Physics informed Gaussian Process Regression for Particle Tracking Data Assimilation

John M. Lawson[1]

[1] *Department of Aeronautics and Astronautics, University of Southampton, SO17 1BJ, UK**

(Dated: December 18, 2025)

We introduce a physics-informed Gaussian Process Regression (GPR) method for data assimilation and uncertainty quantification of Particle Tracking Velocimetry (PTV) data. Unlike traditional methods based on regression, our approach transparently incorporates statistical information and physics such as mass conservation, boundary conditions, and statistical symmetries directly into the regression model. Furthermore, GPR quantifies prediction uncertainty and provides physics-constrained estimates of the two-point velocity covariance, a quantity of primary interest in turbulent flows. The methodology is demonstrated using synthetic and experimental data from three canonical turbulent flows: homogeneous isotropic turbulence (HIT), turbulent channel flow (TCF), and the turbulent wake behind a square prism (SPW). In all cases, we make comparisons relative to the performance of the vortex in cell method, VIC+. For HIT, the model leverages isotropy to learn the velocity correlation function from even very noisy and sparse data, achieves a factor of two improvement over VIC+ in velocity prediction error, and accurately quantifies the prediction uncertainty. For TCF, we introduce a novel and scalable approach to train a high-dimensional GP model that respects wall-bounded flow physics. GPR significantly outperforms VIC+ in terms of accuracy, uncertainty estimation, and resolution in this case. In the SPW case, GPR demonstrates improved accuracy in velocity prediction and improved coherence of the vorticity field obtained from independent snapshots of tracers. Our approach lays the groundwork for extensions to time-resolved data, inclusion of acceleration measurements, and reduced-parameter models based on resolvent analysis.

I. INTRODUCTION

The past decade has seen widespread adoption of Lagrangian particle tracking (LPT) and particle tracking velocimetry (PTV) to measure flows [1]. These techniques are capable of tracking hundreds of thousands of tracer particles in space and time, which provide access to velocity information sampled at irregular locations in the flow along particle trajectories. Its advantages include a fundamental improvement in spatial resolution over conventional particle image velocimetry techniques, because flow information can be resolved to the position of individual particles with sub-pixel precision [2]. A central issue is the interpolation of these scattered data, usually onto Cartesian grids, for subsequent analysis with the least possible error [1, 3]. It is desirable to assimilate known flow physics, e.g. mass continuity and boundary conditions, to improve the accuracy of the reconstruction of the flow field. Furthermore, it is desirable to quantify the uncertainty in the reconstruction. Often, the reconstruction process is used to facilitate further statistical analysis. This commonly includes quantifying the mean velocity field, Reynolds stresses, two-point correlations and identification of proper orthogonal decomposition (POD) modes.

As showcased in the first LPT data assimilation challenge [3], the majority of solutions to this problem so far are based upon formulating and solving nonlinear regression problems [4–9]. These minimise a cost function which penalises the residual between observations of velocity (optionally, acceleration) sampled on particle trajectories. The minimisation is performed with respect to a set of weights which parametrise the velocity and Lagrangian acceleration fields. The velocity field is usually a linear function of the weights, whereas the acceleration field (if evaluated) has a quadratic nonlinearity. Penalties may be introduced into the cost function to enforce smoothness of the solution [5, 7] or to introduce soft constraints, e.g. boundary conditions [4, 10] and mass and momentum conservation [4, 5, 7]. Alternatively, hard constraints may be introduced either in the formulation of the model [7, 8, 10] or in the determination of the model coefficients [6, 11].

By incorporating these physical constraints, spatial resolution can be improved by a factor of three to four over naïve linear interpolation [3]. For data assimilation with data at a single time instant, the majority of this improvement is attributed to the incorporation of acceleration data [5, 8]. This limits the benefits for two-pulse or four-pulse LPT [12, 13], where acceleration information is not available or may be too noisy to improve the solution. Moreover, these approaches neglect the statistical information available to predict the flow field and do not quantify the uncertainty in the reconstruction. This uncertainty can be large: for a cylinder wake flow inside a turbulent boundary layer, Sciacchitano et al. [3] reported typical velocity errors between 3 and 12% of the bulk velocity.

* j.m.lawson@soton.ac.uk

Gaussian process regression (GPR), or Kriging, is a well-established spatial interpolation technique which aims to predict the conditional expectation and variance of a variable at an unobserved location (e.g. the velocity field at a point) given some other random variables (e.g. the velocity sampled on particle tracks) which are known [14, 15]. This prescribes a linear regression model where the velocity field is a linear function of the input data. The Gaussian process (GP) framework provides a means to obtain the optimal coefficients for any such linear model which is unbiased and minimises the mean squared error of the prediction, i.e. the best linear unbiased prediction [14, 15]. Therefore, we expect suitably constructed GP models to outperform other linear regression models [4]. Furthermore, GPR provides a measure of the prediction uncertainty, i.e. the conditional variance of the prediction given the measured data. Linear and non-linear constraints can be incorporated into the regression model to ensure that the resultant estimate of the velocity field satisfies known physics [16–19]. The regression model can also be formulated incorporate to known statistical symmetries.

GPR is closely related to linear stochastic estimation (LSE) [20] and extended POD [21]. Stochastic estimation has long been used within the fluid mechanics community for the identification and interpretation of coherent flow features [20, 22, 23], whereas extended POD has found application in the interpolation of flow fields from sparse measurements [24, 25]. All techniques implement a shallow autoencoder neural network which provides an interpretable explanation of the output velocity field in terms of POD coefficients of the input data and POD modes of the output. This provides an advantage over "black-box" deep learning approaches, which are harder to interpret [23]. There are two significant distinctions between GPR and extended POD. The first is that GPR provides an estimation of the uncertainty in addition to the conditional mean. The second is that extended POD and LSE are typically performed with a truncated set of POD modes which are not derived from the sparse measurements [24, 25]. In contrast, GPR predictors typically build the covariance model and mean field from the sparse input data themselves.

Recently, several works have addressed the mean flow estimation problem for PTV data [11, 26, 27]. In contrast, the velocity covariance is a much more complex quantity to learn: a 3×3 tensor field in up to six spatial dimensions and two temporal dimensions. Since this is not known *a-priori*, it must be learned from the data by fitting it to a model. The construction and training of this model is the main obstacle to implementing GPR for PTV data assimilation. The problem is twofold. Firstly, the number of training data are large: a typical PTV measurement might yield tens of thousands of point velocity measurements at a single time instant, of which there might be thousands. Since the computational cost of training and inference scale cubically with the number of data points, approximation strategies must be employed [28]. Secondly, the formulation of the model itself is an open question. Very recently, Tirelli et al. [29] have begun to address this by fitting unconstrained radial basis functions to PTV data to learn the covariance. However, no flow physics are encoded in the model and as a consequence there are a very large number of hyperparameters to learn.

To summarise: Gaussian process regression provides an attractive alternative to the dominant PTV data assimilation approaches, because it can transparently leverage statistical information to interpolate scattered data optimally whilst incorporating physics-based constraints and quantifying measurement uncertainty. When constructing a GP model for PTV data, estimates of the two-point velocity covariance function and mean velocity field are learned from the data. Therefore, a GP model for PTV data assimilation necessarily contains quantities of primary interest: estimates of the mean velocity field and two-point velocity covariance, which can be used to obtain Reynolds stresses and POD modes of the flow [30].

In this article, we demonstrate how to construct interpretable, physics-informed Gaussian process models for the reconstruction and uncertainty quantification of incompressible velocity fields obtained from PTV data. We restrict ourselves to the context of GPR with PTV data which are not time-resolved, i.e. the input data correspond to a single time instant. This targets the case of reconstructing velocity fields from two-pulse or four-pulse PTV data, where acceleration information is not available or is too noisy to be useful. We develop models for two prototypical flows: homogeneous isotropic turbulence (HIT) and turbulent channel flow (TCF). These cases demonstrate how to encode fundamental statistical symmetries and incorporate linear constraints such as boundary conditions and incompressibility into the model. We test the accuracy of these models using synthetic particle tracking data obtained from direct numerical simulations. In both cases, we quantify the accuracy of the GPR predictions of the velocity field and its estimated uncertainty, as well as the error in the estimates of the mean flow and two-point velocity covariance. As a real-world test of the method, we train and predict dense flow fields in the wake of a square prism using the GP model derived for turbulent channel flow. In all cases, we compare the performance of the GP predictors to interpolation using VIC+ under varying levels of seeding concentration and measurement noise.

The article is structured as follows. We revise Gaussian process regression in section §II A and revise methods to build models with linear PDE, boundary condition and statistical symmetry constraints in §II C. We then describe how to construct and train physics-informed models of the velocity covariance and mean flow in HIT and TCF in sections §II D- §II G. These models are trained on synthetic and experimental datasets described in §II H. We quantify the accuracy of GPR prediction, uncertainty estimation and the estimated velocity covariance in §III. We present conclusions and outlook for future work in §IV.

II. METHODOLOGY

A. Gaussian Process Regression

To apply GPR to PTV data, we proceed as follows. Consider the Reynolds decomposition of the instantaneous velocity field $\boldsymbol{U}(\boldsymbol{x},t) = \overline{\boldsymbol{U}} + \boldsymbol{u}$ into the mean flow field $\overline{\boldsymbol{U}}(\boldsymbol{x})$ and velocity fluctuation $\boldsymbol{u}(\boldsymbol{x},t)$. We will treat the statistics of this field as stationary in time t. The spatial coordinate within the measurement domain Ω is $\boldsymbol{x} \in \Omega \subset \mathbb{R}^3$. PTV data are available for each of $t = 1 \dots T$ snapshots, with N_t particles per snapshot at input points $\boldsymbol{X}_t = \{\boldsymbol{x}_{n,t}\}_{n=1}^{N_t}$. These data are observations $\boldsymbol{y}_t = (\boldsymbol{v}_1; \boldsymbol{v}_2; \dots; \boldsymbol{v}_{N_t}) \in \mathbb{R}^{3N_t}$ of the velocity fluctuation $\boldsymbol{v}_{n,t} = \boldsymbol{u}(\boldsymbol{x}_n, t) + \boldsymbol{\epsilon}_{n,t}$. We treat the additive Gaussian noise $\boldsymbol{\epsilon}_{n,t}$ noise as independent of the data, statistically stationary, spatially white and isotropic with variance σ^2.

We will suppose that these realisations are statistically independent of one another, so inference of the underlying velocity field at time t is performed using data from time t only. The Gaussian process framework provides the means to predict the distribution of the underlying random field $\boldsymbol{z} = (\boldsymbol{u}(\boldsymbol{x}_{1\star}, t); \dots; \boldsymbol{u}(\boldsymbol{x}_{Q\star}, t))$ at Q query points $\boldsymbol{X}_\star = \{\boldsymbol{x}_{q\star}\}_{q=1}^{Q}$ denoted with the subscript $\star$. The velocity field and noise are treated as joint Gaussian random processes. The conditional mean of $\boldsymbol{z}$ given the observations $\boldsymbol{X}_t, \boldsymbol{y}_t$ is

$$\boldsymbol{\mu}_{\boldsymbol{z}} = E[\boldsymbol{z}|\boldsymbol{y}_t] = \mathbf{K}_{\boldsymbol{z}\boldsymbol{y}_t}\mathbf{K}_{\boldsymbol{y}_t\boldsymbol{y}_t}^{-1}\boldsymbol{y}_t \tag{1}$$

whereas the conditional variance is

$$\boldsymbol{\Sigma}_{\boldsymbol{z}\boldsymbol{z}} = \mathrm{Cov}[\boldsymbol{z}|\boldsymbol{y}_t] = \mathbf{K}_{\boldsymbol{z}\boldsymbol{z}} - \mathbf{K}_{\boldsymbol{z}\boldsymbol{y}_t}\mathbf{K}_{\boldsymbol{y}_t\boldsymbol{y}_t}^{-1}\mathbf{K}_{\boldsymbol{y}_t\boldsymbol{z}} \tag{2}$$

where $\mathbf{K}_{\boldsymbol{y}_t\boldsymbol{y}_t}$, $\mathbf{K}_{\boldsymbol{z}\boldsymbol{z}}$ and $\mathbf{K}_{\boldsymbol{z}\boldsymbol{y}_t} = \mathbf{K}_{\boldsymbol{y}_t\boldsymbol{z}}^\dagger$ represent the (modelled) prior covariance of the observations and the underlying velocity field at the query points. Equation (1) provides the best, linear unbiased prediction of $\boldsymbol{z}$ in the sense that it minimises the sum-of-squares error $\|\boldsymbol{\mu}_{\boldsymbol{z}} - \boldsymbol{z}\|_2^2$ when the covariances are known exactly. The measurement uncertainty, i.e. the variance of $\boldsymbol{z}$ which cannot be explained by the covariates in $\boldsymbol{y}_t$, is given by the diagonal entries of $\boldsymbol{\Sigma}_{\boldsymbol{z}\boldsymbol{z}}$. These expressions for the conditional mean (1) and variance (2) of $\boldsymbol{z}$ hold regardless of whether the underlying process is in fact Gaussian [14, 15].

Alternatively, we can express the GP from the weight-space view as a Bayesian linear model

$$\boldsymbol{u}(\boldsymbol{x}) = \boldsymbol{\Phi}^\dagger(\boldsymbol{x})\boldsymbol{w} \tag{3}$$

where $\boldsymbol{w} \sim \mathcal{N}(0, \mathbf{K}_{\boldsymbol{w}\boldsymbol{w}})$ are a set of normally distributed weights with covariance $\mathbf{K}_{\boldsymbol{w}\boldsymbol{w}}$ and $\boldsymbol{\Phi}(\boldsymbol{x})$ is the feature vector which encodes the input into a higher-dimensional space. An advantage of this formulation is that it is generative: we can create synthetic realisations of the Gaussian process with the same covariance structure as our training data by drawing samples of $\boldsymbol{w}$. We can represent this process as being measured on the input points $\boldsymbol{X}_t$ as

$$\boldsymbol{y}_t = \boldsymbol{\Phi}_t^\dagger \boldsymbol{w}_t + \boldsymbol{\epsilon}_t \tag{4}$$

Given the Gaussian prior on the distributions of the weights and the noise, the solution for the weights which maximises the log-likelihood

$$\log p(\boldsymbol{y}_t|\mathbf{K}_{\boldsymbol{w}\boldsymbol{w}}, \sigma^2) = -\frac{1}{2}\boldsymbol{w}^\dagger\mathbf{K}_{\boldsymbol{w}\boldsymbol{w}}^{-1}\boldsymbol{w} - \frac{1}{2\sigma^2}\|\boldsymbol{y}_t - \boldsymbol{\Phi}_t^\dagger\boldsymbol{w}\|_2^2 + \text{const.} \tag{5}$$

is

$$\boldsymbol{\mu}_{\boldsymbol{w}} = \mathbf{K}_{\boldsymbol{w}\boldsymbol{w}}\boldsymbol{\Phi}_t\mathbf{K}_{\boldsymbol{y}_t\boldsymbol{y}_t}^{-1}\boldsymbol{y}_t \tag{6}$$

with $\mathbf{K}_{\boldsymbol{y}_t\boldsymbol{y}_t} = \boldsymbol{\Phi}_t^\dagger\mathbf{K}_{\boldsymbol{w}\boldsymbol{w}}\boldsymbol{\Phi}_t + \mathbf{K}_{\boldsymbol{\epsilon}\boldsymbol{\epsilon}}$ [31]. Equations (6) and (1) provide identical solutions for $\boldsymbol{\mu}_{\boldsymbol{z}} = \boldsymbol{\Phi}_\star^\dagger\boldsymbol{\mu}_{\boldsymbol{w}}$. Equation (4) represents the general form of linear models which choose different features $\boldsymbol{\Phi}(\boldsymbol{x})$ e.g. splines [5] or radial basis functions [4, 7–9, 32] to represent the velocity field. However, each of these works only retains the sum-of-squares penalty in (5), so prior knowledge of the statistics of the velocity field is not incorporated.

To use (1) and (2), we must first train a model to learn the mean flow field $\overline{\boldsymbol{U}}$, the two-point velocity covariance $\mathbf{R}(\boldsymbol{x}, \boldsymbol{x}') = \overline{\boldsymbol{u}(\boldsymbol{x},t)\boldsymbol{u}^\dagger(\boldsymbol{x}',t)}$ and the measurement noise σ^2. For instance, the covariance of $\boldsymbol{y}_t$ is

$$\mathbf{K}_{\boldsymbol{y}_t\boldsymbol{y}_t} = \begin{bmatrix} \mathbf{R}(\boldsymbol{x}_{1,t}, \boldsymbol{x}_{1,t}) & \mathbf{R}(\boldsymbol{x}_{1,t}, \boldsymbol{x}_{2,t}) & \dots \\ \mathbf{R}(\boldsymbol{x}_{2,t}, \boldsymbol{x}_{1,t}) & \mathbf{R}(\boldsymbol{x}_{2,t}, \boldsymbol{x}_{2,t}) & \dots \\ \vdots & \vdots & \ddots \end{bmatrix} + \mathbf{K}_{\boldsymbol{\epsilon}\boldsymbol{\epsilon}} \tag{7}$$

where $\mathbf{K}_{\boldsymbol{\epsilon}\boldsymbol{\epsilon}} = \sigma^2\mathbf{I}$ is the covariance of the additive measurement noise. Various techniques exist for learning the mean flow field from PTV data [11, 26, 27]. However, the velocity covariance tensor field $\mathbf{R}(\boldsymbol{x}, \boldsymbol{x}')$ is generally a complex thing to learn: a 3×3 tensor field in six spatial dimensions. This tensor field is positive definite, and therefore must be represented as the sum of matrix valued, positive definite kernels $k(\boldsymbol{x}, \boldsymbol{x}'; \boldsymbol{\theta})$ with hyperparameters $\boldsymbol{\theta}$. This condition implies that $\mathbf{K}_{\boldsymbol{zz}}$ is a positive semi-definite matrix for any set of query points.

