

PIGLET: Physics Informed Gaussian Process Regression for the Learning and Estimation of Turbulence

J. M. Lawson ^{*1}

1: Dept. of Aeronautics and Astronautics, University of Southampton, UK ^{*}Corresponding author: j.m.lawson@soton.ac.uk

Keywords: PTV Techniques, Artificial Learning Methods, Turbulence, Volumetric PIV/PTV

ABSTRACT

We introduce PIGLET: an open source framework for physics-informed Gaussian process regression (GPR), which can be used to learn statistical descriptions of turbulent flows and interpolate and quantify uncertainty in turbulent flow fields obtained from particle tracking velocimetry (PTV) and Lagrangian particle tracking (LPT) data. Building on our previous work (Lawson, 2026), we describe how Gaussian process models can be formulated to incorporate Navier-Stokes dynamics in addition to incompressibility, boundary conditions and statistical symmetries we have previously documented. We describe a simple, finite-dimensional Gaussian process model, suitable for the generation of probable realisations of a homogeneous turbulent flow field given velocity information sampled by passive tracers. The model can incorporate acceleration information using a Bayesian update. This formulation mitigates the main bottleneck of GPR, allowing reconstruction of flow fields using $\mathcal{O}(100,000)$ particle tracks. The result is a technique which can incorporate both non-linear and linear physical constraints as well as statistical symmetries into the estimation and uncertainty quantification of turbulent flow fields from LPT and PTV data. We train and test this model upon synthetic data from homogeneous isotropic turbulence (HIT), the turbulent duct flow from the 2nd Data Assimilation Challenge, and experimental data from a square prism wake (SPW). PIGLET performs favourably in comparison to the state-of-the-art.

]

1. Introduction

The past decade has seen widespread adoption of Lagrangian particle tracking (LPT) and particle tracking velocimetry (PTV) to measure flows (Schröder & Schanz, 2023). These techniques are capable of tracking hundreds of thousands of tracer particles in space and time, which provide access to velocity and acceleration information sampled at irregular locations in the flow along particle trajectories. A central issue is the interpolation of these scattered data, usually onto Cartesian grids, for subsequent analysis with the least possible error (Sciacchitano et al., 2021; Schröder & Schanz, 2023). It is desirable to incorporate known flow physics, e.g. mass, momentum conservation 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 and two-point correlations. We refer to the problem of combining PTV/LPT data with these known physics to yield volumetric flow fields and statistics as the Data Assimilation (DA) problem.

As showcased in the first LPT data assimilation challenge (Sciacchitano et al., 2021), the majority of solutions to the DA problem so far are based upon formulating and solving nonlinear regression problems (Casa & Krueger, 2013; Gesemann et al., 2016; Godbersen et al., 2024; Jeon et al., 2018; Schneiders & Scarano, 2016; Sperotto et al., 2022). These minimise a cost function which penalises the residual between observations of velocity and acceleration sampled on particle trajectories. The minimisation is performed with respect to a set of weights which parametrise the velocity and Lagrangian acceleration fields. Penalties may be introduced into the cost function to enforce smoothness of the solution (Gesemann et al., 2016; Jeon et al., 2018) or to introduce soft constraints, e.g. boundary conditions (Casa & Krueger, 2013; Cakir et al., 2022) and mass and momentum conservation (Casa & Krueger, 2013; Gesemann et al., 2016; Jeon et al., 2018). Alternatively, hard constraints may be introduced either in the formulation of the model (Schneiders & Scarano, 2016; Jeon et al., 2018; Cakir et al., 2022) or in the determination of the model coefficients (Godbersen et al., 2024; Ratz & Mendez, 2024). By incorporating these physical constraints, spatial resolution can be improved by a factor of three to four over naïve linear interpolation (Sciacchitano et al., 2021).

The main deficiency of these approaches is that they do not incorporate statistical information about the flow field. Physics Informed Neural Networks (PINNs) represent one way of doing so (Shin & Schröder, 2025; Mons et al., 2022; Steinfurth et al., 2024; Zhou & Grauer, 2023). This involves training neural networks (usually multilayer perceptrons) with a loss function which penalises not only the closeness of fit to training data but also the degree to which known physical constraints are violated. PINNs have demonstrated promising performance over the above regression methods in preliminary results from the second LPT data assimilation challenge. However, the uncertainty of the resulting output remains unquantified by both PINNs and regression approaches.

Recently, we introduced physics-informed Gaussian process regression to address the DA problem (Lawson, 2026). By leveraging statistical information and knowledge of boundary conditions, physics-informed GPR can provide an order of magnitude improvement in the mean-square velocity error over to the vortex in cell method (Jeon et al., 2018). Furthermore, GPR provides closed form, interpretable expressions for the most likely flow field given the input data as well as a statistical description of the underlying flow, which is often the primary objective of measurement. This provides an advantage over “black-box” deep learning approaches, including PINNs, which are harder to interpret (Brunton et al., 2019). Moreover, GPR also provides closed form expressions to draw samples from a statistical distribution of velocity fields which are consistent with the ob-

served LPT data using the Matheron update rule (Wilson et al., 2020). This allows us to generate a Monte-Carlo sampling of the most likely velocity fields which are consistent with the data, suitable for uncertainty quantification of post-processed outputs.

