k}$ are identical. We also define $\bar{e}_{k,j}$ for the individual models $\mathcal{M}_j$ by using $\mu_u(x_l|\mathcal{X}_k, \mathcal{M}_j)$ in definition (21) instead of the mean value from (12a); similarly for $v(x)$ and $i(x)$. Using equation (21), the prior prediction error is $e_0 = 0.092$ m/s while the posterior error is $e_{155} = 0.037$ m/s, a 60% improvement compared to $e_0$. Moreover, given the posterior knowledge that model 7 is the most likely model, we have $\bar{e}_{0,7} = 0.052$ m/s which is still 29% higher than the posterior error value $e_{155}$. This demonstrates that the real flow field can be best predicted by systematically combining physical models and empirical data.[5] In Figure 7, we plot separately for $u, v$ and $i$, the prior errors of individual models as well as the prior and posterior models. It can be seen that the solutions using the RSM models, including model 7, generally have smaller errors; see also Table 1.

## 5. Conclusion

We proposed a physics-based method to learn environmental fields using a mobile robot. Specifically, we constructed GP models of the flow properties and used numerical simulations to inform their prior mean. Then, utilizing Bayesian inference, we incorporated measurements of flow properties into these GPs. To collect the measurements, we controlled a custom-built mobile robot sensor through a sequence of waypoints that maximize the information content of the measurements. We showed that, compared to purely data-driven methods that are common in the literature, our method can produce high-fidelity global estimations using only sparse measurements. To the best of our knowledge, this is the first physics-based framework for active learning of environmental flow fields that has also been effectively demonstrated in practice. An additional contribution of this work is that it provides new insights into how physics-based models can be efficiently used to learn high-fidelity statistical models of complex dynamical systems modeled by PDEs using only sparse measurements.

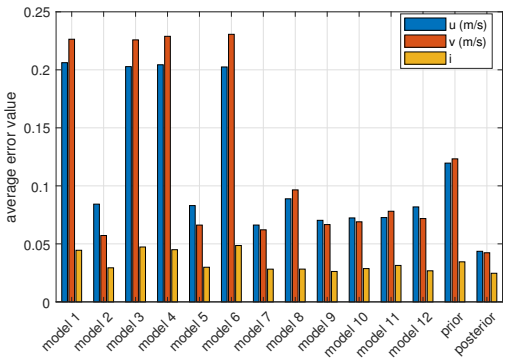



**Figure 7:** Prior error values for individual models along with their averaged prior as well as the posterior errors $e_{u,155}, e_{v,155}$, and $e_{i,155}$.

5. Note that, in practice, knowledge of the best model is not available *a priori*.

## Acknowledgments

This work is supported in part by ONR under agreement #N00014-18-1-2374, and by NSF under awards CNS-1837499 and CNS-1932011.

## References

Felix Berkenkamp and Angela P Schoellig. Safe and robust learning control with Gaussian processes. In *European Control Conference*, pages 2496–2501. IEEE, 2015.

Felix Berkenkamp, Riccardo Moriconi, Angela P Schoellig, and Andreas Krause. Safe learning of regions of attraction for uncertain, nonlinear systems with Gaussian processes. In *Proceedings of IEEE Conference on Decision and Control*, pages 4661–4666. IEEE, 2016.

Luke Calkins, Reza Khodayi-mehr, Wilkins Aquino, and Michael Zavlanos. Stochastic model-based source identification. In *Proceedings of IEEE Conference on Decision and Control*, pages 1272–1277, 2017.

Daniel Andre Duecker, Andreas Rene Geist, Edwin Kreuzer, and Eugen Solowjow. Learning environmental field exploration with computationally constrained underwater robots: Gaussian processes meet stochastic optimal control. *Sensors*, 19(9):2094, 2019.

Matthew Dunbabin and Lino Marques. Robots for environmental monitoring: Significant advancements and applications. *IEEE Robotics & Automation Magazine*, 19(1):24–39, 2012.

Joseph R Edwards, Joshua Smith, Andrew Girard, Diana Wickman, Pierre FJ Lermusiaux, Deepak N Subramani, Patrick J Haley, Chris Mirabito, Chinmay S Kulkarni, and Sudip Jana. Data-driven learning and modeling of AUV operational characteristics for optimal path planning. In *OCEANS 2017-Aberdeen*, pages 1–5. IEEE, 2017.

Experiment. Instantaneous velocity vector, a. https://vimeo.com/281186744.

Experiment. Flow visualization, b. https://vimeo.com/281498120.

Charles Freundlich, Philippos Mordohai, and Michael M Zavlanos. Optimal path planning and resource allocation for active target localization. In *Proceedings of American Control Conference*, pages 3088–3093. IEEE, 2015.