B. A general model for incompressible flow

To motivate physics informed Gaussian process regression, consider the following "naïve" GP for an incompressible flow [19]. The mean flow $\overline{\boldsymbol{U}}(\boldsymbol{x}) = \boldsymbol{\Phi}^\dagger(\boldsymbol{x})\overline{\boldsymbol{w}}$ and velocity fluctuation field $\boldsymbol{u}(\boldsymbol{x}) = \boldsymbol{\Phi}^\dagger(\boldsymbol{x})\boldsymbol{w}$ are described by a sum of matrix-valued radial basis functions

$$\boldsymbol{\Phi}(\boldsymbol{x}) = \left\{ \phi(r_{\star n})\hat{\boldsymbol{r}}_{\star n}\hat{\boldsymbol{r}}_{\star n}^\dagger + \left(\phi(r_{\star n}) + \frac{1}{2} r_{\star n} \frac{\partial \phi}{\partial r} \right) \left(\mathbf{I} - \hat{\boldsymbol{r}}_\star \hat{\boldsymbol{r}}_\star^\dagger \right) \right\}_{n=1}^{N} \tag{8}$$

centered at N inducing points $\boldsymbol{X}_\star = \{\boldsymbol{x}_\star\}_{n=1}^N$ with weights $\overline{\boldsymbol{w}}, \boldsymbol{w} \in \mathbb{R}^{3N}$ for the mean flow and velocity fluctuations. Here, $r_\star = \|\boldsymbol{x} - \boldsymbol{x}_\star\|_2$ is the distance to each inducing point and $\hat{\boldsymbol{r}}_\star$ is the associated unit vector. These are defined in terms of a scalar RBF

$$\phi(r_\star) = \exp\left(-\frac{r_\star^2}{2\ell^2} \right) \tag{9}$$

for which we choose the popular Gaussian RBF with uniform scale ℓ. Already, this contains some flow physics: (8) always yields $\nabla \cdot \mathbf{u} = 0$, regardless of the choice of weights. This stands in contrast with other RBF based models for PTV data assimilation [4, 5, 7, 9, 11], where the solenoidal constraint is enforced by constraints upon the weights. The number of inducing points scales with the number of particles per snapshot; literature suggests [5, 7, 8, 32] that as many as 5-20 times inducing points as input data are necessary provide enough flexibility to reconstruct small scale flow features.

Under the Gaussian process Ansatz, the prior distribution of the weights is $\boldsymbol{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{K}_{\boldsymbol{ww}})$ with covariance $\mathbf{K}_{\boldsymbol{ww}} = \mathbf{L}_{\boldsymbol{w}}\mathbf{L}_{\boldsymbol{w}}^\dagger$, where $\mathbf{L}_{\boldsymbol{w}}$ is a lower triangular matrix with real and positive diagonal entries. This definition ensures the prior covariance is positive definite. The model has $9N(N+1)/2 + 1$ hyperparameters $\boldsymbol{\theta} = \{\overline{\boldsymbol{w}}, \mathbf{L}_{\boldsymbol{w}}, \sigma^2\}$, which encode the prior mean, covariance and model noise. The conventional approach to training GP models is to maximise the marginal likelihood of the observations

$$J_{rbf}(\boldsymbol{\theta}) = \sum_{t=1}^{T} \log p(\boldsymbol{y}_t|\boldsymbol{\theta}) = -\frac{1}{2} \sum_{t=1}^{T} \left[\boldsymbol{y}_t{}^\dagger \mathbf{K}_{\boldsymbol{y}_t\boldsymbol{y}_t}^{-1} \boldsymbol{y}_t{}^\dagger + \log|\mathbf{K}_{\boldsymbol{y}_t\boldsymbol{y}_t}| + N_t \log 2\pi \right] \tag{10}$$

over the hyperparameters $\boldsymbol{\theta}$ (the dependence on $\overline{\boldsymbol{w}}$ is implicit in the definition of $\boldsymbol{y}_t$).

Large GP models can be very expensive to train using (10). The bottleneck is due to the inversion and determinant of the kernel matrices $\mathbf{K}_{\boldsymbol{y}_t\boldsymbol{y}_t}$, which require $\mathcal{O}\left(TN_t^3\right)$ operations and $\mathcal{O}\left(N_t^2\right)$ memory. Evaluating the gradient with respect to the hyperparameters requires a further $\mathcal{O}\left(TN_tN^2\right) + \mathcal{O}\left(TN^3\right)$ operations. Considering that a typical PTV experiment might track tens of thousands of particles for thousands of timesteps, this can be very costly: with $N = N_t = 50,000$ particles per snapshot and as many inducing points, there are $\sim 10^{10}$ hyperparameters and $\sim 10^{15}$ operations per matrix-matrix multiplication. Therefore, we need a means of reducing the complexity of the model which retains enough flexibility to fit the data but has fewer hyperparameters to learn.

C. Physics Informed Covariance Models

The premise of physics informed Gaussian process regression is to construct a covariance model which satisfies known statistical symmetries, boundary conditions and incompressibility constraints. This not only ensures that the GPR prediction (1) satisfies the boundary conditions and incompressibility, but it also reduces the degrees of freedom within the covariance model so that it may be learned empirically. For instance, statistical homogeneity implies $\mathbf{R}(\boldsymbol{x}, \boldsymbol{x}') = \mathbf{R}(0, \boldsymbol{r})$ where $\boldsymbol{r} = \boldsymbol{x}' - \boldsymbol{x}$, reducing its spatial dimension by three. It can be seen from equation (1) that linear constraints upon the covariance tensor imply linear constraints upon the estimated velocity field and vice-versa. For instance, the prediction $\boldsymbol{\mu}(\boldsymbol{x}_\star)$ (1) will satisfy incompressibility $\nabla_\star \cdot \boldsymbol{\mu}(\boldsymbol{x}_\star) = 0$ for arbitrary $\boldsymbol{y}_t$ only if the covariance model satisfies $\nabla_{\boldsymbol{x}} \cdot \mathbf{R}(\boldsymbol{x}, \boldsymbol{x}') = \nabla_{\boldsymbol{x}'} \cdot \mathbf{R}(\boldsymbol{x}, \boldsymbol{x}') = 0$. Boundary conditions can also be applied, e.g. the no-slip condition $\boldsymbol{u}(\boldsymbol{x}_{\partial\Omega}, t) = 0$ for a set of points $\boldsymbol{x}_{\partial\Omega} \in \partial\Omega$ on the boundary can only be satisfied if $\mathbf{R}(\boldsymbol{x}_{\partial\Omega}, \boldsymbol{x}') = 0$.

These physics can be efficiently encoded into the model using the transformed covariance kernel approach [16]. Homogeneous linear operator constraints of the form $\mathcal{L}_{\boldsymbol{x}}\mathbf{R}(\boldsymbol{x},\boldsymbol{x}') = \mathcal{L}_{\boldsymbol{x}'}\mathbf{R}(\boldsymbol{x},\boldsymbol{x}') = 0$ can be enforced by choosing a set of kernels which lie in the null space of the linear operator $\mathcal{L}_{\boldsymbol{x}}$ such that $\mathcal{L}_{\boldsymbol{x}}k(\boldsymbol{x},\boldsymbol{x}';\boldsymbol{\theta}) = \mathcal{L}_{\boldsymbol{x}'}k(\boldsymbol{x},\boldsymbol{x}';\boldsymbol{\theta}) = 0$, regardless of the choice of hyperparameters $\boldsymbol{\theta}$. Statistical homogeneity implies we should choose kernels of the form $k(\boldsymbol{x},\boldsymbol{x}';\boldsymbol{\theta}) = k(\boldsymbol{r};\boldsymbol{\theta})$, whereas isotropy implies $k(\boldsymbol{r};\boldsymbol{\theta})$ is invariant under rotations. In the following sections, we describe how to construct and fit covariance models for the velocity field in two prototypical flows featuring these symmetries: homogeneous isotropic turbulence and turbulent channel flow.

D. A model for homogeneous and isotropic turbulence

In statistically stationary, homogeneous, isotropic turbulence, the two-point, single-time covariance tensor of any solenoidal vector field is defined by a single, scalar valued, longitudinal autocorrelation function $R_{LL}(r)$ as [22]

$$\mathbf{R}(\boldsymbol{x},\boldsymbol{x}') = R_{LL}(r)\hat{\boldsymbol{r}}\hat{\boldsymbol{r}}^\dagger + \left(R_{LL}(r) + \frac{1}{2}r\frac{\partial R_{LL}}{\partial r}\right)\left(\mathbf{I} - \hat{\boldsymbol{r}}\hat{\boldsymbol{r}}^\dagger\right) \tag{11}$$

This defines the correlation between a pair of points separated space by a distance r corresponding to the displacement $\boldsymbol{r} = \boldsymbol{x}' - \boldsymbol{x} = r\hat{\boldsymbol{r}}$ between the two points. The incompressibility condition is satisfied by construction. For convenience, we write $\mathbf{R}(\boldsymbol{r}) = \mathbf{R}(0, \boldsymbol{x}' - \boldsymbol{x})$. The corresponding longitudinal energy spectrum

$$E_{11}(\kappa) = \frac{1}{\pi}\int_{-\infty}^{\infty} R_{LL}(r)e^{-\mathrm{i}\kappa r}\mathrm{d}r \geq 0 \tag{12}$$

is real-valued and non-negative. Furthermore, it is an even function of the wavenumber κ and is monotonic decreasing in κ for $\kappa \geq 0$ [22].

We fit the autocorrelation function to the data with the linear model

$$R_{LL}(r) = \theta_m K_m(r) \tag{13}$$

for the hyperparameters $\theta_m \geq 0$ and scalar valued kernel functions $K_m(r)$. A natural choice is

$$K_m(r) = \mathrm{sinc}(\kappa_m r) \tag{14}$$

corresponding to a uniform spectral energy density over a band of wavenumbers $[-\kappa_m, \kappa_m]$.We note that the matrix valued kernel (11) corresponding to (14) is of the same form proposed by Narcowich and Ward [19] for the interpolation of incompressible vector fields. Here, however, the kernel has a specific physical interpretation in terms of the energy spectrum. The constraint $\theta_m \geq 0$ conveniently ensures that E_{11} is non-negative and monotonic decreasing in κ and that the transverse autocorrelation function

$$R_{NN}(r) = R_{LL} + \frac{1}{2}r\frac{\partial R_{LL}}{\partial r} = K_m^\perp(r)\theta_m \tag{15}$$

is also positive definite. Here, $K_m^\perp(r) = \frac{1}{2}(K_m(r) + \cos(\kappa_m r))$ is the kernel for the transverse autocorrelation function. We note that the covariance model (11) readily generalises to the two-point, two-time covariance by choosing kernels of the form $\mathrm{sinc}(\kappa r)\cos(\omega(t'-t))$ for a set of frequencies ω and wavenumbers κ.

The standard approach to training the hyperparameters is to maximise the marginal probability $p(\boldsymbol{y}|\boldsymbol{\theta})$ of all the observations $\boldsymbol{y} = (\boldsymbol{y}_1; \ldots; \boldsymbol{y}_t)$ given the hyperparameters $\boldsymbol{\theta} = \{\sigma^2, \theta_1, \ldots \theta_M\}$ [15]. Given the T sets of observations $\boldsymbol{y}_t \in \mathbb{R}^{3N_t}$, this naïvely involves T inversions of $3N_t \times 3N_t$ matrices, which is computationally infeasible. Therefore, we take a product-of-experts approach by splitting our input data into a set of $p = 1 \ldots P$ particle pairs. There are N_t particles per snapshot, so $P \sim T\overline{N_t^2}$. Each pair constitutes an observation $\boldsymbol{y}_p = (\boldsymbol{v}(\boldsymbol{x}_p, t); \boldsymbol{v}(\boldsymbol{x}_p', t))$ of the noisy isotropic turbulence process at coordinates $\boldsymbol{x}_p$ and $\boldsymbol{x}_p'$, separated by displacement $\boldsymbol{r}_p = \boldsymbol{x}_p' - \boldsymbol{x}_p$ and measured at the same time instant. In general, these should be drawn from distinct particle tracks to ensure that $\boldsymbol{x}_p$ and $\boldsymbol{x}_p'$ are independent. We approximate the log-likelihood of our model given the observations as

$$J_{iso}(\boldsymbol{\theta}) = \log p(\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots \boldsymbol{y}_P|\boldsymbol{\theta}) = \sum_{p=1}^{P} \log p(\boldsymbol{y}_p|\boldsymbol{\theta}). \tag{16}$$

The log-likelihood of the observation $\boldsymbol{y}_p$ is

$$\log p(\boldsymbol{y}_p|\boldsymbol{\theta}) = -\log 2\pi - \frac{1}{2}\log|\mathbf{K}_{\boldsymbol{y}_p\boldsymbol{y}_p}| - \frac{1}{2}\boldsymbol{y}_p{}^\dagger\mathbf{K}_{\boldsymbol{y}_p\boldsymbol{y}_p}^{-1}\boldsymbol{y}_p \tag{17}$$

where the $\mathbf{K}_{\boldsymbol{y}_p\boldsymbol{y}_p}$ is the modelled covariance of each pair of observations.

After some algebra, the determinant of this matrix is

$$|\mathbf{K}_{\boldsymbol{y}_p\boldsymbol{y}_p}| = \frac{1}{64} S_{LL}(r_p) D_{LL}(r_p) S_{NN}^2(r_p) D_{NN}^2(r_p) \tag{18}$$

where we have introduced the longitudinal and transverse structure functions

$$\begin{aligned} D_{LL}(r) &= 2(R_{LL}(0) - R_{LL}(r)) \quad D_{NN}(r) = 2(R_{NN}(0) - R_{NN}(r)) \\ S_{LL}(r) &= 2(R_{LL}(0) + R_{LL}(r)) \quad S_{NN}(r) = 2(R_{NN}(0) + R_{NN}(r)) \end{aligned} \tag{19}$$

which describe the second moments of velocity differences $\boldsymbol{\Delta}_p = \boldsymbol{v}(\boldsymbol{x}'_p, t_p) - \boldsymbol{v}(\boldsymbol{x}_p, t_p)$ and sums $\boldsymbol{S}_p = \boldsymbol{v}(\boldsymbol{x}_p, t_p) + \boldsymbol{v}(\boldsymbol{x}'_p, t_p)$. The data fit term is

$$\boldsymbol{y}_p{}^\dagger \mathbf{K}^{-1}_{\boldsymbol{y}_p\boldsymbol{y}_p} \boldsymbol{y}_p = \frac{(\boldsymbol{\Delta}_p \cdot \hat{\boldsymbol{r}}_p)^2}{D_{LL}(r_p) + 2\sigma^2} + 2\frac{|\boldsymbol{\Delta}_p|^2 - (\boldsymbol{\Delta}_p \cdot \hat{\boldsymbol{r}}_p)^2}{D_{NN}(r_p) + 2\sigma^2} + \frac{(\boldsymbol{S}_p \cdot \hat{\boldsymbol{r}}_p)^2}{S_{LL}(r_p) + 2\sigma^2} + 2\frac{|\boldsymbol{S}_p|^2 - (\boldsymbol{S}_p \cdot \hat{\boldsymbol{r}}_p)^2}{S_{NN}(r_p) + 2\sigma^2} \tag{20}$$

To accelerate finding the hyperparameters $\boldsymbol{\theta}$ which maximise the log likelihood J_{iso}, we approximate the sum in (16) using a fine-grained histogram of the separation r_p and corresponding observations of longitudinal and transverse velocity components. The maximum is found subject to the linear constraints $\theta_m \geq 0$ and $\sigma \geq 0$ with a standard interior-point method. The cost of this approach is dominated by the $\mathcal{O}(T\overline{N_t^2})$ operations to generate the fine-grained histogram once, which represents a dramatic reduction in complexity compared to the naïve $\mathcal{O}(T\overline{N_t^3})$ cost of maximising the log-likelihood for our full dataset.

The mean flow field is trivial. In periodic box simulations it is identically zero [22]. In grid turbulence [22] or more sophisticated "zero-mean flow" apparatuses [33], temporal variations in the velocity field far exceed spatial variations in the mean flow field within a particular region of interest, so to a good approximation the mean flow is uniform. In general, one might estimate the mean flow field using ensemble PTV methods [11, 26, 27].

E. A model for turbulent channel flow

A less restrictive set of symmetries is embodied by turbulent channel flow: turbulent flow between infinite parallel plates ("walls"). We consider a channel flow with Cartesian coordinates (x_1, x_2, x_3) aligned with the streamwise, wall-normal and spanwise directions respectively, bounded by no-slip walls at $x_1 = \pm h$. This flow possesses statistical stationarity, statistical homogeneity in the spanwise and streamwise directions, and reflection symmetries about the planes $x_1 = 0$ and $x_3 = 0$.

We consider a cuboidal measurement domain $x_1 \in [0, L_1], x_2 \in [a, b], x_3 \in [0, L_3]$. We therefore require a model of the two-point covariance for $x_2, x'_2 \in [a, b]$ and $x'_1 - x_1 \in [-L_1, L_1]$, $x'_3 - x_3 \in [-L_3, L_3]$. Due to homogeneity, the two-point, single-time covariance function is of the form $\mathbf{R}(\boldsymbol{x}, \boldsymbol{x}') = \mathbf{R}(x_2, x'_2, r_1, r_3)$ [30]. We write a covariance model satisfying these symmetries as

$$\mathbf{R}(x_2, x'_2, r_1, r_3) = \sum_{l=-L}^{L} \sum_{n=-N}^{N} \boldsymbol{\Phi}_2^\dagger(x_2) \mathbf{H}_{\boldsymbol{\kappa}} \boldsymbol{\Phi}_2(x'_2) e^{\mathrm{i}\lambda_l r_1 + \mathrm{i}\nu_n r_3} \tag{21}$$

where $\boldsymbol{\Phi}_2(x_2) \in \mathbb{R}^{3M \times 3}$ are a set of B-spline features

$$\boldsymbol{\Phi}_2(x) = \begin{bmatrix} \boldsymbol{B}(x) & 0 & 0 \\ 0 & \boldsymbol{B}(x) & 0 \\ 0 & 0 & \boldsymbol{B}(x) \end{bmatrix}, \ \boldsymbol{B}(x) = \{B_1(x), B_2(x), \dots B_M(x)\}^\dagger \tag{22}$$

derived from M basis splines $B_m(x)$. This represents the covariance tensor as a weighted sum over wavenumbers $\boldsymbol{\kappa} = [\lambda_l, 0, \nu_n]$ with $\lambda_l = \pi l / L_1$ and $\nu_n = \pi n / L_3$ using a combination of spline and Fourier kernels with coefficients $\mathbf{H}_{\boldsymbol{\kappa}} \in \mathbb{C}^{3M \times 3M}$. For each wavenumber, these matrix valued kernels encode spatial inhomogeneity in the wall-normal direction using spline features $\boldsymbol{\Phi}_2(x_2)$. Since $\mathbf{R}$ is real, conjugate symmetry implies that $\mathbf{H}_{-\boldsymbol{\kappa}} = \mathbf{H}_{\boldsymbol{\kappa}}^\dagger$. The eigendecomposition of $\mathbf{H}_{\boldsymbol{\kappa}}$ specifies the proper orthogonal decomposition (POD) modes [30, 34] of the flow. We note that the largest wavelength represented in (21) is *twice* the measurement domain dimension. The model (21) readily generalises to more homogeneous dimensions (e.g. time) by increasing the dimension of the Fourier kernel, or to more inhomogeneous dimensions by increasing the dimension of the spline kernel.