A limitation of our previous work was that the Gaussian process regression model did not leverage Navier-Stokes dynamics to reconstruct the flow. This is significant because, for data assimilation with regression methods, the incorporation of acceleration data has improved reconstruction accuracy (Gesemann et al., 2016; Schneiders & Scarano, 2016). A further limitation of our previous work is the computational bottleneck of GPR: with N input data, it is necessary to invert a dense kernel matrix with N^2 elements which requires $\sim N^3$ operations. In double precision with $N = 10^5$ particles, this would require 720GB of memory and $\sim 10^{16}$ floating-point operations per reconstruction, which necessitates high-performance computing. Our insight is that the GPR problem can be formulated as a Bayesian inference problem which is solved using standard optimisation techniques using much less memory and processor time. The problem can be augmented to include acceleration information, which is a non-linear function of the underlying Gaussian process. However, to do this, the model must be formulated as a finite dimensional Gaussian process.

In this work, we describe a physics-informed Gaussian process regression model which incorporates Navier-Stokes dynamics through Bayesian inference. We restrict ourselves to the case of two or four frame LPT, where the objective is to recover the flow field from short tracks providing a measurement of the position, velocity and acceleration of tracers at a single time instant. However, the principles described here can be extended to time series LPT measurements. The model represents the underlying velocity fluctuations as statistically homogeneous and incompressible, but with inhomogeneous mean flow. It is applicable to both homogeneous and inhomogeneous turbulent flows. The model is implemented in our open source software, PIGLET. We train and test this model upon synthetic and experimental data from homogeneous isotropic turbulence (HIT), a square prism wake (SPW) and the turbulent duct flow from the 2nd Data Assimilation Challenge.

2. Methodology

This section is structured as follows. In §2.1, we explain the theoretical framework of physics-informed Gaussian process regression to apply GPR to LPT data. In §2.2 we describe a Gaussian process models for homogeneous turbulence using this framework. In §2.3, we explain how this model can be trained upon experimental data to learn the statistical properties of the flow and noise. Finally, we describe our experimental and numerical datasets in §2.4.

2.1. Physics Informed Gaussian Process Regression of LPT data

Consider the instantaneous velocity field $U(\mathbf{x}, t)$ and acceleration field $A(\mathbf{x}, t)$. We will treat the statistics of this field as stationary in time t . The spatial coordinate within the measurement domain Ω is $\mathbf{x} \in \Omega \subset \mathbb{R}^3$. Velocity data are available for each of $t = 1 \dots T$ snapshots, with N_t particles per snapshot at input points $\mathbf{X}_t = \{\mathbf{x}_{n,t}\}_{n=1}^{N_t}$. At a given time t , the velocity data are noisy observations $\mathbf{y}_t = (\mathbf{V}_1; \mathbf{V}_2; \dots; \mathbf{V}_{N_t}) \in \mathbb{R}^{3N_t}$ of the velocity field $\mathbf{V}_n = U(\mathbf{x}_n, t) + \epsilon_n$ sampled on the tracers. We treat the additive Gaussian noise ϵ_n noise as independent of the data, statistically stationary, spatially white and isotropic with variance σ_u^2 . Likewise, we also have observations $\mathbf{z}_t = (\mathbf{B}_1; \mathbf{B}_2; \dots; \mathbf{B}_{N_t}) \in \mathbb{R}^{3N_t}$ of the noisy material acceleration vector $\mathbf{B}_n = A(\mathbf{x}_n, t) + \zeta_n$ with noise variance σ_a^2 .

We suppose that a M-dimensional Gaussian Process has generated the velocity fields

$$U(\mathbf{x}) = \Phi^\dagger(\mathbf{x})\mathbf{W} \quad (1)$$

where the feature vector $\Phi(\mathbf{x})$ represents a set of basis functions parametrised by a latent variable of weights $\mathbf{W} \in \mathbb{R}^M$ drawn from the normal distribution $\mathcal{N}(\overline{\mathbf{W}}, \mathbf{K}_{ww})$ with mean $\overline{\mathbf{W}}$ and covariance $\mathbf{K}_{ww}$ with rank R . It is possible for the process to be degenerate so that $R < M$. For convenience, we write the Reynolds decomposition $\mathbf{W} = \overline{\mathbf{W}} + \mathbf{w}$, where $\mathbf{w}$ is the fluctuating component of the weight. Samples of this prior distribution can be easily generated as

$$\mathbf{W} = \overline{\mathbf{W}} + \mathbf{L}_{ww}\boldsymbol{\eta} \quad (2)$$

where $\boldsymbol{\eta} \in \mathbb{R}^R$ is drawn from the standard normal distribution and $\mathbf{L}_{ww}$ is some factorisation of the covariance such that $\mathbf{K}_{ww} = \mathbf{L}_{ww}\mathbf{L}_{ww}^\dagger$. Note that this Gaussian prior completely specifies the mean flow $\overline{U}(\mathbf{x}) = \Phi^\dagger(\mathbf{x})\overline{\mathbf{W}}$ and the two-point velocity correlation tensor $\mathbf{R}_{uu}(\mathbf{x}, \mathbf{x}') = \Phi^\dagger(\mathbf{x})\mathbf{K}_{ww}\Phi(\mathbf{x}')$ of the turbulence. We provide physics-informed choices of $\Phi(\mathbf{x})$ and methods to learn $\mathbf{K}_{ww}$ and $\overline{\mathbf{W}}$ for homogenous turbulent flows in §2.2.