Charles Freundlich, Soomin Lee, and Michael M Zavlanos. Distributed active state estimation with user-specified accuracy. *IEEE Transactions on Automatic Control*, June 2017.

Geoffrey A Hollinger, Arvind A Pereira, Jonathan Binney, Thane Somers, and Gaurav S Sukhatme. Learning uncertainty in ocean current predictions for safe and reliable navigation of underwater vehicles. *Journal of Field Robotics*, 33(1):47–66, 2016.

Marco F Huber, Tim Bailey, Hugh Durrant-Whyte, and Uwe D Hanebeck. On entropy approximation for Gaussian mixture random vectors. In *Multisensor Fusion and Integration for Intelligent Systems*, pages 181–188. IEEE, 2008.

Dylan Jones and Geoffrey A Hollinger. Planning energy-efficient trajectories in strong disturbances. *IEEE Robotics and Automation Letters*, 2(4):2080–2087, 2017.

Reza Khodayi-mehr and Michael M Zavlanos. Physics-based learning for robotic environmental sensing. *arXiv preprint arXiv:1812.03894*, 2021.

Reza Khodayi-mehr, Yiannis Kantaros, and Michael M. Zavlanos. Distributed state estimation using intermittently connected robot networks. *IEEE Transactions on Robotics*, 35(3):709–724, June 2019. doi: 10.1109/TRO.2019.2897865.

Andreas Krause and Carlos Guestrin. Nonmyopic active learning of gaussian processes: an exploration-exploitation approach. In *Proceedings of the 24th international conference on Machine learning*, pages 449–456, 2007.

Dhanushka Kularatne, Subhrajit Bhattacharya, and M Ani Hsieh. Time and energy optimal path planning in general flows. In *Robotics: Science and Systems*, June 2016.

Xiaodong Lan and Mac Schwager. Planning periodic persistent monitoring trajectories for sensing robots in Gaussian random fields. In *Proceedings of IEEE International Conference on Robotics and Automation*, pages 2415–2420. IEEE, 2013.

Xiaodong Lan and Mac Schwager. Learning a dynamical system model for a spatiotemporal field using a mobile sensing robot. In *Proceedings of American Control Conference*, pages 170–175. IEEE, 2017.

Ki Myung Brian Lee, Chanyeol Yoo, Ben Hollings, Stuart Anstee, Shoudong Huang, and Robert Fitch. Online estimation of ocean current from sparse GPS data for underwater vehicles. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 3443–3449. IEEE, 2019.

Julia Ling and J Templeton. Evaluation of Machine Learning algorithms for prediction of regions of high Reynolds averaged Navier Stokes uncertainty. *Physics of Fluids*, 27(8):085103, 2015.

Kai-Chieh Ma, Lantao Liu, and Gaurav S Sukhatme. An information-driven and disturbance-aware planning method for long-term ocean monitoring. In *Proceedings of IEEE International Conference on Intelligent Robots and Systems*, pages 2102–2108, 2016.

Linh Nguyen, Sarath Kodagoda, Ravindra Ranasinghe, and Gamini Dissanayake. Mobile robotic sensors for environmental monitoring using Gaussian Markov random field. *Robotica*, pages 1–23, 2020.

Chris J Ostafew, Angela P Schoellig, and Timothy D Barfoot. Learning-based nonlinear model predictive control to improve vision-based mobile robot path-tracking in challenging outdoor environments. In *Proceedings of IEEE International Conference on Robotics and Automation*, pages 4029–4036. IEEE, 2014.

Carl Edward Rasmussen and Christopher KI Williams. *Gaussian processes for Machine Learning*, volume 1. Cambridge: MIT press, 2006.

Hongchuan Wei, Wenjie Lu, Pingping Zhu, Silvia Ferrari, Robert H Klein, Shayegan Omidshafiei, and Jonathan P How. Camera control for learning nonlinear target dynamics via Bayesian nonparametric Dirichlet-process Gaussian-process (DP-GP) models. In *Proceedings of IEEE International Conference on Intelligent Robots and Systems*, pages 95–102. IEEE, 2014.

David C Wilcox. *Turbulence modeling for CFD*, volume 2. DCW Industries La Canada, CA, 1993.

Yunfei Xu, Jongeun Choi, Sarat Dass, and Tapabrata Maiti. Efficient bayesian spatial prediction with mobile sensor networks using Gaussian Markov random fields. *Automatica*, 49(12):3520–3530, 2013.

Junku Yuh. Design and control of autonomous underwater robots: A survey. *Autonomous Robots*, 8(1):7–24, 2000.