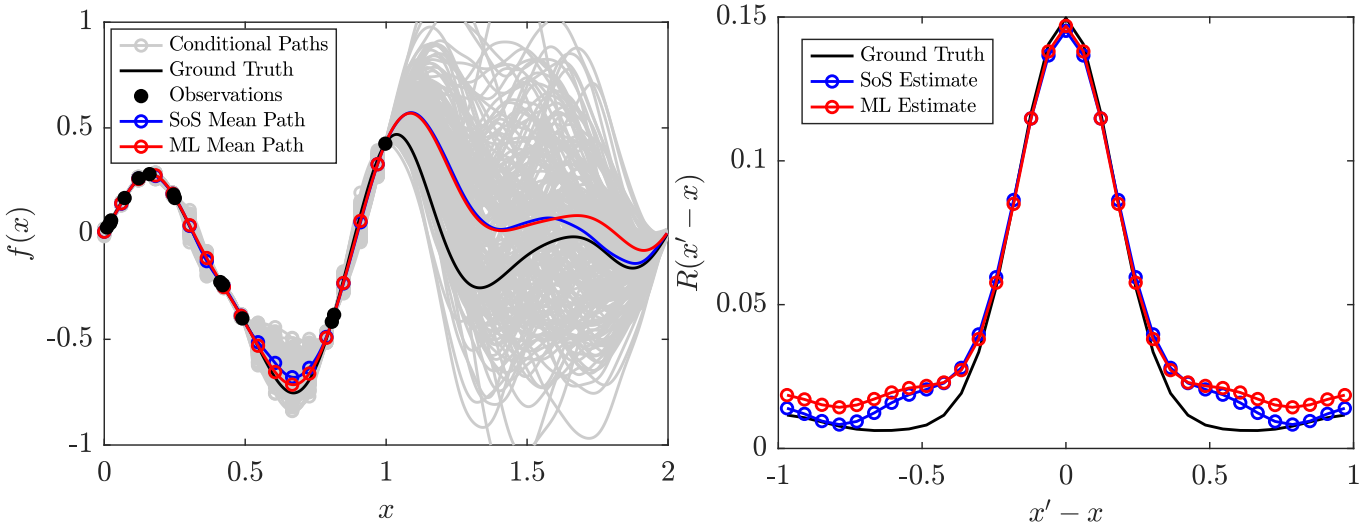

FIG. 1: (a) A realisation of a 1D, stationary Gaussian process with period $2L_1 = 2$ and 16 noisy observations sampled at random locations on the interval $[0, 1]$. Open markers correspond to samples at inducing points. (b) The maximum likelihood (27) and least-squares (SoS) (29) estimates of the autocorrelation function inferred from 100 realisations of this process after 30 epochs, in comparison to ground truth.

The kernel (21) is associated with the weight-space formulation

$$\boldsymbol{u}(\boldsymbol{x}) = \sum_{\boldsymbol{\kappa}} e^{-\mathrm{i}\boldsymbol{\kappa}\cdot\boldsymbol{x}} \boldsymbol{\Phi}_2^\dagger(x_2)\boldsymbol{\pi}_{\boldsymbol{\kappa}} \tag{23}$$

which represents the velocity field as a sum of spline/Fourier features and complex weights $\boldsymbol{\pi}_{\boldsymbol{\kappa}}$ for each wavenumber. Since $\boldsymbol{u}$ is real, weights at wavenumbers $\boldsymbol{\kappa}$ and $-\boldsymbol{\kappa}$ are complex conjugate pairs. The coefficients $\mathbf{H}_{\boldsymbol{\kappa}}$ correspond to the covariance of the weights $\overline{\boldsymbol{\pi}_{\boldsymbol{\kappa}}\boldsymbol{\pi}_{\boldsymbol{\kappa}}^\dagger}$. To express this in terms of a set of real weights $\boldsymbol{w}_{\boldsymbol{\kappa}} \in \mathbb{R}^{3M}$ for each wavenumber, we can choose $\boldsymbol{\pi}_{\boldsymbol{\kappa}} = (\boldsymbol{w}_{\boldsymbol{\kappa}} + \boldsymbol{w}_{-\boldsymbol{\kappa}})/2 + \mathrm{i}(\boldsymbol{w}_{\boldsymbol{\kappa}} - \boldsymbol{w}_{-\boldsymbol{\kappa}})/2$.

Linear constraints $\mathbf{H}_{\boldsymbol{\kappa}}\mathbf{A}_{\boldsymbol{\kappa}} = \mathbf{0}$ upon the coefficients specified by $\mathbf{A}_{\boldsymbol{\kappa}} \in \mathbb{C}^{3M\times P}$ can be introduced to satisfy boundary conditions or incompressibility at discrete points. For instance $\mathbf{H}_{\boldsymbol{\kappa}}\boldsymbol{\Phi}(-h) = 0$ describes the no-slip boundary condition at $x_2 = -h$ whereas

$$\mathbf{A}_{\boldsymbol{\kappa}} = \begin{bmatrix} \mathrm{i}\lambda_l \boldsymbol{B}(x_m) \\ \boldsymbol{B}'(x_m) \\ \mathrm{i}\nu_n \boldsymbol{B}(x_m) \end{bmatrix} \tag{24}$$

imposes the incompressibility constraint $\partial R_{ij}/\partial x'_j = \partial R_{ij}/\partial x_i = 0$ on the plane $x_2 = x_m$. We can construct a set of coefficients $\mathbf{H}_{\boldsymbol{\kappa}} = \mathbf{Q}_{\boldsymbol{\kappa}}\boldsymbol{\Theta}_{\boldsymbol{\kappa}}\boldsymbol{\Theta}_{\boldsymbol{\kappa}}^\dagger\mathbf{Q}_{\boldsymbol{\kappa}}^\dagger$, where $\boldsymbol{\Theta}_{\boldsymbol{\kappa}} \in \mathbb{C}^{(3M-P)\times(3M-P)}$ is a lower-triangular matrix of kernel hyperparameters with real and positive diagonal entries and $\mathbf{Q}_{\boldsymbol{\kappa}} \in \mathbb{C}^{3M\times(3M-P)}$ is an orthonormal matrix chosen so that $\mathbf{Q}_{\boldsymbol{\kappa}}^\dagger\mathbf{A}_{\boldsymbol{\kappa}} = 0$. Formulating the model in this way, although non-linear in the hyperparameters $\boldsymbol{\Theta}_{\boldsymbol{\kappa}}$, automatically satisfies the constraints $\mathbf{H}_{\boldsymbol{\kappa}}\mathbf{A}_{\boldsymbol{\kappa}} = \mathbf{0}$ and guarantees that $\mathbf{H}_{\boldsymbol{\kappa}}$ is positive semi-definite. Furthermore, for each wavenumber, it reduces the number of coefficients needed to specify $\mathbf{H}_{\boldsymbol{\kappa}}$ from $9M^2$ to $(3M - P)^2$.

The mean flow field varies in the wall-normal direction only. The spanwise and wall-normal components are $\overline{U}_3 = \overline{U}_2 = 0$. We fit the streamwise velocity component

$$\overline{U}_1(\boldsymbol{x}) = \overline{U}_{1,m} B_m(x_2) \tag{25}$$

with coefficients $\overline{U}_{1,m}$ using the same set of basis splines by minimising the sum-of-squares loss $J_U = \sum_p (\overline{U}_1(x_{2,p}) - U_1(\boldsymbol{X}_q, t_q))^2$, subject to the no-slip boundary condition $\overline{U}_1 = 0$ at the wall. This mean flow field satisfies incompressibility by construction.

F. Training models with Expectation-Maximisation and the Matheron Update Rule

We use an iterative approach to infer the model hyperparameters based on a modified Expectation-Maximisation (EM) algorithm and Matheron's update rule [35, 36]. For the purpose of exposition, consider the stationary, 1D

Gaussian process generated by $f(x) = \mathbf{\Phi}^\dagger(x)\boldsymbol{w}$ with noisy observations $\boldsymbol{X}_t, \boldsymbol{y}_t$ (4). To make analogy with (21), we pick Fourier features $\mathbf{\Phi}(x) = \exp(-\mathrm{i}\boldsymbol{\nu}x)$ for a set of $2N+1$ wavenumbers $\nu_n = \pi n/L_3$ with $n = -N \dots N$. The weights $w_{-n} = w_n^\dagger$ are complex Fourier coefficients of this periodic process and their covariance $\mathbf{K}_{\boldsymbol{ww}}$ is a real-valued diagonal matrix describing the power spectral density. Thus, we have the autocorrelation function

$$R(x, x') = \sum_{n=-N}^{N} \theta_n \exp(i\nu_n(x' - x)) \tag{26}$$

and constrained hyperparameters $\theta_{-n} = \theta_n \geq 0$. Figure 1a illustrates one such realisation: black dots show the noisy observations sampled on the left half of the domain, whereas the solid black line shows the noise-free realisation of $f(x)$.

The algorithm, described in pseudocode in Appendix A, proceeds as follows. Rather than directly maximising the log likelihood of observations to infer $\boldsymbol{\theta}$ as in (16), we introduce latent variables $\boldsymbol{z}_t = \mathbf{\Phi}_\star^\dagger \boldsymbol{w}_t$ which represent the process sampled at fixed inducing points $\boldsymbol{X}_\star = \{n\Delta x\}_{n=0}^N$, conditional on observations $\boldsymbol{y}_t$. A sample of conditional process paths drawn from $p(f|\boldsymbol{y}_t)$ for our 1D example is illustrated by grey lines in figure 1, with observations at inducing points shown by open markers. By construction, the inducing points cover the region supported by the data ($0 \leq x \leq 1$) and do not include points which are not well informed by the observations. We train the model by generating an empirical sampling of the process $\mathbf{Z}$ and noise $\mathbf{E}$ by making S draws of the latent variable $\boldsymbol{z}_t$ and noise $\boldsymbol{\epsilon}_t$ from each of $t = 1 \dots T$ conditional distributions $p(\boldsymbol{z}_t, \boldsymbol{\epsilon}_t|\boldsymbol{y}_t, \boldsymbol{\theta})$ using the Matheron update rule [35]. Thus, $\mathbf{Z}$ and $\mathbf{E}$ have ST columns. These samples are used to infer updated hyperparameters. The process then repeats, with the updated hyperparameters used to create updated draws of the latent variables at each iteration.

Crucially, the samples correspond to partial observations of the underlying process on *fixed* inducing points. The standard approach to learning the hyperparameters is to maximise the log-likelihood

$$\log p(\mathbf{Z}|\boldsymbol{y}, \boldsymbol{\theta}) = -\frac{1}{2}\mathrm{Tr}\left(\mathbf{Z}^\dagger \mathbf{K}_{\boldsymbol{zz}}^{-1}\mathbf{Z}\right) + -\frac{ST}{2}\log|\mathbf{K}_{\boldsymbol{zz}}| + \text{const.} \tag{27}$$

However, this objective function requires $\mathcal{O}(NST)$ memory to store $\mathbf{Z}$ during the optimisation and costs at least $\mathcal{O}\left(N^2ST\right)$ to evaluate, which very quickly becomes infeasible when the number of realisations T and sampled paths S is large. The maximum likelihood estimate of the noise is simply $\mathrm{Tr}\left(\mathbf{E}^\dagger\mathbf{E}\right)/ST$ which requires only $\mathcal{O}(N)$ storage.

As an alternative to maximum likelihood estimation, consider that we can completely determine the autocorrelation sequence (26) and hence $\boldsymbol{\theta}$ from samples of $\boldsymbol{z} = [z_0, \dots z_N]$ as

$$\widetilde{R}_j = \frac{1}{N - |j| + 1}\sum_{n=0}^{N-|j|} \overline{z_n z_{n+|j|}} \tag{28}$$

This can be implemented with a storage cost $\mathcal{O}(N)$. However, the resulting empirical autocorrelation sequence $\widetilde{R}_j$ is not guaranteed to be positive definite. Therefore, we minimise the weighted sum-of-squares distance

$$J_R(\boldsymbol{\theta}) = \sum_{j=-N}^{N}(N - |j| + 1)(\widetilde{R}_j - R(j\Delta x, 0))^2 \tag{29}$$

subject to the constraints $\theta_n \geq 0$, which costs $\mathcal{O}(N \log N)$. Figure 1 demonstrates that estimates of the autocorrelation $R(x - x')$ for this 1D example obtained from the approximate inference using (29) do not differ much from the exact, maximum likelihood inference (27). Secondly, the maximum likelihood estimates of the process (red and blue lines in Figure 1a) are also similar.

Samples of $\boldsymbol{z}_t$ from its conditional distribution $\mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{z}}, \mathbf{\Sigma}_{\boldsymbol{zz}})$ can be efficiently generated using the Matheron update rule [35], without the need to compute $\boldsymbol{\mu}_{\boldsymbol{z}}$ or $\mathbf{\Sigma}_{\boldsymbol{zz}}$ from (1) and (2) explicitly. Let $\boldsymbol{w}_\star$ be an unconditional realisation of the weights and $\boldsymbol{y}_\star = {\mathbf{\Phi}_t}^\dagger \boldsymbol{w}_\star + \boldsymbol{\epsilon}_\star$ be the associated noisy observation with noise realisation $\boldsymbol{\epsilon}_\star$. Then the conditional realisation $\boldsymbol{z}_{t\star} = \mathbf{\Phi}_\star^\dagger \boldsymbol{w}_{t\star}$ is distributed as $\mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{z}}, \mathbf{\Sigma}_{\boldsymbol{zz}})$ when we choose

$$\boldsymbol{w}_{t\star} = \boldsymbol{w}_\star - \mathbf{K}_{\boldsymbol{w}\boldsymbol{y}_t}\mathbf{K}_{\boldsymbol{y}_t\boldsymbol{y}_t}^{-1}(\boldsymbol{y}_t - \boldsymbol{y}_\star) \tag{30}$$

where $\mathbf{K}_{\boldsymbol{w}\boldsymbol{y}_t} = \mathbf{K}_{\boldsymbol{ww}}\mathbf{\Phi}_t^\dagger$ is the covariance of the weights with the data. Likewise, the conditional realisation of the noise is $\boldsymbol{\epsilon}_{t\star} = \boldsymbol{y}_t - \mathbf{\Phi}_t^\dagger \boldsymbol{w}_{t\star}$. We note that, after training, these samples can also be used for uncertainty propagation using Monte-Carlo methods.

G. Training the turbulent channel flow model

The training of the turbulent channel flow model proceeds analogously to our 1D example. In each iteration, or epoch, we evaluate (30) to draw $S = 201$ samples of the velocity field at a set of inducing points defined on a regular, Cartesian grid for each of the T realisations of the flow. There are $L+1, M$ and $N+1$ inducing points in the x_1, x_2, x_3 directions respectively. The grid is uniformly spaced over in the x_1 and x_3 directions $x_{1,l} = lL_1/(2L+1)$ and $x_{3,n} = nL_3/(2N+1)$ with $l = 0 \dots L$ and $n = 0 \dots N$. The inducing points in the x_2 direction are chosen to coincide with the spline knot points. Since we use cubic splines, there are only $M-2$ unique knot points, creating an underdetermined problem. We therefore add an extra pair of inducing points equidistant between the first two and last two knot pairs. We use these samples to evaluate the empirical covariance $\widetilde{\mathbf{R}}(\boldsymbol{x}_\star, \boldsymbol{x}'_\star)$ based on the inducing points $\boldsymbol{x}_\star = [0, x_{2,m}, 0]^\dagger$, $\boldsymbol{x}'_\star = [x'_{1,l}, x'_{2,m'}, x'_{3,n}]^\dagger$. We then minimise the weighted sum-of-squares residual

$$J_R = \sum_{l=-L}^{L} \sum_{m=1}^{M} \sum_{m'=1}^{M} \sum_{n=-N}^{N} \rho_{ln} ||\mathbf{R}(\boldsymbol{x}_g, \boldsymbol{x}'_g) - \widetilde{\mathbf{R}}(\boldsymbol{x}_g, \boldsymbol{x}'_g)||_F^2 \tag{31}$$

over the hyperparameters $\boldsymbol{\Theta}_{\boldsymbol{\kappa}}$ using weights $\rho_{ln} = (L-|l|+1)(N-|n|+1)$. We use the L-BFGS-B algorithm to solve this large optimisation problem with approximately $(3M-P)^2(L+1)(N+1)/2$ degrees of freedom.

To bootstrap this procedure, we need an initial guess for the velocity covariance function. We obtain this from a Gaussian-blob model of the velocity covariance, which is projected onto the constraints and tuned so that the initial guess of the velocity variance $\mathbf{R}(\boldsymbol{x}, \boldsymbol{x})$ matches the least-squares regression estimate of the velocity variance obtained from a spline regression like (25). We provide details of this procedure in Appendix B.

H. Datasets

In this section we describe the generation of synthetic and real experimental PTV data to test our GP models. The synthetic data are based on direct numerical simulations of passive tracers in homogeneous, isotropic turbulence and turbulent channel flow. The experimental data are obtained from time-resolved Shake-The-Box particle tracking in the wake of a square prism.