We represent this process as being measured on the input points $\mathbf{X}_t$ as

$$\mathbf{y}_t = \Phi_t^\dagger \mathbf{W}_t + \epsilon_t \quad (3)$$

where $\Phi_t \in \mathbb{R}^{M \times 3N_t}$ represents the feature vector sampled on the input points $\mathbf{X}_t$. Therefore, the underlying process and the noisy observations are joint normally distributed

$$\begin{bmatrix} \mathbf{W} \\ \mathbf{y}_t \end{bmatrix} \sim \mathcal{N} \left(\begin{bmatrix} \overline{\mathbf{W}} \\ \Phi_t^\dagger \overline{\mathbf{W}} \end{bmatrix}, \begin{bmatrix} \mathbf{K}_{ww} & \mathbf{K}_{wy_t} \\ \mathbf{K}_{y_t w} & \mathbf{K}_{y_t y_t} \end{bmatrix} \right) \quad (4)$$

where $\mathbf{K}_{y_t w} = \mathbf{K}_{wy_t}^\dagger = \Phi_t^\dagger \mathbf{K}_{ww}$ represents the covariance of the velocity measurements and the weights and $\mathbf{K}_{y_t y_t} = \Phi_t^\dagger \mathbf{K}_{ww} \Phi_t + \sigma_u^2 \mathbf{I}$ represents the covariance of the noisy velocity data.

Since $\mathbf{W}$ and $\mathbf{y}_t$ are joint normally distributed, the conditional distribution of the weights $\mathbf{W}$ given the observations $\mathbf{y}_t$ is also a normal distribution with mean

$$\boldsymbol{\mu}_w = \overline{\mathbf{W}} + \mathbf{K}_{w\mathbf{y}_t} \mathbf{K}_{\mathbf{y}_t\mathbf{y}_t}^{-1} (\mathbf{y}_t - \Phi_t^\dagger \overline{\mathbf{W}}). \quad (5)$$

This represents an update of the unconditional mean weight $\overline{\mathbf{W}}$ with the velocity information in $\mathbf{y}_t$ to generate the conditional mean $\boldsymbol{\mu}_w$. The conditional variance is

$$\Sigma_{ww} = \mathbf{K}_{ww} - \mathbf{K}_{w\mathbf{y}_t} \mathbf{K}_{\mathbf{y}_t\mathbf{y}_t}^{-1} \mathbf{K}_{\mathbf{y}_t w} \quad (6)$$

which represents an update to the unconditional variance $\mathbf{K}_{ww}$ due to the information in $\mathbf{y}_t$.

The conditional mean weights $\boldsymbol{\mu}_w$ (5) are the most likely set of weights given only the velocity data and that this choice minimises the mean squared error $|\mathbf{U}(\mathbf{x}) - \boldsymbol{\mu}_u(\mathbf{x})|^2$ (Bishop, 2006). Therefore, we can use (5) to reconstruct the most likely velocity field $\boldsymbol{\mu}_u(\mathbf{x}) = \Phi^\dagger(\mathbf{x}) \boldsymbol{\mu}_w$ based on the velocity data. However, the great advantage of Gaussian Process Regression is that we have also specified the *distribution* of the possible values of the velocity field around this mean. We can efficiently draw samples from this conditional distribution using the Matheron update rule (Wilson et al., 2020). From this easily verified rule it follows that

$$\widehat{\mathbf{W}} = \mathbf{W} + \mathbf{K}_{w\mathbf{y}_t} \mathbf{K}_{\mathbf{y}_t\mathbf{y}_t}^{-1} (\mathbf{y}_t - \Phi_t^\dagger \mathbf{W} - \boldsymbol{\epsilon}_t) \quad (7)$$

is distributed with mean $\boldsymbol{\mu}_w$ (5) and variance Σ_{ww} (6), where $\mathbf{W}$ and $\boldsymbol{\epsilon}_t$ are drawn from their respective, prior distributions as described above. Conditional realisations of the velocity field are obtained as $\widehat{\mathbf{U}}(\mathbf{x}) = \Phi^\dagger(\mathbf{x}) \widehat{\mathbf{W}}$. We can use these realisations to derive local estimates of the measurement uncertainty or propagate uncertainty using Monte-Carlo techniques.

A disadvantage of (5) is that it requires the computationally expensive formulation and inversion of the $3N_t \times 3N_t$ kernel $\mathbf{K}_{\mathbf{y}_t\mathbf{y}_t}$. To avoid this, we note that the log-likelihood of the normalised weights $\boldsymbol{\eta}$ and noise $\boldsymbol{\epsilon}_t$ given the data $\mathbf{y}_t$ and the Gaussian prior $\boldsymbol{\theta} = \{\overline{\mathbf{W}}, \mathbf{L}_{ww}, \sigma_u\}$ is

$$\log p(\boldsymbol{\eta}|\mathbf{y}_t) = \log \frac{p(\mathbf{y}_t|\boldsymbol{\eta})p(\boldsymbol{\eta})}{p(\mathbf{y}_t)} = -\frac{1}{2} \boldsymbol{\eta}^\dagger \boldsymbol{\eta} - \frac{1}{2\sigma_u^2} \left\| \mathbf{y}_t - \Phi_t^\dagger (\overline{\mathbf{W}} + \mathbf{L}_{ww} \boldsymbol{\eta}) \right\|_2^2 + \text{const.} \quad (8)$$

is maximised when we choose $\boldsymbol{\eta}$ such that $\boldsymbol{\mu}_w = \overline{\mathbf{W}} + \mathbf{L}_{ww} \boldsymbol{\eta}$. In (8), the first term represents the prior probability of the weights whereas the second term represents the probability of the noise $\boldsymbol{\epsilon}_t$ given the data and the weights. Therefore, we solve (5) numerically by maximising (8) using the L-BFGS algorithm. To accelerate this, the initial condition can be obtained from (5) using a subset of the available input data. A similar optimisation problem can be written for (7).