In all simulations, the position $\boldsymbol{X}(t)$ of passive tracers advects with the flow as $\dot{\boldsymbol{X}} = \boldsymbol{u}(\boldsymbol{X}, t)$. Velocity data sampled on tracer trajectories are used to generate a synthetic two-pulse PTV measurement at times $t_0 \pm \Delta t/2$. Since the desired PTV time separation Δt is small compared to the flow timescales, we locally approximate the trajectories of tracers at time t_0 with a first order Taylor series expansion $\boldsymbol{X}(t; \boldsymbol{X}_0, t_0) = \boldsymbol{X}_0 + \boldsymbol{u}(\boldsymbol{X}_0, t_0)(t - t_0)$. Additive white Gaussian noise is added to these trajectories to model position uncertainty. Finally, a first order polynomial fit is applied to each position pair to measure the (noisy) tracer position and velocity at t_0.

1. Homogeneous isotropic turbulence

We simulate homogeneous, isotropic turbulence at $R_\lambda \approx 122$ in a 512^3 periodic box of side-length 2π using TuRTLE [37]. TuRTLE implements a standard pseudo-spectral method to solve the incompressible Navier-Stokes equations in vorticity form. Statistical stationarity is maintained with band-passed Lundgren forcing in the wavenumber range $[2, 4]$. The trajectories of 2^{22} tracers are integrated from $t = 0$ to a statistically stationary state at $t = 16$ corresponding to a simulated duration of $23.4T_{int}$ integral timescales. The spatial resolution is $k_{max}\eta = 1.87$, where $k_{max} = 256$ is the maximum resolvable wavenumber and η is the Kolmogorov lengthscale. This ensures that small scales are well resolved. Particular care is taken over the time integration of tracers: cubic splines were used to interpolate the underlying velocity field and time-stepping was performed using a fourth-order Adams-Bashforth method.

The snapshot at $t = 16$ is subdivided into 8^3 cubic sub-volumes of side-length $\pi/4 \approx 107\eta \approx 1.4L_{int}$, where L_{int} is the integral lengthscale. To recreate datasets of differing seeding density, we downsample the data to contain on average 512, 2048 or 8192 particles within each sub-volume. We define the seeding concentration ρ_p to be the average number of particles per unit volume and associate a characteristic length-scale $\ell_p = \rho_p^{-1/3}$. The seeding concentrations tested correspond to $\ell_p = 13.4, 8.4$ and 5.3η. The PTV timestep $\Delta t = 0.15\tau_\eta$ is chosen so that the mean square displacement between timesteps is 0.85η, around 0.8% of the measurement domain size. This corresponds to a typical displacement for PTV. For technical reasons with VIC+, trajectories which enter or leave the volume over the tracked interval are excluded, ensuring that tracks always contain two points within the measurement domain. Subsequently, additive noise is added to the trajectories. Three levels of measurement noise are considered, which

correspond to an RMS velocity error of $\sigma = 0, 0.01$ and $0.03u'$ in the PTV measurement, where u' is the RMS velocity fluctuation.

The covariance model (11) is trained on this data to learn 129 coefficients corresponding to wavenumbers $\kappa_m = 0, 2, \ldots 256$. This includes large-scale features whose wavelength exceeds four times the measurement domain size. To accelerate querying the model, cubic splines are used to interpolate the longitudinal and transverse correlation functions (13,15) on a regular grid with spacing 0.167η, rather than evaluating the series exactly. This introduces some small numerical error which can cause $\mathbf{K}_{\boldsymbol{yy}}$ to become indefinite. To mitigate this, we truncate $\mathbf{K}_{\boldsymbol{yy}}$ down to the r largest eigenvalues ordered $\sigma_1 \geq \sigma_2 \geq \ldots$ necessary to capture 99.99% of the signal energy, i.e. $\sum_{i=1}^{r} \sigma_i \geq 0.9999\mathrm{Tr}\left(\mathbf{K}_{\boldsymbol{yy}}\right)$.

To provide a comparison to a naïve GP approach, we also train the incompressible RBF GP model (8) upon this dataset by maximising the marginal likelihood (10) over 25 iterations using the L-BFGS-B algorithm. To reduce the complexity, we assume a zero mean flow and optimise the hyperparameters $\boldsymbol{\theta} = \{\mathbf{L}_{\boldsymbol{w}}, \sigma^2\}$. The model uses $N = 8^3, 13^3$ and 20^3 inducing points evenly spaced on a uniform Cartesian grid corresponding to a $1:1$ ratio between inducing points and input points. The RBF scale is $\ell = 1.5h$, where h is the spacing between adjacent grid points. Training with $N = 20^3$ inducing points takes around $19,000$ core-hours, which is about 10^4 times more expensive than training the isotropic model.

2. Turbulent channel flow

We simulate turbulent channel flow in a $4\pi \times 2 \times 2\pi$ domain, periodic in the streamwise and spanwise directions, at $\mathrm{Re}_\tau \approx 180$ using spectralDNS [38]. The channel half height is $h = 1$ and the skin friction velocity is $u_\tau = 0.064$. spectralDNS implements a standard spectral-Galerkin method discretised onto $192 \times 128 \times 192$ intervals using Fourier basis functions for the streamwise and spanwise directions and Chebyshev polynomials in the wall-normal direction. Second-order accurate integration in time is performed using the Crank-Nicolson method for the linear terms and an Adams-Bashforth method for the non-linear terms with a constant time step 5×10^{-3} corresponding to a Courant-Friedrichs-Lewy number of 0.076. The simulation is initialised from laminar conditions with a small perturbation added to trigger transition to turbulence and is initially allowed to evolve to a fully turbulent state with a dynamically adjusted driving force to maintain the bulk velocity $U_b = 1$. After $tU_b/h = 60$ convective flow through times, the forcing is turned off and the flow is driven by a uniform pressure gradient. After a further 210 convective flow through times to establish stationarity, 2^{22} tracer particles are uniformly seeded throughout the domain and their trajectories are integrated using a second-order Heun scheme. The tracer velocity is interpolated using a fourth order barycentric Lagrange interpolation scheme.

At 21 time steps evenly spaced over $tU_b/h \in [250, 300]$, the tracers are subdivided into 72 equal measurement volumes of dimension $\pi/6 \times 1 \times 2\pi/3$ covering the bottom half of the channel $x_2 \in [-1, 0]$. To recreate datasets of differing seeding density, we downsample the data to contain on average 910, 3640 or 14600 particles. This corresponds to seeding concentrations $\rho_p = \ell_p^{-3}$ with $\ell_p = 19, 12$ and $7.7\delta_\nu$, where δ_ν is the skin friction length scale. As for the HIT case, trajectories are selected so that all tracks are entirely within the measurement domain. Finally, a normally distributed position error is added to trajectories, corresponding to a velocity measurement noise of $\sigma = 0.01U_b$. This level is representative of typical PTV errors [39].

The velocity is reconstructed on a uniform grid with $33 \times 65 \times 17$ elements with corresponding grid spacing $11.8\delta_\nu$, $2.8\delta_\nu$ and $5.9\delta_\nu$ in the streamwise, wall-normal and spanwise directions. Correspondingly, the covariance model is discretised with $M = 67$ spline basis functions, 65 Fourier modes in the streamwise direction up to a maximum wavenumber $46.5/h$ and 33 modes in the spanwise direction up to $90.4/h$. This corresponds to a domain $[-\pi/6, \pi/6]$, $[0, 1]$ and $[-2\pi/3, 2\pi/3]$ in each principal direction and matches the resolution of the underlying simulation in the streamwise and spanwise directions. It has approximately 1.9×10^7 hyperparameters to learn. These are trained on up to 1512 snapshots of the tracer field.

Evaluation of equation (23) is accelerated using the nonuniform fast Fourier transform library from the Flatiron Institute [40]. After solving for the weights $\boldsymbol{w}_{\boldsymbol{\kappa}}$, the model can be queried in a few seconds on a single core. The bottleneck lies in evaluating $\mathbf{K}_{\boldsymbol{yy}}$ and solving for $\mathbf{K}_{\boldsymbol{yy}}^{-1}\boldsymbol{y}$. To accelerate this, cubic splines are used to approximate the covariance model (21) on this grid in the streamwise and spanwise directions, rather than evaluating the Fourier series exactly. As in the HIT case, this introduces a small numerical error in $\mathbf{K}_{\boldsymbol{yy}}$, which we remedy with the same method.

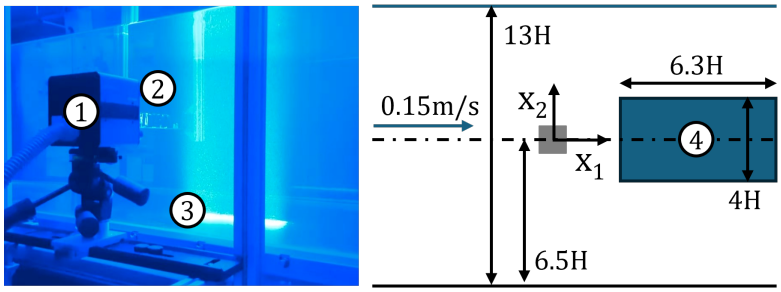

FIG. 2: Photograph (left) and sketch (right) the experimental setup, showing (1) Minishaker system, (2) square prism, (3) LED illumination and (4) the particle tracking measurement volume within the Recirculating Water Tunnel. The coordinate system origin is at the centre of the prism at the mid-span of the tunnel and is aligned with the streamwise (x_1) and transverse (x_2) directions, with x_3 pointing out of the page.

3. *Square Prism Wake Flow*

We measured the flow in the wake of a $H = 50$mm wide square prism at $\mathrm{Re}_H \approx 7850$ in the University of Southampton's Recirculating Water Tunnel using time-resolved Lagrangian particle tracking. Figure 2 shows the experimental configuration. A square prism was suspended in the 1.2m wide test section, centred at 325mm from the floor. The freestream velocity was maintained at 0.150 ± 0.003 m/s at a water depth of 0.65 m and temperature of 22 ± 0.1 °C. The flow was illuminated using a Lavision LED-Flashlight 300 lamp using 335 μs long pulses at 121.2 Hz. Two thousand time-resolved burst recordings of the flow were obtained at 3.5 s intervals using a four-camera LaVision Minishaker 2M PTV system. For comparison, we estimate the vortex shedding period is ~ 2.4 s based on a Strouhal number of 0.14 [41]. In each burst, 25 images were recorded with a resolution of 1984×1264 px.The uncropped field of view covers a 230 mm tall and 83 mm deep region between 105 and 445 mm downstream of the prism centreline, corresponding to a spatial resolution of 0.165 mm/vx.

The cameras were calibrated based on a third-order polynomial fit to a single view of a two-step LaVision calibration plate (model 309-15-3) at the mid-plane of the measurement volume. Subsequently, we applied a volume self-calibration to refine the calibration, resulting in an average disparity error of 0.03 voxels. Particle tracking was performed in DaViS 10 using multi-pass Shake-The-Box processing: two passes forward and backward in time, followed by a reconnect pass. We used four outer loop iterations to add particles and four inner loop iterations to refine particle positions with a shaking amplitude of 0.1 voxels. Particles closer than 1.0 voxel were rejected. The average number of tracked particles is $\sim 14,000$, corresponding to an image concentration of 0.0056 ppp.

To evaluate the performance of GPR on two-pulse PTV data, we extracted the positions of particles at two consecutive frames in each burst. The average streamwise displacement of particles between these frames is 5.0 vx. The measurement volume is truncated to $316 \times 197 \times 79$ mm to eliminate regions of low particle concentration near the edges. In each burst, particle tracks are randomly partitioned into training and cross-validation data. Training data are used to train and evaluate the model, whereas cross-validation data are used to check the accuracy of the velocity field predicted by the model. To test the effect of seeding concentration upon reconstruction accuracy, we withheld 90, 50 or 10% of the data for cross-validation, leaving $\sim$ 1400, 6700 or 12,500 particles used to reconstruct each snapshot.

Using these data, we train a model of the form described in §II E using $M = 38$ spline basis functions with uniformly spaced knots in the transverse direction, and $2L+1 = 65$ and $2N+1 = 17$ Fourier modes up to a maximum wavenumber of 0.303 mm^{-1} and 0.260 mm^{-1} in the streamwise and spanwise directions, respectively. For the most sparsely sampled dataset (10% training, 90% cross-validation), we use $M = 27, 2L+1 = 33$ and $2N+1 = 9$ modes up to a maximum wavenumber of 0.154 mm^{-1} and 0.141 mm^{-1} in the streamwise and spanwise directions, commensurate with the lower seeding concentration. The model is trained with 120 snapshots of the particle velocities over six epochs. In contrast to the channel flow, where the mean flow is fit using B-spline regression, we instead update the mean flow field after each epoch based on the sample mean of the inferred weights $\boldsymbol{w}_t$. We found that restricting the inducing points to only the $L+1$ and $N+1$ grid points within the measurement domain led to poor convergence during training. Using the full $2L+1$ and $2N+1$ inducing points, corresponding to the whole grid, resulted in better performance. Furthermore, to reflect the anisotropic nature of the measurement error, we fit a different noise variance for each velocity component.

4. VIC+ processing

To provide a comparison to an industry standard method, velocity fields were reconstructed on a regular Cartesian grid using VIC+ [7, 8] in DaViS 10. Due to a limitation of DaViS, the grid resolution must be the same in each direction. Therefore, in the TCF case, the velocity field is reconstructed on a uniform Cartesian grid with resolution $2.9\delta_\nu$ ($62\times129\times33$ vectors), which is downsampled to match the GPR reconstruction. The first grid point of the VIC+ reconstruction is at $x_2 = -0.9918h$ ($y^+ = 1.47$). In the HIT case, the resolution matches the GPR reconstruction: $65 \times 65 \times 65$ vectors on a Cartesian grid with spacing 1.68η. Likewise, in the prism wake case, we use a Cartesian grid of $38 \times 61 \times 16$ vectors with 32 voxel (5.27mm) spacing in each direction. This is matched exactly to the grid used for GPR. In all cases, the flow field is solved using 40 iterations with de-noising factor of 0.001.

III. RESULTS

A. Homogeneous Isotropic Turbulence

How well does the training of the model recover the underlying covariance function in HIT? Figure 3a shows the longitudinal (13) and transverse (15) correlations obtained from fitting (16) to the most sparse dataset $\ell_p = 13.4\eta$ with the largest additive measurement noise tested, $\sigma = 0.03u'$. This case is chosen to illustrate the robustness of the training. The ground truth of $R_{LL}(r) = R_{ii}(r\boldsymbol{e}_i)$ and $R_{NN} = R_{jj}(r\boldsymbol{e}_i)$ ($j \neq i$, no summation implied) is uncertain due to anisotropy at the largest scales of the motion; the shaded region reflects the variation in these estimates over the principal directions $\boldsymbol{e}_i$. We observe good agreement between the estimates and the ground truth. The maximum likelihood estimate is also consistent with the histogram estimates of R_{LL} and R_{NN} used to train the model. The most significant deviations occur at the largest scales, where the statistical convergence is the poorest and the isotropy assumption begins to break down. The inset of figure 3a shows that there is good agreement at small scales $r \leq 10\eta$, despite the fact that this training dataset contains fewer than one particle per $(13.4\eta)^3$.

Figure 3b shows the model estimate of the dissipation rate $\epsilon = -15\nu R''_{LL}(0)$ and the maximum likelihood estimate of the noise as a function of the seeding concentration ρ_p and true additive measurement noise σ. Even with very sparse data $\ell_p = 13.4\eta$, the dissipation rate estimate is in error by less than 25% and under 1.1% at higher seeding densities. The estimated measurement noise is also in good agreement. This demonstrates the ability of the product-of-experts ML estimator (16) to recover fundamental statistics of the turbulent flow and correctly infer the measurement noise with sparse data.

The fitting of the HIT regression model is insensitive to the number of hyperparameters used. We tested models with half ($\kappa_m = 0, 4, \ldots 256$) and twice as many ($\kappa_m = 0, 1, \ldots 256$) hyperparameters for the $\sigma = 0.03u'$ noise level. The estimates of the dissipation rate and noise differ by less than 0.3% and 0.6% respectively; there is negligible ($< 0.1\%$) change in the mean-square velocity error. We also tested a model with half the wavenumber resolution ($\kappa_m = 0, 2, \ldots 128, \kappa_{max}\eta = 0.94$). For the highest seeding concentration cases, this results in a small change in the model estimates of dissipation and noise (less than 0.7% and 1.6%, respectively). It improves the accuracy of the dissipation and noise estimates in the lowest seeding concentration case, where the relative error in dissipation reduces from 25% to 2.7% and the relative error in noise reduces from 19% to 14%. However, this results in negligible ($< 0.1\%$) change in the mean-square velocity error. We conclude that a model with $\kappa_{\max}\eta \geq 0.94$ is sufficient to resolve the mean dissipation rate and that a lower resolution is preferable when training data are sparse.

Figure 4a shows an example of the reconstructed velocity field for a seeding concentration $\ell_p = 5.3\eta$ and additive measurement noise corresponding to $\sigma = 0.03u'$. There is a close agreement between the ground truth and reconstruction, even in the presence of noise. The reconstruction uncertainty is shown in Figure 4b. Both the error and uncertainty are least near particle locations, i.e. where the velocity field is sampled, and greatest in the gaps between the input data. We also observe greater measurement uncertainty near the edges of the domain, where fewer data are available nearby to predict the velocity field.

We quantify the accuracy of the GP reconstruction and uncertainty estimate as a function of seeding concentration and measurement noise in Figure 5a. The accuracy of VIC+ reconstruction, applied to the same data, is included to provide a comparison to the industry standard. In all cases, we observe that the mean square measurement error with GPR is lower than VIC+; it is 50% smaller than VIC+ at the highest seeding concentration. Additive measurement noise has little effect upon the reconstruction accuracy for the range of seeding concentrations tested here. Interpolation error dominates: the measurement uncertainty scales approximately as ρ_p^{-1}. The spatial distribution of the measurement error, averaged over the x_2 and x_3 directions, is shown in Figure 5b. We find reasonable agreement between the uncertainty predicted by GPR and the mean square error; for a typical level of additive measurement noise, the model estimates the mean square error to within 22% of the true value at the highest seeding concentration.

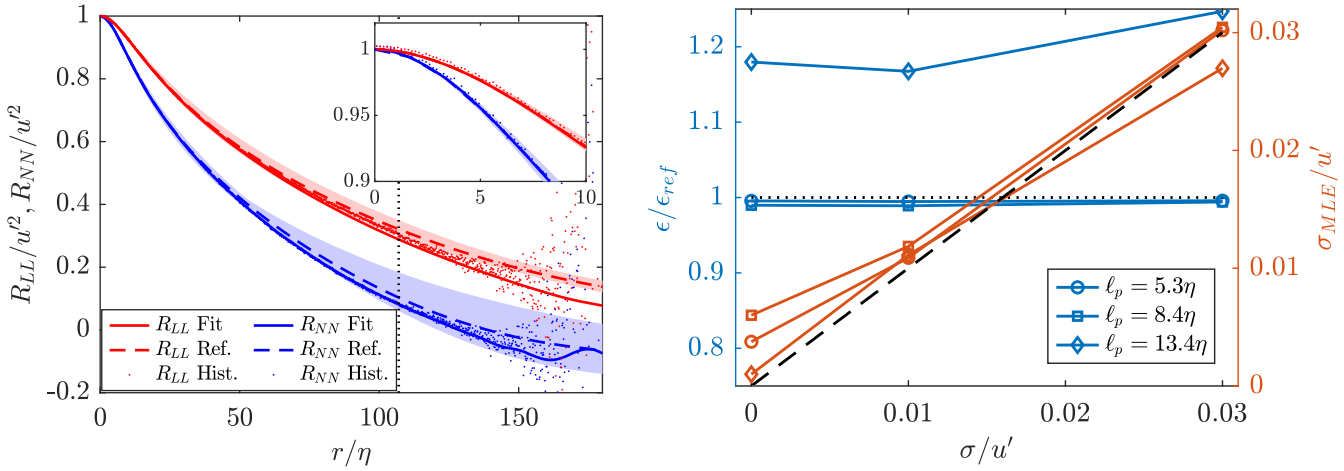

FIG. 3: (a) Comparison of maximum likelihood estimates of longitudinal and transverse autocorrelation function to ground truth in HIT for $\sigma = 0.03u'$ and $\ell_p = 13.4\eta$. The dotted line shows the domain size. The shaded region shows confidence interval for ground truth, based on the variation of R_{LL} and R_{NN} with orientation $\hat{\boldsymbol{r}}$. The inset shows the region $r \leq 10\eta$. (b) Estimated dissipation rate (ϵ, left) and noise (σ_{MLE}, right) as a function of seeding concentration and noise.

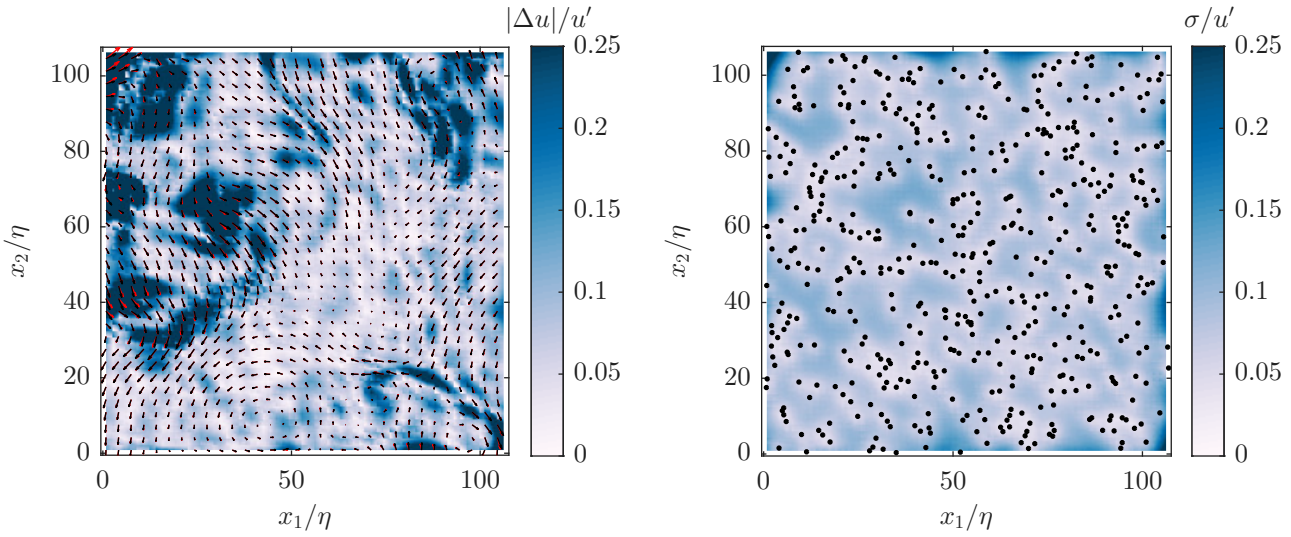

FIG. 4: GP reconstruction of HIT with 8192 particles ($\ell_p = 5.3\eta$) and additive measurement noise $\sigma = 0.03u'$ across the plane $x_3 = L/2$. (a) Error in reconstructed velocity field (black vectors) compared to ground truth (red vectors) and (b) estimated uncertainty. Black markers show particles within $x_3 \pm 3.3\eta$.

The uncertainty near the domain edges is largest, but affects only a small proportion of the measurement domain. As the seeding concentration increases, the extent of the affected region is reduced.

To test the local estimate of uncertainty, we define a standardised z-score for the error $z = \Delta u/\sigma_u$, where Δu is the measurement error in a velocity component and σ_u is the associated uncertainty modelled by (2). Under the Gaussian process Ansatz, z follows the standard normal distribution. One therefore expects that 95.45% of errors lie within $-2 < z < 2$. Figure 6 shows the cumulative distribution function (CDF) of z for a noise level of $\sigma = 0.03u'$. We observe that the CDF is closely approximated by the standard normal distribution. The $\pm 2\sigma_u$ interval captures at 93% of error values for $\ell_p = 5.3\eta$, rising to 94.4% for $\ell_p = 13.4\eta$. The model therefore generates a good prediction of the true 95% confidence interval.

Let us consider the spectral decomposition of error in the reconstructed velocity field. Figure 7 presents the spectral coherence of the reconstructed velocity signal $|E_{\hat{u}u}(\kappa)|^2/E_{\hat{u}\hat{u}}(\kappa)E_{uu}(\kappa)$, where $E_{\hat{u}u}$, E_{uu} and $E_{\hat{u}\hat{u}}$ are the cross-spectral density and spectral density of the ground truth $u(x_1, \ldots)$ and reconstruction $\hat{u}(x_1, \ldots)$, respectively. This can be thought of as the correlation coefficient between reconstructed and ground truth Fourier modes of the velocity signal.

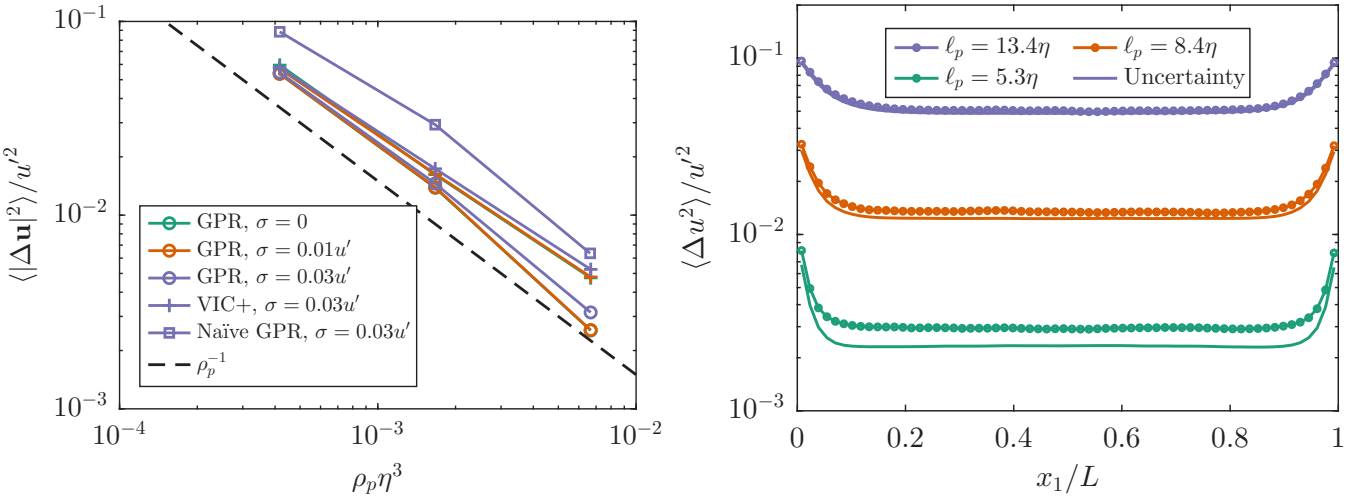

FIG. 5: (a) Volume average mean square error of GPR compared to compared to VIC+ and (b) the spatial profile of the mean square error in comparison to the uncertainty predicted by GPR with $\sigma = 0.03u'$ additive measurement noise.

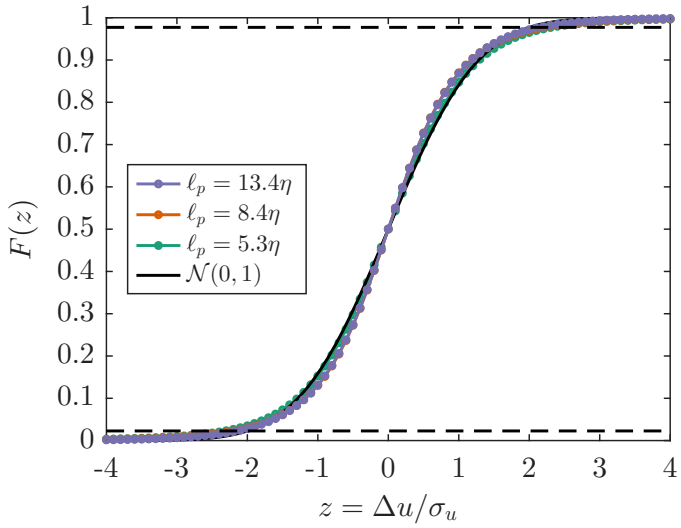

FIG. 6: Cumulative distribution of the standardised error $z = \Delta u/\sigma_u$ for GPR of HIT with $\sigma = 0.03u'$ at different seeding concentrations, with the standard normal distribution shown in black. Dashed lines show the 95.45% confidence interval corresponding to $-2 \leq z \leq 2$ for the normal distribution.

There is a sharp cutoff in the spectral coherence near a wavenumber π/ℓ_p, which corresponds to the Nyquist frequency of a signal sampled the at the characteristic spacing of particles. There is a spurious peak at large wavenumber, which is due to spectral leakage of the 64-point Hann window used. Gaussian Process Regression and VIC+ show a similar frequency response, with VIC+ showing slightly better coherence at small scales but worse coherence at largest scales. We conclude that the cutoff wavenumber for both VIC+ and GPR in this case is $\kappa_c \approx \pi/\ell_p$.

B. Turbulent Channel Flow

The first step in training the TCF model is to estimate the mean velocity field. Figure 8a compares the ground truth mean velocity profile to the velocity profile obtained from the least-squares fit of (25) to the noisy synthetic PTV data. A total of $T = 360$ snapshots were used. We observe that mean flow profile obtained from least-squares regression is in close agreement with the ground truth, even at the lowest seeding concentration levels. The typical error in the mean flow is very small, below 0.1% of the bulk velocity. In contrast, the mean velocity profile obtained from VIC+

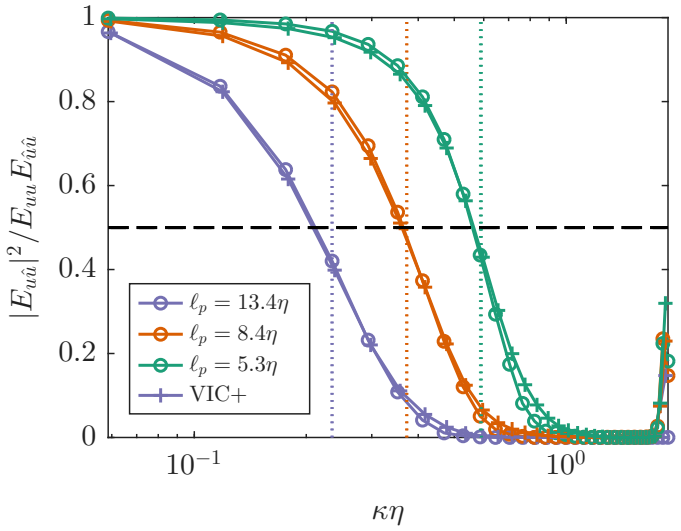

FIG. 7: Spectral coherence of the reconstructed velocity field $\hat{u}(x_1, \ldots)$ and ground truth $u(x_1, \ldots)$ in HIT, for varying seeding concentrations at a noise level $\sigma = 0.03u'$. The cutoff wavenumber $\kappa_c = \pi/\ell_p$ is shown with a dotted line for each seeding concentration.

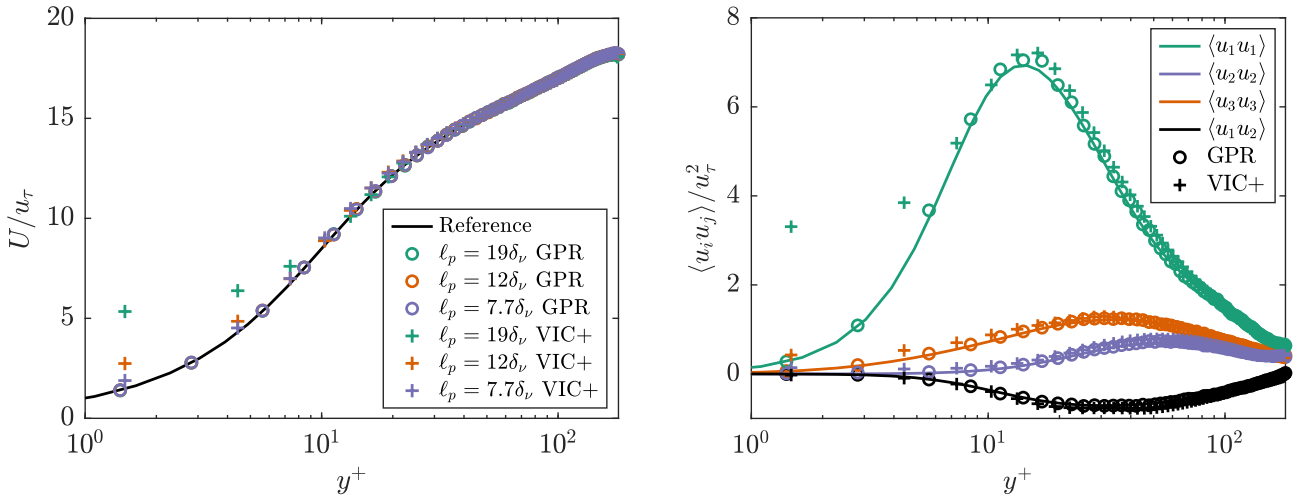

FIG. 8: (a) Mean velocity profile and (b) Reynolds stress profile in turbulent channel flow obtained from least-squares fitting (25) and VIC+ for a seeding concentration $\ell_p = 12\delta_\nu$, in comparison to ground truth. The abscissa shows the distance from the wall $y^+ = (x_2 + h)/\delta_\nu$.

interpolation exhibits large discrepancies near the wall. However, we caution that the strong agreement seen with GPR here is likely difficult to replicate in practice, because our simulation conditions do not capture the increased measurement errors and reduced seeding concentration often found in near-wall PTV [2]. Figure 8b compares the ground truth Reynolds stress profile to the Reynolds stress profile obtained from a least-squares fit in the same form as (25), which is used to bootstrap the model using the procedure in Appendix B. The least squares fit closely captures the near wall turbulence, which exhibits a prominent peak near $y^+ = (x_2 + h)/\delta_\nu = 15$. In contrast, VIC+ substantially overestimates velocity fluctuations near the wall.

We bootstrap the GP model using the mean flow field and velocity variance profiles shown in figure 8 and iteratively train the model following §II E. Each pass of the training dataset through this procedure constitutes one epoch. Figure 9 shows the convergence of the mean square velocity residual as the number of training epochs increases. Here, the mean square residual is obtained by comparison to known ground truth. In practice, convergence can be identified by cross-validation against a withheld portion of the dataset [15]. The model converges quickly: there are diminishing returns for training beyond three epochs. Figure 9 also shows the effect of training with more or fewer training data. Decreasing the number of training snapshots from $T = 360$ to 72 reduces the overall accuracy and results in a

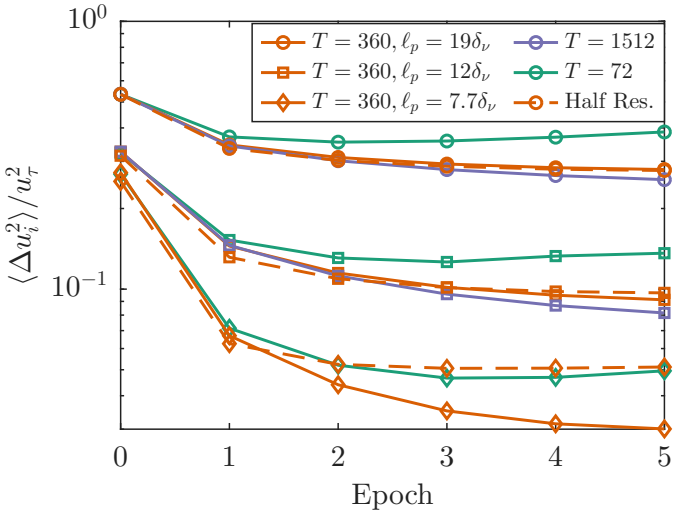

FIG. 9: Convergence of mean square residual of reconstructed velocity fields with increasing training epochs. Markers correspond to different seeding concentrations whereas colours represent the number of snapshots used for training. Dashed lines show the convergence of a half resolution model.

tendency to over-fit the data to noise, evidenced by a decrease in the model accuracy with further training epochs. However, increasing the number of training data from $T = 360$ to 1512 snapshots results in a marginal increase in reconstruction accuracy, despite a fourfold increase in computational cost. Therefore, in the following, we consider models trained on 360 snapshots for five epochs.