Thus far, we have not specified any physics. The premise of physics informed Gaussian process regression (Lawson, 2026) is to construct a covariance model which satisfies known statistical symmetries, boundary conditions and governing equations of the system we aim to reconstruct. This not only ensures that the GPR prediction (1,5) satisfies the boundary conditions and governing

equations, but it also reduces the degrees of freedom within the covariance model so that it may be learned empirically. For instance, statistical homogeneity implies $\mathbf{R}_{uu}(\mathbf{x}', \mathbf{x}) = \mathbf{R}_{uu}(\mathbf{x}' - \mathbf{x}, 0)$, reducing its spatial dimension by three. This property implies specific choices of $\Phi(\mathbf{x})$ and the structure of $\mathbf{K}_{ww}$, which we will describe shortly. Similarly, the solenoidal constraint upon the velocity field $\nabla \cdot \mathbf{U} = 0$ implies linear constraints of the form $\mathbf{C}\mathbf{W} = \mathbf{0}$ and $\mathbf{C}\mathbf{K}_{ww} = \mathbf{0}$. This has the benefit of reducing the number of model hyperparameters to be learned.

To incorporate the acceleration information, we note that the estimate of the acceleration field is uniquely specified by the normalised process weights $\boldsymbol{\eta}$. Furthermore, the velocity and acceleration error ϵ_t and ζ_t are independently and normally distributed for Gaussian white position noise. Therefore, the Bayesian estimate of the process weights $\boldsymbol{\eta}$ given the velocity and acceleration observations and Gaussian prior is

$$\log p(\boldsymbol{\eta} | \mathbf{y}_t, \mathbf{z}_t) = -\frac{1}{2} \boldsymbol{\eta}^\dagger \boldsymbol{\eta} - \frac{1}{2\sigma_u^2} \left\| \mathbf{y}_t - \Phi_t^\dagger(\overline{\mathbf{W}} + \mathbf{L}_{ww}\boldsymbol{\eta}) \right\|_2^2 - \frac{1}{2\sigma_a^2} \left\| \mathbf{z}_t - \mathbf{A}_t(\boldsymbol{\eta}) \right\|_2^2 + \text{const.} \quad (9)$$

where $\mathbf{A}_t(\boldsymbol{\eta})$ is a vector representing the acceleration field evaluated at the input points $\mathbf{X}_t$. This adds an additional term to (8) representing the probability of the acceleration error given the data and the model.

2.2. A Model for Homogeneous Turbulence

GPR requires us build and learn a model of the two-point velocity covariance and mean flow. In statistically stationary, homogeneous turbulence, the two-point, single-time covariance tensor of any vector field may be written

$$\mathbf{R}(\mathbf{x}', \mathbf{x}) \equiv \langle \mathbf{u}(\mathbf{x}', t) \mathbf{u}^\dagger(\mathbf{x}, t) \rangle = \mathbf{R}(\mathbf{x} - \mathbf{x}', 0) = \iiint \Psi(\boldsymbol{\kappa}) e^{i\boldsymbol{\kappa} \cdot (\mathbf{x}' - \mathbf{x})} d\boldsymbol{\kappa} \quad (10)$$

where $\Psi(\boldsymbol{\kappa})$ is the velocity spectrum tensor (Pope, 2000). The solenoidal condition implies that $\Psi(\boldsymbol{\kappa})\boldsymbol{\kappa} = 0$. Unfortunately, (10) implies an infinite-dimensional weight vector in (1). However, we can approximate (10) using a Fourier series

$$\mathbf{R}_{uu}(\mathbf{x}', \mathbf{x}) = \sum_{\boldsymbol{\kappa}} \Psi_{\boldsymbol{\kappa}} \exp(i\boldsymbol{\kappa} \cdot (\mathbf{x}' - \mathbf{x})) \quad (11)$$

where $\Psi_{\boldsymbol{\kappa}}$ is the spectral energy density tensor associated with each wavenumber $\boldsymbol{\kappa}$. This formulation implies a Fourier series representation of the velocity field

$$\mathbf{U}(\mathbf{x}) = \sum_{\boldsymbol{\kappa}} \mathbf{W}_{\boldsymbol{\kappa}} \exp(i\boldsymbol{\kappa} \cdot \mathbf{x}) \quad (12)$$

where the basis functions $\Phi^\dagger(\mathbf{x}) = (\mathbf{I} \exp(i\boldsymbol{\kappa}_1 \cdot \mathbf{x}), \dots, \mathbf{I} \exp(i\boldsymbol{\kappa}_M \cdot \mathbf{x})) \in \mathbb{C}^{3 \times 3M}$ are Fourier modes and the (complex) weights $\mathbf{W} = (\mathbf{W}_{\boldsymbol{\kappa}_1}; \dots; \mathbf{W}_{\boldsymbol{\kappa}_M})$ are Fourier coefficients. The associated variance

$\mathbf{K}_{ww}$ is a block diagonal matrix, where there is zero correlation between different wavenumbers and the velocity spectrum tensor $\Psi_{\kappa} \equiv \langle \mathbf{w}_{\kappa} \mathbf{w}_{\kappa}^{\dagger} \rangle$ is the variance of mode κ . The Fourier weights are related to real-valued, normalised process weights $\boldsymbol{\eta}$ as $\mathbf{W}_{\kappa} = \overline{\mathbf{W}}_{\kappa} + \Phi_{\kappa}^{1/2} \mathcal{F}_{\kappa} \boldsymbol{\eta}$, where $\mathcal{F}_{\kappa}$ represents the 3D Fourier transform operator acting on the (flattened) real-valued 3D field represented by $\boldsymbol{\eta}$. The mean flow field will satisfy continuity provided that we choose $\overline{\mathbf{W}}_{\kappa} \cdot \boldsymbol{\kappa} = 0$. Note that velocity fluctuations satisfy continuity since $\mathbf{w}_{\kappa} \sim \mathcal{N}(0, \Psi(\boldsymbol{\kappa}))$ ensures that $\mathbf{w}_{\kappa} \cdot \boldsymbol{\kappa} = 0$.

2.3. Training of model hyperparameters

We use the expectation-maximisation algorithm combined with the Matheron update rule (Lawson, 2026) to learn the hyperparameters by iteratively refining an initial guess for the mean and variance of the model parameters. In brief: given T realisations of the velocity measurements $(\mathbf{X}_t, \mathbf{y}_t)$, we draw ST samples of the underlying process weights $\widehat{\mathbf{W}}$ (7) and noise ϵ of the flow under the prior $\boldsymbol{\theta}_0$. From this, we construct a new sample mean $\overline{\mathbf{W}} = \langle \widehat{\mathbf{W}} \rangle$ and variance $\Psi_{\kappa} = \langle \widehat{\mathbf{w}}_{\kappa} \widehat{\mathbf{w}}_{\kappa}^{\dagger} \rangle$ of the Fourier coefficients and the noise variance $\sigma_u^2 = \langle \epsilon^2 \rangle$. The input velocity and acceleration data are obtained from finite differences of the form $\mathbf{V} = \sum c_i \mathbf{x}_i$ and $\mathbf{B} = \sum d_i \mathbf{x}_i$. Under the assumption that the position noise is white-in-time, the acceleration uncertainty is $\sigma_a^2 = \sigma_u^2 \sum d_i^2 / \sum c_i^2$. This specifies a new set of hyperparameters $\boldsymbol{\theta}_1$, which can be iteratively refined.

To create an initial guess of the velocity covariance, we use a Fourier series to approximate the isotropic correlation function

$$\mathbf{R}_{uu}(\mathbf{x}', \mathbf{x}) = R_{LL}(r) \frac{\mathbf{r} \mathbf{r}^{\dagger}}{r^2} + \left(R_{LL}(r) + \frac{1}{2} r \frac{\partial R_{LL}}{\partial r} \right) \left(\mathbf{I} - \frac{\mathbf{r} \mathbf{r}^{\dagger}}{r^2} \right) \quad (13)$$

where $R_{LL}(r)$ is the longitudinal autocorrelation function, which is a scalar function of the separation $r = |\mathbf{r}| = |\mathbf{x}' - \mathbf{x}|$. The corresponding energy spectral density is (Pope, 2000)

$$\Psi_{\kappa} = \frac{E(\kappa)}{4\pi\kappa^2} \left(\mathbf{I} - \frac{\boldsymbol{\kappa} \boldsymbol{\kappa}^{\dagger}}{\kappa^2} \right) \quad (14)$$

where $\boldsymbol{\kappa}$ is the wavenumber and $E(\kappa)$ is the 3D energy spectrum for $\kappa = |\boldsymbol{\kappa}|$. The incompressibility condition is satisfied by construction. For this initial guess, we choose

$$R_{LL} = \langle V^2 \rangle \exp \left(-\frac{r^2}{2\lambda^2} \right) \Leftrightarrow E(\kappa) = \frac{\langle V^2 \rangle}{\sqrt{2\pi}} \lambda^5 \kappa^4 \exp \left(-\frac{\kappa^2 \lambda^2}{2} \right) \quad (15)$$

where $\langle V^2 \rangle$ is the mean-square velocity sampled on the particles and λ is chosen to match an estimate of the Taylor microscale of the flow. The noise is estimated as $\sigma_u = 0.01 \langle V^2 \rangle^{1/2}$.

For the case of homogeneous, isotropic turbulence, we have shown that it is possible to robustly learn the velocity spectrum and noise directly from particle tracks (Lawson, 2026). In this case, we

use the product-of-experts method to fit the noise and the autocorrelation function to the data with the linear model

$$R_{LL}(r) = \sum_{n=1}^N \theta_n \exp\left(-\frac{r^2}{2\lambda_n^2}\right) \Leftrightarrow E(\kappa) = \frac{1}{\sqrt{2\pi}} \sum_{n=1}^N \theta_n \lambda_n^5 \kappa^4 \exp\left(-\frac{\kappa^2 \lambda_n^2}{2}\right) \quad (16)$$

for N hyperparameters $\theta_n \geq 0$, each associated with wavelength λ_n . In contrast to the sinc kernel we have previously used, this choice of squared-exponential kernel ensures that the 3D energy spectrum is smooth and differentiable everywhere. The fit is obtained by maximising the marginal probability of observations of the velocity field given the kernel hyperparameters. To do this, we split our input data into $p = 1 \dots P$ observations of particle pairs $\mathbf{y}_p = (v(\mathbf{x}_p, t), v(\mathbf{x}'_p, t))$ which sample the fluctuating component of the velocity field at coordinates $\mathbf{x}_p$ and $\mathbf{x}'_p$. There are N_t particles per snapshot, so we can form $P \approx TN_t^2/2$ such pairs. Under the product-of-experts approximation, the log likelihood of the hyperparameters $\boldsymbol{\theta} = \{\sigma_u^2, \theta_1, \dots, \theta_N\}$ given the data is