To test the effect of model resolution, we also train a half-resolution model with 35 spline basis functions and half as many Fourier modes (33 streamwise modes up to $\kappa_1 h = 23.3$ and 17 spanwise modes up to $\kappa_3 h = 45.2$). This reduces the number of hyperparameters sixteen-fold and the resulting model is around 40-60% cheaper to train and query. The dashed lines in Figure 9 show the mean-square velocity error of the half resolution model trained with 360 snapshots. There is a modest difference in accuracy for $\ell_p = 19\delta_\nu$ and $\ell_p = 12\delta_\nu$, but for $\ell_p = 7.7\delta_\nu$ the mean-square velocity error is a factor of two larger. This arises from the half resolution model's inability to capture high-frequency information from dense training data, which we address presently.

Does training the model recover the underlying correlation function? Figure 10 illustrates the correlation function of streamwise velocity fluctuations $R_{11}(\boldsymbol{x}, \boldsymbol{x}')$ trained on 360 snapshots from the most sparse dataset $\ell_p = 19\delta_\nu$. Despite the sparsity of the training data, the spatial coherence of the velocity fluctuations is captured well in comparison to the ground truth. As a more quantitative test, Figure 11 compares the streamwise velocity spectrum $E_{11}(\kappa_1)$ to the ground truth obtained from DNS at three representative wall-normal positions. These are obtained from the streamwise velocity autocorrelation model $R_{11}(\boldsymbol{x}, \boldsymbol{x}')$. Figure 11a shows that the spectral content of the near-wall turbulence is adequately captured by the trained model at the lowest wavenumbers. However, the comparison is less favourable further from the wall where the turbulence intensity is lower and fine-scale details are more difficult to discern from the measurement noise. Figure 11b shows the relative magnitude of the trained model and ground truth energy spectrum at different wall-normal distances, with wavenumber scaled by the seeding density scale as $\kappa\ell_p$. The model and ground truth are in good agreement below a cutoff wavenumber of around $k_c\ell_p \approx 1 - 2$, beyond which there is insufficient detail in the training data to discern flow features from noise.

Figure 12 compares typical reconstructions of the flow field made at a low seeding concentration $\ell_p = 19\delta_\nu$ and additive measurement noise $\sigma = 0.01U_b$. The GPR-based reconstruction retains the spatial coherence of the hairpin-like vortical structures near the wall. These vortex signatures are continuous and exhibit the characteristic elongated shapes and tilted heads that are expected in wall-bounded turbulence. This reflects the ability of GPR incorporate prior assumptions about smoothness and correlation scales in a statistically consistent manner. In contrast, the VIC+ reconstruction fails to capture these features as clearly: the isosurfaces appear fragmented and less coherent in space. Consequently, the GPR results display more physically realistic near-wall flow patterns, whereas VIC+ tends to produce spurious features that obscure the true flow dynamics.

How accurately can GPR reconstruct velocity fields from PTV data in turbulent channel flow? Figure 13a shows the mean square error in the velocity field obtained from GPR and VIC+ for different seeding concentrations. The reconstruction error has a strong spatial variation and is largest near the peak in turbulence intensity at $x_2 = -0.91h$. In all cases, GPR results in a significantly smaller prediction error: the volume averaged mean square error is smaller by

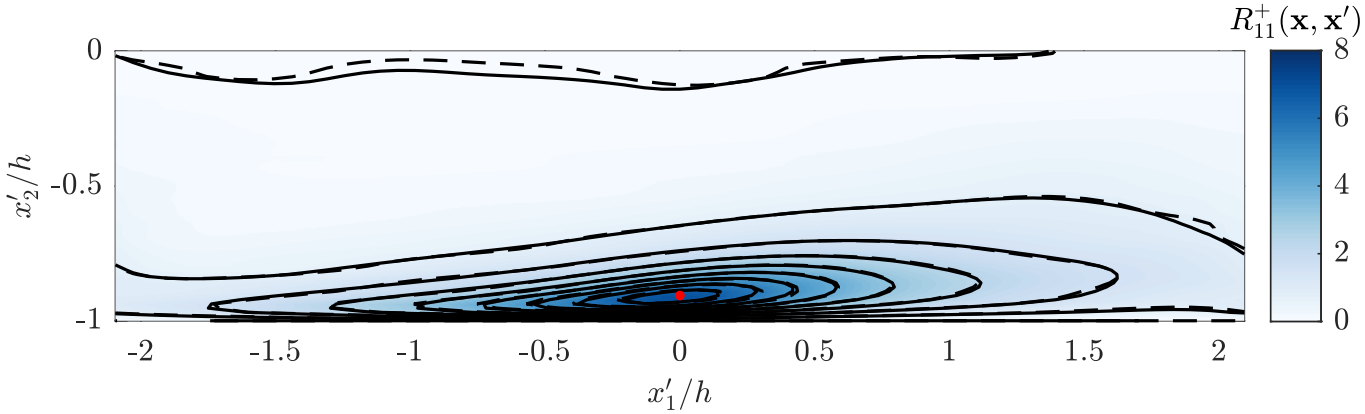

FIG. 10: (a) The model of the streamwise velocity covariance (dashed contours and heatmap) in the $x_1' - x_2'$ plane for $x_2 = -0.91h, x_1 = x_3 = 0$ (red dot), trained on the most sparse data $\ell_p = 19\delta_\nu$. Solid contours show the ground truth.

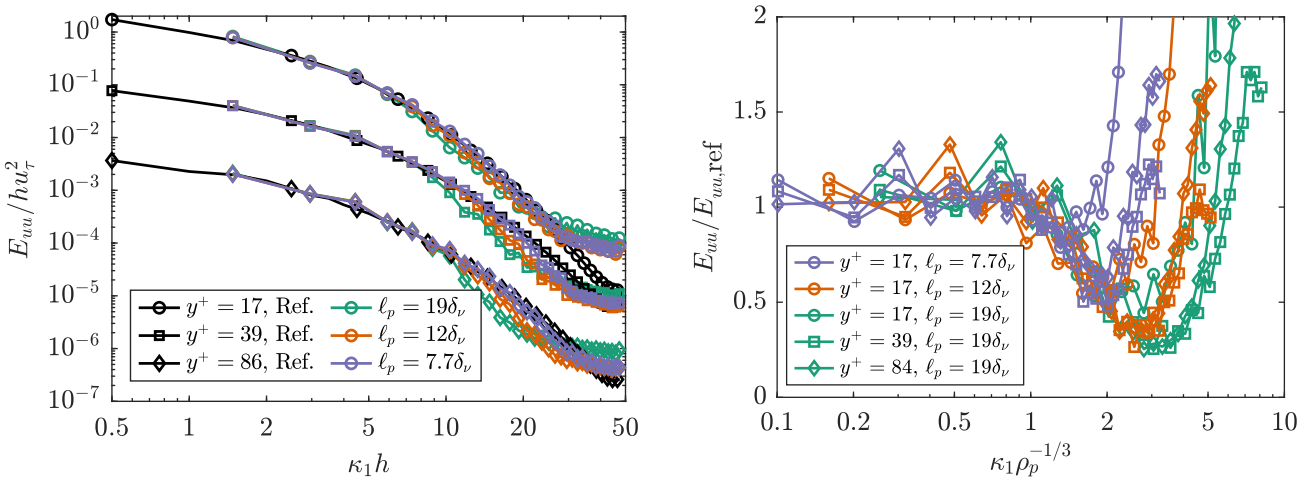

FIG. 11: (a) Energy spectrum $E_{11}(\kappa_1)$ of streamwise velocity fluctuations at different seeding concentrations (colours) and wall-normal distances y^+ ($x_2/h = -0.91, -0.78, -0.53$). For clarity, the spectra at $y^+ = 39$ and 86 have been offset by a factor of 10^{-1} and 10^{-2}. (b) Model spectrum, relative to ground truth.

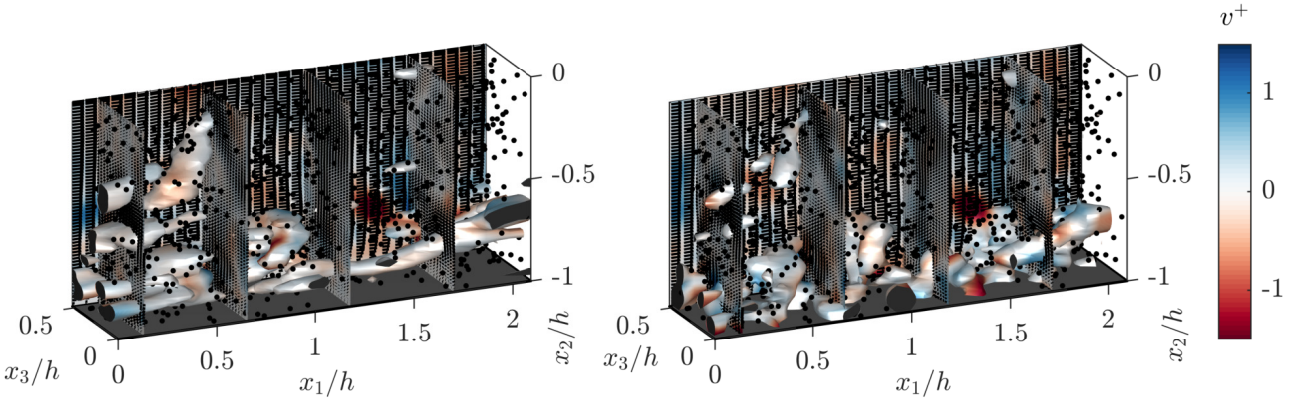

FIG. 12: GPR (left) and VIC+ (right) reconstruction of flow field with from 903 tracer particles in channel flow with $\sigma = 0.01U_b$. Isosurfaces show contours of Q-criterion $Q^+ = 0.0075$ coloured by wall-normal velocity to identify near-wall vortex structures.

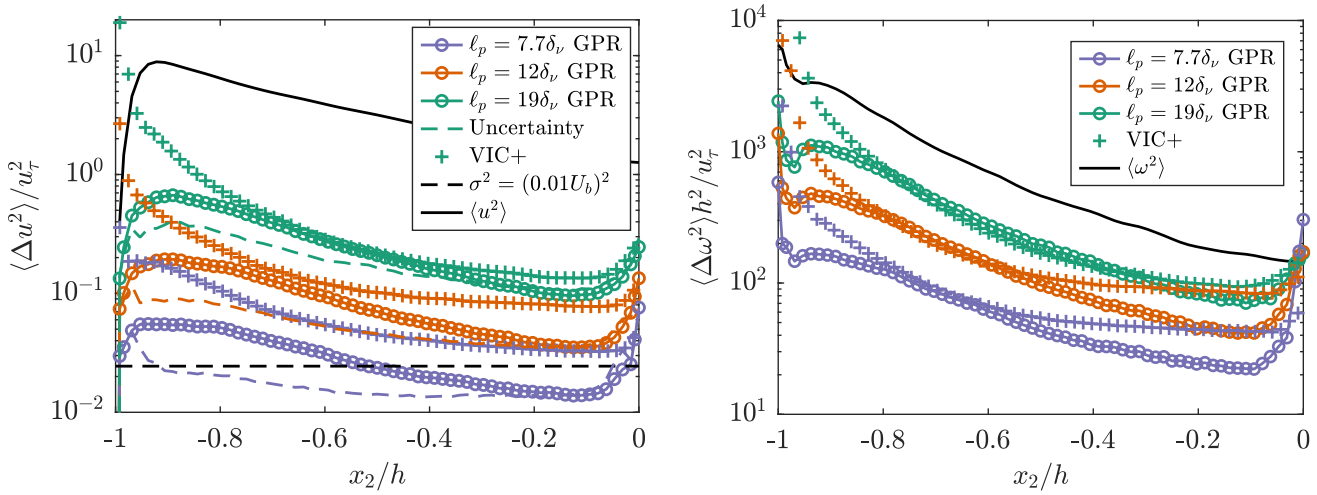

FIG. 13: Profiles of the mean square (a) velocity error Δu and (b) vorticity error $\Delta\omega$ from GPR and VIC+ reconstructions of the flow field. The black lines show the mean turbulent kinetic energy $\langle u^2 \rangle$ and mean enstrophy $\langle \omega^2 \rangle$ of the ground truth to contextualise the reported error.

a factor of between 2.1 and 2.7 compared to VIC+. As in the isotropic case, there is a sharp increase in reconstruction uncertainty near the edge of the domain at $x_2 = 0$ due to the lack of data beyond the edge. The reconstruction with VIC+ also exhibits very large error near the wall. A benefit of Gaussian Process Regression is evident in this region, where awareness of the boundary condition and local statistics reduce the residual by over an order of magnitude in comparison. A further benefit of GPR is apparent towards the centre of the channel at the highest seeding concentration tested. Since GPR is a predictor, not an interpolator, the prediction accuracy can surpass the noise floor.

How well is measurement uncertainty estimated by the model? The dashed lines in Figure 13a show the spatially averaged prediction of the measurement uncertainty. The model tends to be overconfident: the prediction uncertainty near the wall is underestimated by up to 60%, relative to the observed mean square error. However, the uncertainty near the centreline is captured correctly and the trend with seeding concentration is correct. To understand why this is the case, consider that the uncertainty lies within the smallest-scale (highest wavenumber) flow features of the underlying velocity field. However, figure 11b demonstrates that the model fails to accurately capture the covariance at wavenumbers beyond $\kappa_c \approx \ell_p^{-1}$. As a result, the true uncertainty is underestimated. This discrepancy is particularly pronounced near the wall, where turbulent flow features are more energetic and smaller in scale.

In practice, we are often interested in the reconstruction of small-scale quantities such as velocity gradients. Figure 13b shows profiles of the mean square residual in the vorticity field evaluated using GPR and VIC+. Again, the advantage of GPR is most pronounced near the wall, where the turbulence is strongest, and near the centre, where measurement noise begins to limit the VIC+ reconstruction. Averaged over the measurement volume, the mean square residual in the vorticity is between 1.7 and 2.6 times smaller with GPR in comparison to VIC+.

To quantify the spatial resolution of the reconstructed velocity field, we evaluate the spectral coherence of reconstructed velocity fluctuations with respect to the ground truth in the streamwise direction. Figure 14a shows this comparison at the near-wall turbulence peak, where the mean-square error in the reconstruction is the largest. In all cases, we find Gaussian Process Regression offers an improvement in the spectral coherence of velocity fluctuations reconstructed near the wall in comparison to VIC+. The spectral coherence improves with increasing seeding concentration. Figure 14b shows that the coherence approximately collapses when the wavenumber is scaled with the seeding concentration as $\kappa_1 \ell_p$. However, the coherence varies with wall-normal distance and a single cutoff-wavenumber does not describe the frequency response well. To illustrate this variation, we calculate an average spectral coherence over the measurement volume. Based on the variation seen between the near-wall and volume average spectral coherence, we conclude that a cutoff-wavenumber between $1 < \kappa_1 \ell_p < 2$ adequately describes the frequency response.

The dashed lines in Figure 14 show the spectral coherence of velocity fields reconstructed using the half resolution model. The half resolution model exhibits almost identical spectral coherence below its maximum streamwise wavenumber $\kappa_1 h = 23.3$, but drops off sharply thereafter. In the lowest seeding concentration cases, Figures 14 and 11 demonstrate that high-frequency information is not effectively learned, so this cutoff poses no significant detriment. However, in the highest seeding concentration case, this simpler model neglects high-frequency information which can be learned from the data. This highlights the importance of adjusting the model resolution to avoid over-fitting sparse

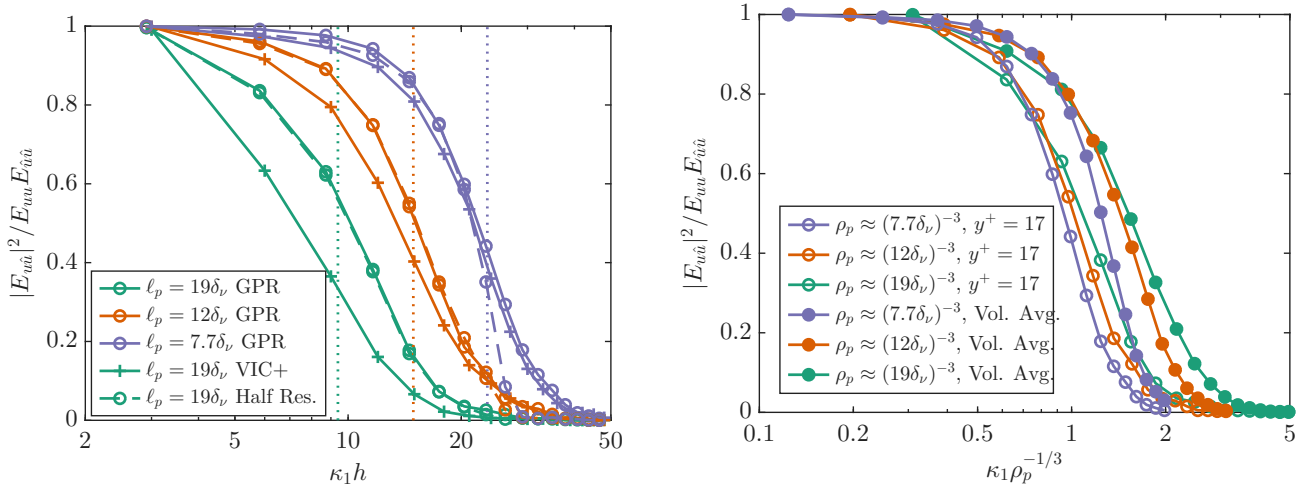

FIG. 14: Spectral coherence of reconstructed velocity fluctuations in turbulent channel flow with respect to the ground truth (a) evaluated at the near-wall turbulence peak $x_2 = -0.93h$ with GPR and VIC+ and (b) when scaled by the average seeding concentration $\kappa_1 \rho_p^{-1/3}$. The dotted lines in (a) mark the wavenumber $\rho_p^{1/3}$.

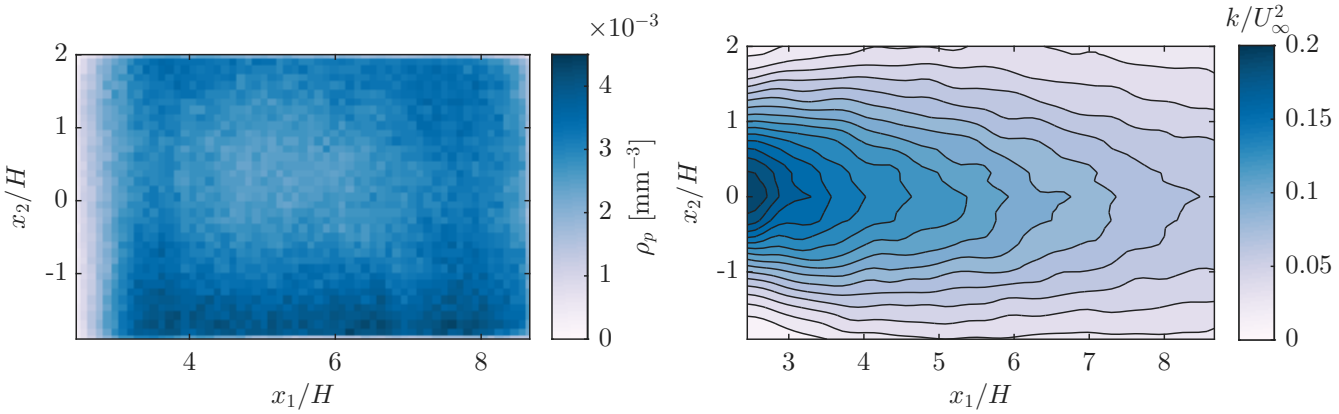

FIG. 15: Spatial inhomogeneity of (a) seeding concentration and (b) turbulent kinetic energy in the turbulent wake of a square prism.

data by matching the model resolution to the cutoff wavenumber.

C. Turbulent Square Prism Wake

We test the performance of GPR under real-world conditions by applying it to the reconstruction of flow fields in the turbulent wake of a square prism, as described in §II H. This dataset is particularly noisy and exhibits strong inhomogeneities in seeding concentration throughout the measurement volume, as illustrated in figure 15a. Furthermore, as the distribution of turbulent kinetic energy shown in figure 15b demonstrates, this flow does not exhibit the high degree of statistical symmetry found in our synthetic tests. This flow therefore represents a challenging test case for GPR, since the model we have used assumes statistical homogeneity in the streamwise and spanwise directions.

Figure 16a shows the mean-square cross-validation error obtained using GPR and VIC+ for varying concentrations of the test data. At the lowest seeding concentration, the mean-square cross-validation error with GPR is around 25% lower than VIC+, but is just 2% smaller at the highest seeding concentration. The cross-validation error is close to model's estimate of measurement noise for each velocity component, which corresponds to an RMS error of 11.4, 11.1 and 19.5 mm/s (0.57, 0.56 and 0.98vx) in the streamwise, transverse and spanwise velocity components. Furthermore,

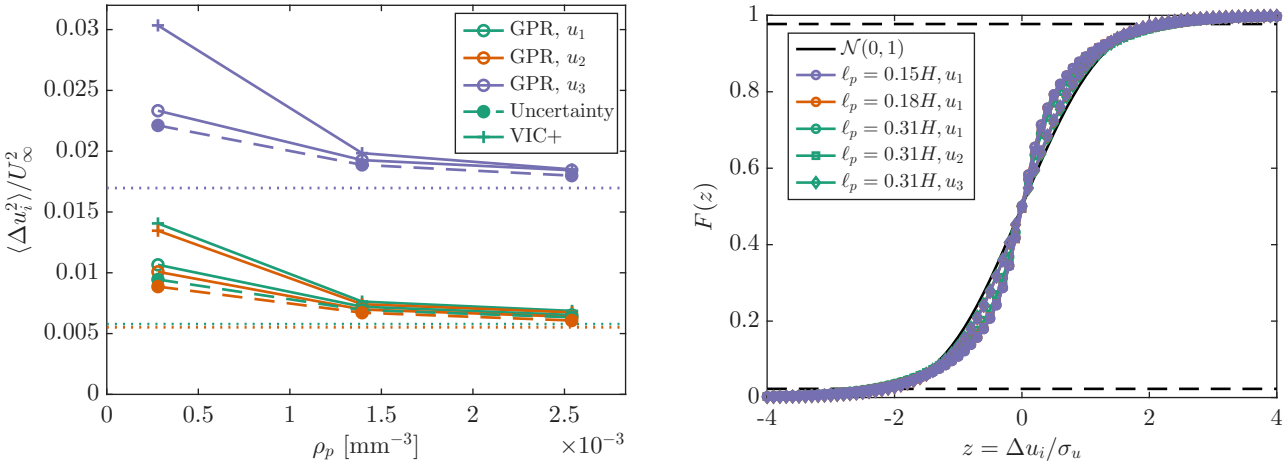

FIG. 16: Cross-validation of velocity field in the wake of a square prism. (a) Mean-square error, obtained using 10, 50 and 90% of the available data. The dotted line shows the noise estimate for each velocity component from GPR. (b) CDF of the standardised cross-validation error for each velocity component at each seeding concentration, obtained from GPR.

for both VIC+ and GPR, we observe that the cross-validation error does not have a strong dependence upon the seeding concentration. In this case, recovery of fine-scale flow details by increasing the seeding concentration does not significantly improve the cross-validation error, because this is dominated by error in the cross-validation data. Filled markers in figure 16a show the mean square uncertainty of the model's prediction. The average uncertainty in the cross-validation data is predicted well by the model. Figure 16b shows the CDF of the standardised cross-validation error. The distribution is more heavy-tailed than the standard normal distribution. Nonetheless, the $2\sigma_u$ confidence interval predicted by the model covers between 92.9% and 94.6% of the cross-validation data.

To quantify the reconstruction of fine-scale flow features, figure 17 shows the mean enstrophy $\langle \boldsymbol{\omega}^2 \rangle$ averaged over the measurement volume. With VIC+, the enstrophy decreases as the seeding concentration increases, indicating that significant noise in the vorticity field is present at low seeding concentration. With GPR, in contrast, the enstrophy increases with increasing seeding concentration, which is expected as more data improves spatial resolution. To estimate the contribution of the vorticity error to these statistics, we perform independent velocity field reconstructions at two times, 16.5ms apart, using four consecutive frames. We do not expect the vorticity to change significantly over this interval, since it is just 0.7% of the vortex shedding period. Note that temporal coherence of the output is not enforced by the reconstruction method, since the velocity fields are obtained from separate inputs. The dashed line in figure 17 shows the mean-square of the change $\delta\boldsymbol{\omega}$ in the vorticity field over this interval. Figure 17 demonstrates that, both in absolute and relative terms, the GP reconstruction of the vorticity field is substantially more coherent in time. The discrepancy between vorticity fields is $2.1 - 5.3$ times smaller in absolute terms, or $1.7 - 3.3$ times smaller relative to the mean enstrophy. This is also demonstrated in the supplementary videos, which show the temporal evolution of the flow field and fine-scale vortices identified using the Q-criterion.

IV. CONCLUSION

Physics-informed Gaussian process regression offers a powerful and transparent method for PTV data assimilation. It incorporates statistical and physical information to predict dense velocity fields from scattered data, accompanied by uncertainty estimates and models of the two-point velocity covariance, enabling further analysis such as POD. Furthermore, given the input data, the statistical distribution of the underlying velocity field is predicted. This can be sampled using the Matheron update rule to train the model, or used for uncertainty propagation. We have introduced methods to create and train physics-informed Gaussian process models of turbulent flows. These methods explicitly incorporate mass continuity, boundary conditions and statistical symmetries such as isotropy and homogeneity into the covariance model, permitting a reduction in complexity compared to more general models (e.g. RBF networks) by exploiting known flow physics. These were tested on synthetic and experimental PTV data with varying levels of seeding concentration and noise for three canonical turbulent flows: homogeneous isotropic turbulence, turbulent channel flow and a square prism wake.

In the case of HIT, we have demonstrated that an isotropic, two-point correlation function can be learned by a

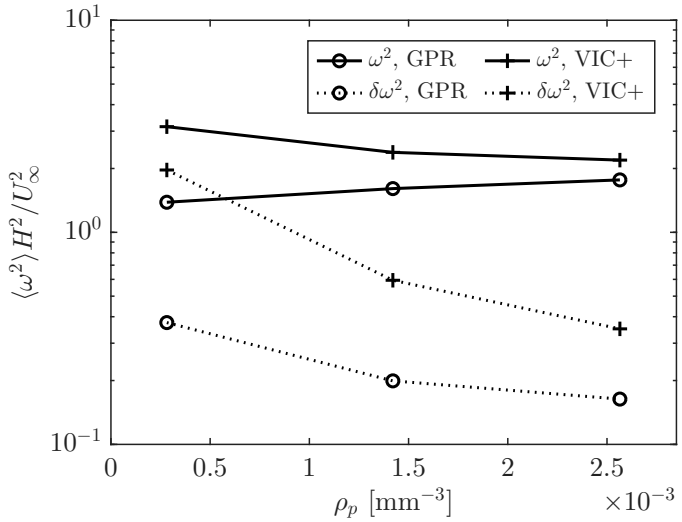

FIG. 17: Volume-average mean-square vorticity magnitude $\langle \boldsymbol{\omega}^2 \rangle$ obtained for different effective seeding concentrations using GPR and VIC+. Dashed lines show the mean-square difference in the vorticity field between reconstructions obtained 16.5ms apart.

maximum-likelihood product-of-experts method with a trivial cost that scales quadratically with the number of data. Even with very noisy and sparse data, the two-point correlation function can be recovered and provides accurate estimates of fine-scale information such as the kinetic energy dissipation rate. Tests with synthetic PTV data showed that the model outperforms the industry standard by up to a factor of two in terms of mean-square error in the velocity field prediction. The uncertainty estimates of these predictions were also shown to be in good agreement with the observed mean-square-error. Measurements of the spectral coherence of the reconstructed flow field against ground truth show that the flow features can be resolved up to the average Nyquist rate $\pi\rho_p^{1/3}$.

In the case of TCF, we have introduced a mixed Fourier/spline feature based GP model that captures spatial inhomogeneity in the wall normal direction, homogeneity in streamwise and spanwise directions, incompressibility and the no-slip boundary condition. Even with these symmetries, the model has millions of hyperparameters, necessitating the use of large datasets to train it. To address this, we have introduced a novel, scalable algorithm to train on very large datasets using a modified expectation-maximisation approach based on sampling conditional realisations of the underlying turbulent process using the Matheron update rule. The complexity of training is the same as inference. This algorithm recovers well the structure of the two-point covariance of the velocity field, even with few, sparse and noisy training data. This model shows order-of-magnitude improvement over VIC+ in reconstructing the near-wall velocity fluctuations in terms of the mean square velocity and vorticity error, and by a factor of $2.1 - 2.7$ on average. It also provides reasonable (although overconfident) estimates of the prediction uncertainty. The spatial resolution of the inferred two-point velocity correlation and flow fields is limited by the seeding concentration ρ_p and the maximum resolvable wavenumber scales as $\rho_p^{1/3}$.

As a real-world test of the method, we applied the GP model developed for turbulent channel flow to LPT measurements of the turbulent wake behind a square prism. This flow does not possess the streamwise statistical homogeneity found in turbulent channel flow. For this particularly noisy dataset, cross-validation confirms that the model closely predicts the uncertainty of its outputs. A modest improvement in the velocity cross-validation error is achieved with GPR in comparison to VIC+. However, GPR demonstrates significant improvement over VIC+ in the temporal coherence of the vorticity fields obtained from two-pulse PTV data.

There are several directions for future research. Firstly, it remains to be quantified how well these models would perform in more complex flows that do not possess the statistical symmetries found here. We have considered inference from a single time instant: as we have shown, these models can be easily extended to multiple timesteps, which would allow the temporal coherence of LPT data to be leveraged. Acceleration information from particle tracks could be incorporated by adding a penalty on the model's prediction of the material derivative to (5). Whilst this becomes a non-linear minimisation problem which precludes using (1) or (2), it allows momentum conservation to be included in the physics and retains the Bayesian framework. Resolvent analysis could be used to prescribe models a-priori, or with fewer parameters: [42] provide models of the covariance structure of turbulent channel flow which require only the mean flow profile to construct. We plan to explore these ideas in future work.

ACKNOWLEDGMENTS

We wish to acknowledge Dr. Daniel Carlson for his assistance in the collection of the experimental dataset. We also acknowledge the use of the IRIDIS High Performance Computing Facility, the Boldrewood Campus Recirculating Water Tunnel, and associated support services at the University of Southampton, in the completion of this work. Pertinent data for this manuscript are available at doi:10.5258/SOTON/D3642 [43].

Appendix A: Modified EM algorithm for training stationary GP models

Input: Noisy observations $\{(\boldsymbol{X}_t, \boldsymbol{y}_t)\}_{t=1}^{T}$, inducing points $\boldsymbol{X}_\star = \{n\Delta x\}_{n=0}^{N}$, Fourier frequencies $\nu_n = \pi n/L_3$, initial hyperparameters $\boldsymbol{\theta}_0$ and σ_0^2, number of iterations I, number of samples per observation S
Output: Learned hyperparameters $\boldsymbol{\theta}$, noise variance σ^2
Initialize $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta}_0$ and $\sigma^2 \leftarrow \sigma_0^2$;
for $i = 1$ **to** I **do**
  Initialize $\Sigma_{\sigma^2} \leftarrow 0, N_{\sigma^2} \leftarrow 0$ and $\widetilde{R}_j \leftarrow 0$ for all $j = -N, \ldots, N$;
  for $t = 1$ **to** T **do**
    for $s = 1$ **to** S **do**
      Sample the prior distribution $\boldsymbol{w}_\star \sim \mathcal{N}(0, \mathbf{K}_{\boldsymbol{ww}})$, $\boldsymbol{\epsilon}_\star \sim \mathcal{N}\left(0, \sigma^2\mathbf{I}\right)$;
      Compute $\boldsymbol{y}_\star \leftarrow \boldsymbol{\Phi}_t^\dagger \boldsymbol{w}_\star + \boldsymbol{\epsilon}_\star$;
      Draw posterior weights (30) $\boldsymbol{w}_{t\star} \leftarrow \boldsymbol{w}_\star - \mathbf{K}_{\boldsymbol{w}\boldsymbol{y}_t}\mathbf{K}_{\boldsymbol{y}_t\boldsymbol{y}_t}^{-1}(\boldsymbol{y}_t - \boldsymbol{y}_\star)$;
      Draw posterior samples $\boldsymbol{z}_{t\star} \leftarrow \boldsymbol{\Phi}_\star^\dagger \boldsymbol{w}_{t\star}$ and $\boldsymbol{\epsilon}_{t\star} \leftarrow \boldsymbol{y}_t - \boldsymbol{\Phi}_t^\dagger \boldsymbol{w}_{t\star}$;
      Accumulate model noise Σ_{σ^2} += $\|\boldsymbol{\epsilon}_{t\star}\|_2^2$, N_{σ^2} += N_t ;
      Accumulate autocorrelation $\widetilde{R}_j$:
        for $j = -N$ **to** N **do**
          for $n = 0$ **to** $N - |j|$ **do**
            $\widetilde{R}_j$ += $z_n^{t\star} z_{n+|j|}^{t\star}/ST(N - |j| + 1)$;
          end
        end
    end
  end
  Update $\boldsymbol{\theta}$ by solving constrained least-squares problem (29) ;
  Update noise $\sigma^2 \leftarrow \Sigma_{\sigma^2}/N_{\sigma^2}$
end

Algorithm 1: Modified Expectation-Maximisation algorithm for training large, stationary 1D GPs using Matheron's update rule

Appendix B: Gaussian blob initial condition

To bootstrap the training of the model hyperparameters, we require an initial guess for the coefficients $\mathbf{H}_{\boldsymbol{\kappa}}$ in (21) which are consistent with the constraints upon the model. The idea is to use a covariance function resembling a Gaussian blob which is projected onto the constraints and then tuned so that the variance matches single-point statistics of the data.