$$J_{iso}(\boldsymbol{\theta}) = \log p(\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_P | \boldsymbol{\theta}) = \sum_{p=1}^P \log p(\mathbf{y}_p | \boldsymbol{\theta}) \quad (17)$$

where the log-likelihood of each particle pair is

$$\log p(\mathbf{y}_p | \boldsymbol{\theta}) = -\log 2\pi - \log |\mathbf{K}_{\mathbf{y}_p \mathbf{y}_p}| - \frac{1}{2} \mathbf{y}_p^\dagger \mathbf{K}_{\mathbf{y}_p \mathbf{y}_p}^{-1} \mathbf{y}_p \quad (18)$$

and $\mathbf{K}_{\mathbf{y}_p \mathbf{y}_p}$ is the input kernel obtained from (13) and (16). The optimal hyperparameters $\boldsymbol{\theta}$ are found by minimising (17) using a standard sequential quadratic programming method in MATLAB, subject to the constraints $\sigma_u \geq 0$ and $\theta_n \geq 0$. The cost of this approach is dominated by the $\mathcal{O}(TN_t^2)$ operations required to compute (17) and takes only minutes on the author's workstation for the datasets presented here.

To create an initial guess of the mean flow, we choose $\overline{\mathbf{W}}_\kappa = \boldsymbol{\Psi}_\kappa^{1/2} \mathcal{F}_\kappa \overline{\boldsymbol{\eta}}$ which ensures that $\overline{\mathbf{W}}$ satisfies incompressibility for any choice of $\overline{\boldsymbol{\eta}}$. We then minimise the regularised sum-of-squares objective

$$J(\overline{\boldsymbol{\eta}}) = \frac{1}{2\langle V^2 \rangle} \sum_{t=1}^T \left\| \mathbf{y}_t - \boldsymbol{\Phi}_t^\dagger \overline{\mathbf{W}} \right\|_2^2 + \frac{1}{2} \overline{\boldsymbol{\eta}}^\dagger \overline{\boldsymbol{\eta}} \quad (19)$$

Conveniently, this is the same form as (8). In fact, it corresponds to an approximation for the log-likelihood of the data given a prior distribution on the mean flow. For the synthetic HIT dataset, the mean flow is trivially zero.

2.4. Datasets

In this section we describe the training of our GP models on synthetic and real experimental PTV data. These have previously been described in Lawson (2026) and we recap only the important

details. The synthetic data are based on direct numerical simulations of passive tracers in homogeneous, isotropic turbulence. The experimental data are obtained from time-resolved Shake-The-Box particle tracking in the wake of a square prism. Additionally, we provide a brief description of the turbulent duct flow dataset from the second Data Assimilation Challenge used to benchmark the performance of PIGLET against other algorithms.

2.4.1. Homogeneous isotropic turbulence

We performed direct numerical simulations of homogeneous, isotropic turbulence ($R_\lambda \approx 122$) in a 512^3 periodic box of side length 2π using the TuRTLE pseudo-spectral solver (Lalescu et al., 2022). The incompressible Navier-Stokes equations were solved in vorticity form with Lundgren forcing applied in the $[2, 4]$ wavenumber range. The resolution $k_{max}\eta = 1.87$ ensures well-resolved small scales. We allow a statistically stationary state to develop after running the simulation for 23.4 integral timescales. However, this forcing adds a stochastic term to the Lagrangian acceleration which cannot be predicted by the Navier-Stokes dynamics. Therefore, we turn off the forcing at this point and integrate the trajectories of 2^{22} tracer particles using a fourth-order Adams-Bashforth method and cubic spline velocity interpolation for a further 52 timesteps ($1.0\tau_\eta$). Velocity data sampled on tracer trajectories are used to generate a synthetic four-pulse LPT measurement using a 2 : 1 : 2 timing scheme with $\Delta t = 0.15\tau_\eta$ between the second and third pulses. Additive white Gaussian noise is added to these trajectories to model position uncertainty. Finally, a second order polynomial fit is applied to each trajectory to measure the (noisy) tracer position, velocity and acceleration.

The snapshot at $23.4T_{int}$ was partitioned into 64 sub-volumes of side-length $\pi/2 \approx 214\eta \approx 2.8L_{int}$. To investigate seeding effects, data was downsampled to 4096, 16384, or 65536 particles per sub-volume, corresponding to a inter-particle separation $\ell_p = \rho_p^{-1/3} = 13.4, 8.4$, and 5.3η where ρ_p is the number density of particles. With an LPT timestep of $\Delta t = 0.15\tau_\eta$, the mean square displacement is 0.85η ($\approx 0.4\%$ of the domain). We introduced additive measurement noise at $\sigma_u/u' = 0.01$ to simulate moderate measurement noise. This data was used to train the covariance model (13) for 257 coefficients across wavenumbers $\kappa_n = 2\pi/\lambda_n = 0, 1, \dots 256$.