After fitting (25) to the data, we fit a profile to the single-point, normal Reynolds stresses to velocity fluctuations as e.g.

$$\langle u_1' u_1' \rangle(\mathbf{x}) = \alpha_{11,m} B_m(x_2) \tag{B1}$$

by minimising the sum-of-squares residual

$$\sum_{t=1}^{T}\sum_{n=1}^{N_t} \left((\mathbf{e}_1 \cdot \mathbf{v}_{n,t})^2 - \alpha_{11,m} B_m(\mathbf{x}_{n,t} \cdot \mathbf{e}_2)\right)^2 \tag{B2}$$

over the coefficients $\alpha_{11,m}$. This initial guess is biased by the noise, but

We specify Gaussian blob model

$$\mathbf{R}_g(x_2, x_2', r_1, r_3) = \sum_{l=-L}^{L} \sum_{n=-N}^{N} \mathbf{\Phi}_g^\dagger(x_2) \mathbf{\Lambda} \mathbf{\Phi}_g(x_2') e^{-(\lambda_l^2 \ell_1^2 + \nu_n^2 \ell_3^2)} e^{\mathrm{i}\lambda_l r_1 + \mathrm{i}\nu_n r_3} \tag{B3}$$

where $\mathbf{\Phi}_g \in \mathbb{R}^{3M \times 3}$ correspond to a set of Gaussian RBF features centred on grid points $x_{2,m}$

$$\mathbf{\Phi}_g(x) = \begin{bmatrix} \boldsymbol{\phi}_g(x) & 0 & 0 \\ 0 & \boldsymbol{\phi}_g(x) & 0 \\ 0 & 0 & \boldsymbol{\phi}_g(x) \end{bmatrix}, \ \boldsymbol{\phi}_g(x) = \{e^{-(x-x_{2,1})^2/\ell_2^2}, \ldots, e^{-(x-x_{2,M})^2/\ell_2^2}\}^\dagger \tag{B4}$$

The parameters ℓ_1, ℓ_2 and ℓ_3 represent characteristic widths of the Gaussian blob in each principal direction. For the turbulent channel flow, we choose $\ell_1 = 0.2h, \ell_2 = 0.05h$ and $\ell_3 = 0.1h$, which roughly correspond to the correlation length-scales found in the flow. The matrix $\mathbf{\Lambda} \in \mathbb{R}^{3M \times 3M}$ is diagonal with non-negative entries. It represents the variance of the Gaussian RBF features and prescribes the spatial dependence of the variance. We now seek to project (B3) onto (21) in a manner which best approximates (B3) but satisfies the constraints in the spline basis. Matching Fourier terms between (B3) and (21), we have the kernels

$$\begin{aligned} \mathbf{R}_{g,\boldsymbol{\kappa}}(x_2, x_2') &= \mathbf{\Phi}_g^\dagger(x_2) \mathbf{\Lambda} \mathbf{\Phi}_g(x_2') e^{-(\lambda_l^2 \ell_1^2 + \nu_n^2 \ell_3^2)} \\ \mathbf{R}_{\boldsymbol{\kappa}}(x_2, x_2') &= \mathbf{\Phi}^\dagger(x_2) \mathbf{Q}_{\boldsymbol{\kappa}} \mathbf{\Theta}_{\boldsymbol{\kappa}} \mathbf{\Theta}_{\boldsymbol{\kappa}}^\dagger \mathbf{Q}_{\boldsymbol{\kappa}}{}^\dagger \mathbf{\Phi}(x_2') \end{aligned} \tag{B5}$$

We would like to approximate $\mathbf{R}_{\boldsymbol{\kappa}}(x_2, x_2') \approx \mathbf{R}_{g,\boldsymbol{\kappa}}(x_2, x_2')$. Since the basis functions are different (constrained splines versus unconstrained Gaussian RBFs), we can at best make an approximate correspondence over a set of grid points. Evaluating the Fourier coefficients of the Gaussian blob kernel (B5) at grid points $x_{2,m}, x_{2,m'}'$ yields the positive definite Hermitian matrix

$$\mathbf{R}_{g,\boldsymbol{\kappa},\boldsymbol{x}} = \begin{bmatrix} \mathbf{R}_{g,\boldsymbol{\kappa}}(x_{2,1}, x_{2,1}) & \ldots & \mathbf{R}_{g,\boldsymbol{\kappa}}(x_{2,1}, x_{2,M}) \\ \vdots & \ddots & \vdots \\ \mathbf{R}_{g,\boldsymbol{\kappa}}(x_{2,M}, x_{2,1}) & \ldots & \mathbf{R}_{g,\boldsymbol{\kappa}}(x_{2,M}, x_{2,M}) \end{bmatrix} \in \mathbb{C}^{3M \times 3M} \tag{B6}$$

We define $\mathbf{R}_{\boldsymbol{\kappa},\boldsymbol{x}}$ similarly for $\mathbf{R}_{\boldsymbol{\kappa}}(x_2, x_2')$. Both can be obtained by evaluating the spline/Gaussian RBF features at the grid points with features $\mathbf{\Phi}_{\boldsymbol{x}} = [\Phi(x_{2,1}), \ldots, \Phi(x_{2,M})]$ and $\mathbf{\Phi}_{g,\boldsymbol{x}} = [\Phi_g(x_{2,1}), \ldots, \Phi_g(x_{2,M})]$. We then find the kernel hyperparameters which minimise the Frobenius norm distance

$$\|\mathbf{R}_{g,\boldsymbol{\kappa},\boldsymbol{x}} - \mathbf{R}_{\boldsymbol{\kappa},\boldsymbol{x}}\|_F^2 \tag{B7}$$

We note that, from Parseval's theorem, minimising (B7) for each wavenumber is equivalent to minimising the sum-of-squares distance between (21) and (B3). The solution is

$$\mathbf{H}_{\boldsymbol{\kappa}} = \mathbf{Q}_{\boldsymbol{\kappa}} \mathbf{P}_{\boldsymbol{\kappa},\boldsymbol{x}}^\dagger \mathbf{R}_{g,\boldsymbol{\kappa},\boldsymbol{x}} \mathbf{P}_{\boldsymbol{\kappa},\boldsymbol{x}} \mathbf{Q}_{\boldsymbol{\kappa}}^\dagger \tag{B8}$$

where $\mathbf{P}_{\boldsymbol{\kappa},\boldsymbol{x}}$ is the Moore-Penrose pseudoinverse of $\mathbf{Q}_{\boldsymbol{\kappa}}{}^\dagger \mathbf{\Phi}_{\boldsymbol{x}}$.

It remains to specify the variance of the Gaussian blob features $\mathbf{\Lambda}$. Substituting the solution (B8) into (21), the covariance of the velocity field at $r_2 = r_3 = 0$ over the grid points $x_{2,m}, x_{2,m}'$ is

$$\mathbf{R}_{\boldsymbol{x}} = \sum_{\boldsymbol{\kappa}} \mathbf{\Phi}_{\boldsymbol{x}}^\dagger \mathbf{Q}_{\boldsymbol{\kappa}} \mathbf{P}_{\boldsymbol{\kappa},\boldsymbol{x}}^\dagger \mathbf{\Phi}_{g,\boldsymbol{x}}^\dagger \mathbf{\Lambda} \mathbf{\Phi}_{g,\boldsymbol{x}} \mathbf{P}_{\boldsymbol{\kappa},\boldsymbol{x}} \mathbf{Q}_{\boldsymbol{\kappa}}{}^\dagger \mathbf{\Phi}_{\boldsymbol{x}} \tag{B9}$$

and the velocity variance corresponds to the diagonal of this matrix. We then set up a linear system of $3M$ equations in $3M$ unknowns corresponding to the diagonal entries of (B9) and $\mathbf{\Lambda}$. This is solved in the least squares sense, subject to the constraint that $\mathbf{\Lambda}$ is non-negative.

---

[1] Andreas Schröder and Daniel Schanz. Annual review of fluid mechanics 3d lagrangian particle tracking in fluid mechanics. *Annu. Rev. Fluid Mech. 2023*, 55:511–540, 2023. doi:10.1146/annurev-fluid-031822. URL `https://doi.org/10.1146/annurev-fluid-031822-`.

[2] Christian J. Kähler, Sven Scharnowski, and Christian Cierpka. On the resolution limit of digital particle image velocimetry. *Experiments in Fluids*, 52:1629–1639, 6 2012. ISSN 0723-4864. doi:10.1007/s00348-012-1280-x. URL http://link.springer.com/10.1007/s00348-012-1280-x.

[3] A Sciacchitano, B Leclaire, and A Schröder. Main results of the first data assimilation challenge. In *14th International Symposium on Particle Image Velocimetry – ISPIV2021*, 8 2021.

[4] L. D.C. Casa and P. S. Krueger. Radial basis function interpolation of unstructured, three-dimensional, volumetric particle tracking velocimetry data. *Measurement Science and Technology*, 24, 2013. ISSN 13616501. doi:10.1088/0957-0233/24/6/065304.

[5] Sebastian Gesemann, Florian Huhn, Daniel Schanz, and Andreas Schröder. From noisy particle tracks to velocity, acceleration and pressure fields using b-splines and penalties. pages 4–7, 2016.

[6] Philipp Godbersen, Sebastian Gesemann, Daniel Schanz, and Andreas Schröder. Flowfit3: Efficient data assimilation of lpt measurements. In *21th International Symposium on Application of Laser and Imaging Techniques to Fluid Mechanics*, 2024.

[7] Y J ; Jeon, J F G ; Schneiders, M ; Müller, D ; Michaelis, B Wieneke, Y J Jeon, J F G Schneiders, M Müller, and D Michaelis. 4d flow field reconstruction from particle tracks by vic+ with additional constraints and multigrid approximation. In *18th International Symposium on Flow Visualization*, 2018. doi:10.3929/ethz-b-000279199. URL https://doi.org/10.3929/ethz-b-000279199.

[8] Jan F.G. Schneiders and Fulvio Scarano. Dense velocity reconstruction from tomographic ptv with material derivatives. *Experiments in Fluids*, 57, 9 2016. ISSN 07234864. doi:10.1007/s00348-016-2225-6.

[9] Pietro Sperotto, Sandra Pieraccini, and Miguel A. Mendez. A meshless method to compute pressure fields from image velocimetry. *Measurement Science and Technology*, 33, 9 2022. ISSN 13616501. doi:10.1088/1361-6501/ac70a9.

[10] Bora O. Cakir, Gabriel Gonzalez Saiz, Andrea Sciacchitano, and Bas van Oudheusden. Dense interpolations of lpt data in the presence of generic solid objects. *Measurement Science and Technology*, 33, 12 2022. ISSN 13616501. doi:10.1088/1361-6501/ac8ec7.

[11] Manuel Ratz and Miguel A. Mendez. A meshless and binless approach to compute statistics in 3d ensemble ptv. *Experiments in Fluids*, 65, 9 2024. ISSN 14321114. doi:10.1007/s00348-024-03878-x.

[12] Matteo Novara, Daniel Schanz, Nico Reuther, Christian J. Kähler, and Andreas Schröder. Lagrangian 3d particle tracking in high-speed flows: Shake-the-box for multi-pulse systems. *Experiments in Fluids*, 57, 8 2016. ISSN 07234864. doi:10.1007/s00348-016-2216-7.

[13] M. Novara, D. Schanz, and A. Schröder. Two-pulse 3d particle tracking with shake-the-box. *Experiments in Fluids*, 64, 5 2023. ISSN 14321114. doi:10.1007/s00348-023-03634-7.

[14] Noel A. C. Cressie. *Statistics for Spatial Data*. Wiley, 9 1993. ISBN 9780471002550. doi:10.1002/9781119115151. URL https://onlinelibrary.wiley.com/doi/book/10.1002/9781119115151.

[15] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, 11 2005. ISBN 9780262256834. doi:10.7551/mitpress/3206.001.0001. URL https://direct.mit.edu/books/book/2320/Gaussian-Processes-for-Machine-Learning.

[16] Laura P. Swiler, Mamikon Gulian, Ari L. Frankel, Cosmin Safta, and John D. Jakeman. A survey of constrained gaussian process regression: Approaches and implementation challenges. *Journal of Machine Learning for Modeling and Computing*, 1:119–156, 6 2020. ISSN 2689-3967. doi:10.1615/JMachLearnModelComput.2020035155. URL http://www.dl.begellhouse.com/journals/558048804a15188a,2cbcbe11139f18e5,0776649265326db4.html.

[17] Maziar Raissi, Paris Perdikaris, and George Em Karniadakis. Machine learning of linear differential equations using gaussian processes. *Journal of Computational Physics*, 348:683–693, 2017. ISSN 0021-9991. doi:https://doi.org/10.1016/j.jcp.2017.07.050. URL https://www.sciencedirect.com/science/article/pii/S0021999117305582.

[18] Carl Jidling, Niklas Wahlström, Adrian Wills, and Thomas B. Schön. Linearly constrained gaussian processes. In *31st Conference on Neural Information Processing Systems (NIPS 2017)*, 3 2017. doi:10.48550/arXiv.1703.00787. URL http://arxiv.org/abs/1703.00787.

[19] Francis J. Narcowich and Joseph D. Ward. Generalized hermite interpolation via matrix-valued conditionally positive definite functions. *Mathematics of Computation*, 63:661, 10 1994. ISSN 00255718. doi:10.2307/2153288. URL https://www.jstor.org/stable/2153288?origin=crossref.

[20] Ronald J Adrian. Stochastic estimation of conditional structure: a review, 1994.

[21] J. Borée. Extended proper orthogonal decomposition: A tool to analyse correlated events in turbulent flows. *Experiments in Fluids*, 35:188–192, 8 2003. ISSN 07234864. doi:10.1007/s00348-003-0656-3.

[22] S B Pope. *Turbulent Flows*. Cambridge University Press, 2000. ISBN 0521598869. doi:10.1088/1468-5248/1/1/702.

[23] Steven L Brunton, Bernd R Noack, and Petros Koumoutsakos. Machine learning for fluid mechanics. *Annu. Rev. Fluid Mech. 2020*, 52:477–508, 2019. doi:10.1146/annurev-fluid-010719. URL https://doi.org/10.1146/annurev-fluid-010719-.

[24] J. Cortina-Fernández, C. Sanmiguel Vila, A. Ianiro, and S. Discetti. From sparse data to high-resolution fields: ensemble particle modes as a basis for high-resolution flow characterization. *Experimental Thermal and Fluid Science*, 120, 1 2021. ISSN 08941777. doi:10.1016/j.expthermflusci.2020.110178.

[25] Douglas W. Carter, Francis De Voogt, Renan Soares, and Bharathram Ganapathisubramani. Data-driven sparse reconstruction of flow over a stalled aerofoil using experimental data. *Data-Centric Engineering*, 2, 5 2021. ISSN 26326736. doi:10.1017/dce.2021.5.

[26] Nereida Agüera, Gioacchino Cafiero, Tommaso Astarita, and Stefano Discetti. Ensemble 3d ptv for high resolution turbu-

lent statistics. *Measurement Science and Technology*, 27, 10 2016. ISSN 13616501. doi:10.1088/0957-0233/27/12/124011.

[27] Philipp Godbersen and Andreas Schröder. Functional binning: Improving convergence of eulerian statistics from lagrangian particle tracking. *Measurement Science and Technology*, 31, 9 2020. ISSN 13616501. doi:10.1088/1361-6501/ab8b84.

[28] Haitao Liu, Yew-Soon Ong, Xiaobo Shen, and Jianfei Cai. When gaussian process meets big data: A review of scalable gps. *IEEE Transactions on Neural Networks and Learning Systems*, 31:4405–4423, 11 2020. ISSN 2162-237X. doi: 10.1109/TNNLS.2019.2957109. URL `https://ieeexplore.ieee.org/document/8951257/`.

[29] Iacopo Tirelli, Miguel Alfonso Mendez, Andrea Ianiro, and Stefano Discetti. A meshless method to compute the pod and its variants from scattered data. In *21th International Symposium on Application of Laser and Imaging Techniques to Fluid Mechanics*, 2024.

[30] Gal Berkooz, Philip Holmes, and John L Lumley. The proper orthogonal decomposition in the analysis of turbulent flows, 1993. URL `www.annualreviews.org/aronline`.

[31] Christopher M Bishop. *Pattern Recognition and Machine Learning*. Springer New York, 8 2006. ISBN 978-0-387-31073-2.

[32] Pietro Sperotto, Bo Watz, and David Hess. Meshless track assimilation (mta) of 3d ptv data. *Measurement Science and Technology*, 35, 8 2024. ISSN 13616501. doi:10.1088/1361-6501/ad3f36.

[33] John M. Lawson and Bharathram Ganapathisubramani. Unsteady forcing of turbulence by a randomly actuated impeller array. *Experiments in Fluids*, 63, 1 2022. ISSN 14321114. doi:10.1007/s00348-021-03364-8.

[34] Vejapong Juttijudata, John L. Lumley, and Dietmar Rempfer. Proper orthogonal decomposition in squire's coordinate system for dynamical models of channel turbulence. *Journal of Fluid Mechanics*, 534:195–225, 7 2005. ISSN 00221120. doi:10.1017/S0022112005004404.

[35] James T. Wilson, Viacheslav Borovitskiy, Alexander Terenin, Peter Mostowsky, and Marc Peter Deisenroth. Pathwise conditioning of gaussian processes. *Journal of Machine Learning Research*, 11 2020. URL `http://arxiv.org/abs/2011.04026`.

[36] Hassan Maatouk, Didier Rullière, and Xavier Bay. *Large Scale Gaussian Processes with Matheron's Update Rule and Karhunen-Loève Expansion*, volume 460, pages 469–487. Springer, 2024. doi:10.1007/978-3-031-59762-6˙23. URL `https://link.springer.com/10.1007/978-3-031-59762-6_23`.

[37] Cristian C. Lalescu, Bérenger Bramas, Markus Rampp, and Michael Wilczek. An efficient particle tracking algorithm for large-scale parallel pseudo-spectral simulations of turbulence. *Computer Physics Communications*, 278, 9 2022. ISSN 00104655. doi:10.1016/j.cpc.2022.108406.

[38] Mikael Mortensen. A spectral-galerkin turbulent channel flow solver for large-scale simulations. 1 2017. URL `http://arxiv.org/abs/1701.03787`.

[39] Markus Raffel, Christian E. Willert, Fulvio Scarano, Christian J. Kähler, Steve T. Wereley, and Jürgen Kompenhans. *Particle Image Velocimetry*. Springer International Publishing, 2018. ISBN 978-3-319-68851-0. doi:10.1007/978-3-319-68852-7. URL `https://link.springer.com/10.1007/978-3-319-68852-7`.

[40] Alexander H. Barnett, Jeremy Magland, and Ludvig af Klinteberg. A parallel nonuniform fast fourier transform library based on an "exponential of semicircle" kernel. *SIAM Journal on Scientific Computing*, 41(5):C479–C504, 2019. doi: 10.1137/18M120885X. URL `https://doi.org/10.1137/18M120885X`.

[41] A. K. Saha, K. Muralidhar, and G. Biswas. Experimental study of flow past a square cylinder at high reynolds numbers. *Experiments in Fluids*, 29:553–563, 12 2000. ISSN 0723-4864. doi:10.1007/s003480000123. URL `http://link.springer.com/10.1007/s003480000123`.

[42] Anagha Madhusudanan, Simon. J. Illingworth, and Ivan Marusic. Coherent large-scale structures from the linearized navier–stokes equations. *Journal of Fluid Mechanics*, 873:89–109, 8 2019. ISSN 0022-1120. doi:10.1017/jfm.2019.391. URL `https://www.cambridge.org/core/product/identifier/S0022112019003914/type/journal_article`.

[43] J. M. Lawson. Dataset for "physics informed gaussian process regression for particle tracking data assimilation", 2026. URL `https://doi.org/10.5258/SOTON/D3642`.