2.4.2. Square Prism Wake Flow

We measured the wake of a $H = 50\text{mm}$ square prism at a freestream velocity of $U_\infty = 0.150 \text{ m/s}$ corresponding to $\text{Re}_H \approx 7850$ in the University of Southampton's Recirculating Water Tunnel. The experimental configuration is shown in Figure 1. This flow lacks the fundamental translation symmetry of homogeneous isotropic turbulence, making it a useful test case for the model.

We recorded $55\mu\text{m}$ polyamide tracers (Vestosint 2517) using a four-camera LaVision Minishaker

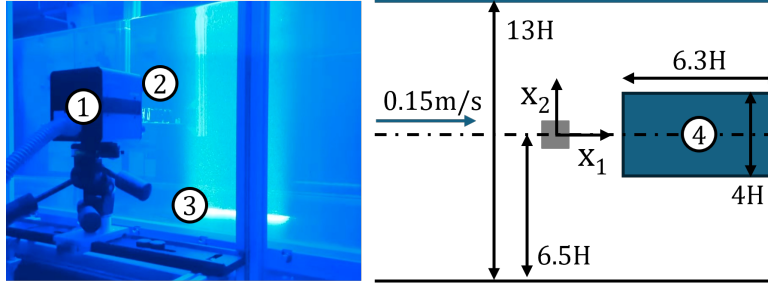


Figure 1. Photograph (left) and sketch (right) the experimental setup, showing (1) Minishaker system, (2) square prism, (3) LED illumination and (4) the particle tracking measurement volume within the Recirculating Water Tunnel. The coordinate system origin is at the centre of the prism at the mid-span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

2M PTV system and LED-pulse illumination at 121.2 Hz. We recorded 2000 bursts of 25 frames each at 3.5s intervals, covering a downstream region 105 to 445 mm from the prism at a resolution of 0.165 mm/vx with a seeding concentration of 0.0056 ppp. For context, the estimated vortex shedding period is ~ 2.4 s. The images have a resolution of 1984×1264 px. The third-order polynomial geometric calibration was refined using volume self-calibration to a 0.03 voxel disparity error. Particle trajectories were reconstructed in DaViS 10 using the Shake-The-Box algorithm.

The model was trained and evaluated on 120 snapshots of four-pulse trajectories extracted from a truncated measurement volume ($316 \times 197 \times 79$ mm). Approximately 10,000 particles were detected for all four timesteps within this volume. To assess seeding density effects, snapshots were downsampled to approximately 1024, 5055, or 9215 particles per snapshot by partitioning the data into training and cross-validation sets. This corresponds to a typical particle separation $\ell_p = \rho_p^{-1/3}$ between 8.1 and 16.9mm. We trained and evaluated the homogeneous covariance model described in §2.2 using $76 \times 120 \times 30$ Fourier modes, corresponding to a periodic domain size which is twice the size of the measurement volume. The spatial resolution of the model is 5.27mm, corresponding to a maximum resolvable wavenumber of 0.60mm^{-1} . This is sufficient to resolve the average Nyquist rate $\pi/\ell_p = 0.39\text{mm}^{-1}$ in the highest seeding concentration case.

2.4.3. Data Assimilation Challenge: Turbulent Duct Flow

We consider the turbulent duct flow test case from the 2nd Data Assimilation Challenge to benchmark PIGLET against other algorithms. The case simulates a synthetic experiment derived from a scaled direct numerical simulation of the incompressible turbulence in a square duct at a friction Reynolds number $Re_\tau \approx 1000$ (Pirozzoli et al., 2018). The flow is aligned with the x_1 direction with no-slip walls at $x_2 = 0$ and $2h$ and $x_3 = 0$ and $2h$, where $h = 0.1\text{m}$ is the half-height of the duct. The bulk flow velocity U_b is 0.2m/s. The synthetic experiment provides the positions of tracers in a cubic box spanning $[0 : h] \times [0 : h] \times [0 : h]$. There are between 1000 and 200,000 tracers,

corresponding to seeding concentrations between 0.001 and 0.200 ppp. The highest seeding concentration represents a significant challenge for exact Gaussian process regression as we have used previously (Lawson, 2026), since the kernel matrix alone would require 2.88TB of memory to store. Therefore, solution using the Bayesian formulation given in (8) or (9) is essential.

We consider two test cases: two-pulse (TP) and four-pulse (FP). In the TP case, two snapshots of the tracer position are provided with a time separation of $\Delta t = 5.916\text{ms}$. In the FP case, four snapshots of the tracer position are available using a 2-1-2 timing scheme with the same separation between second and third pulses. We compute velocity and acceleration from particle tracks using a least-squares regression to a second order polynomial.

We use a homogeneous, isotropic Gaussian process model using $192 \times 384 \times 384$ Fourier modes on a $2h \times 4h \times 4h$ domain. This configuration expands the original measurement domain to account for the effects of periodicity and to impose the solid wall boundary condition. This is achieved by constraining the model to use only odd Fourier modes in the x_2 and x_3 directions, which creates antisymmetric mirror image of the flow in these planes. This antisymmetric configuration also has the benefit of reducing interpolation errors due to the Gibbs phenomenon at the walls, where the mean flow gradient is very steep. Since insufficient training data are available to estimate the mean flow, we initialise the mean flow from that provided publicly by Pirozzoli et al. (2018). In principle, with more training data, it would be possible to initialise an accurate estimate of the mean flow from (19) or through variational DA methods (Cadambi Padmanaban et al., 2026). Then, we learn the isotropic autocorrelation function (16) and noise from a single snapshot for each case. This is possible because only 128 hyperparameters are used to represent the autocorrelation function. However, it comes at the expense of reducing the complex correlation structure of this flow to an extremely simple isotropic model.

3. Results

3.1. Homogeneous Isotropic Turbulence

Figure 2a shows the longitudinal autocorrelation function (13) obtained from the maximum likelihood product-of-experts fit to sparse and noisy data (17). This case corresponds to a moderate seeding concentration $\ell_p = 8.4\eta$ and measurement noise $\sigma_u = 0.01u'$. The maximum likelihood fit (dashed line) is in good agreement with the ground truth (solid line) to within the statistical uncertainty (grey area). The agreement is worse at large scales, which have greater statistical uncertainty since fewer particle pairs are observed at large separations. The model estimates of the noise $\sigma_u = 0.0097u'$ and dissipation rate $\epsilon/\epsilon_{ref} = 0.991$ are very close to the ground truth. The corresponding 1D and 3D energy spectra shown in Figure 2b are in excellent agreement with those obtained directly from the reference DNS data.

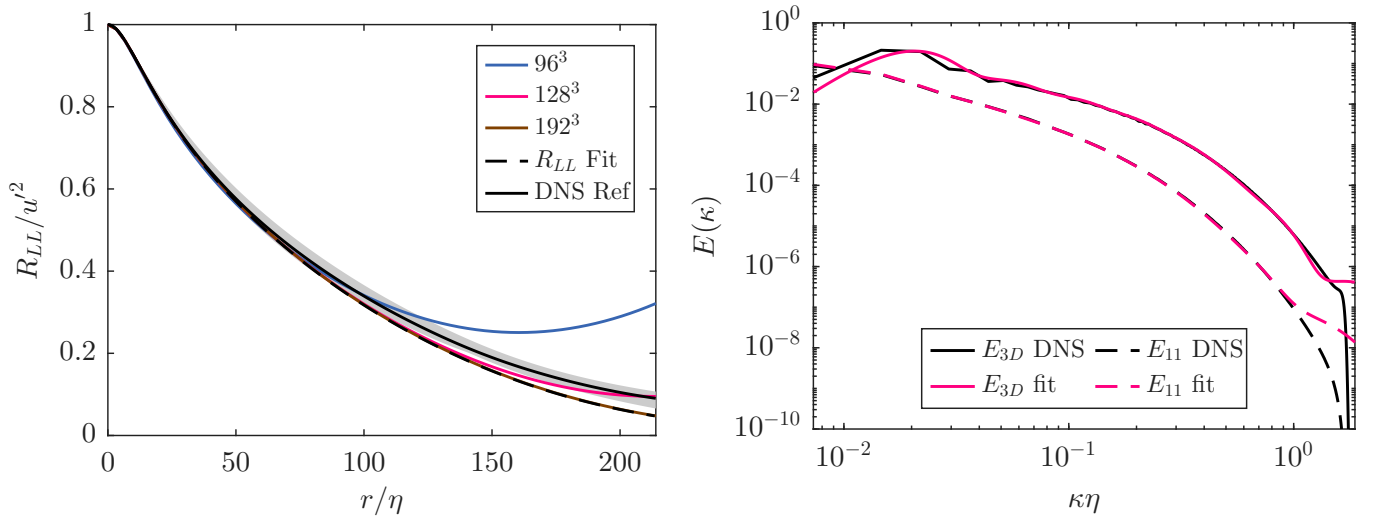


Figure 2. (a) Comparison of the trained model of the longitudinal autocorrelation function (16) and its Fourier series approximation (11) to ground truth in HIT for $\sigma_u = 0.01u'$ and $\ell_p = 8.4\eta$. The shaded region shows confidence interval for ground truth, based on the variation of R_{LL} in each direction. (b) Corresponding 3D and 1D energy spectra (16).

Coloured lines in figure 2a show the Fourier series approximation (11) with $M = 96^3$, 128^3 or 192^3 Fourier modes. The spatial resolution is fixed at $\Delta x = 3.34\eta$ and therefore this discretisation corresponds to a periodic flow field which is 1.5, two or three times larger than the original measurement domain (a cube of side length $L = \pi/2$). For the smallest domain size (96^3 modes, $1.5L$ period), the effect of this implicit periodicity on the approximation (11) is significant. The approximation improves at larger grid sizes. In the following, we use a 128^3 grid corresponding to a periodic domain of side length $2L$ and maximum resolvable wavenumber $\kappa_{max}\eta = 0.94$.

Figure 3a shows an example of the velocity field inferred using PIGLET with $M = 128^3$ Fourier modes at the lowest seeding concentration tested (4000 particles, $\ell_p = 13.4\eta$). Considering the sparse seeding concentration, there is good agreement between the reconstruction and ground truth. The error is least near particle locations and greatest in the regions between them. Figure 3b shows an example of three realisations of the same velocity field drawn from the posterior distribution (7). These realisations show that the uncertainty in the velocity field is spatially correlated: the velocity realisations are always spatially smooth. An estimate of the local uncertainty is also shown, based on the variance over 30 realisations. The uncertainty is least near particle locations and greatest in the regions between them. Note that regions of high uncertainty do not necessarily correspond to high local error. Regions of low uncertainty are localised to individual particles and correspond to regions of low error in the reconstruction.

Figure 4a compares the mean square velocity error of exact inference using (5) and approximate inference of the conditional mean using the linear model (8) after 2000 iterations. The results for the exact and linear model are in good agreement, indicating that the solution has converged to near

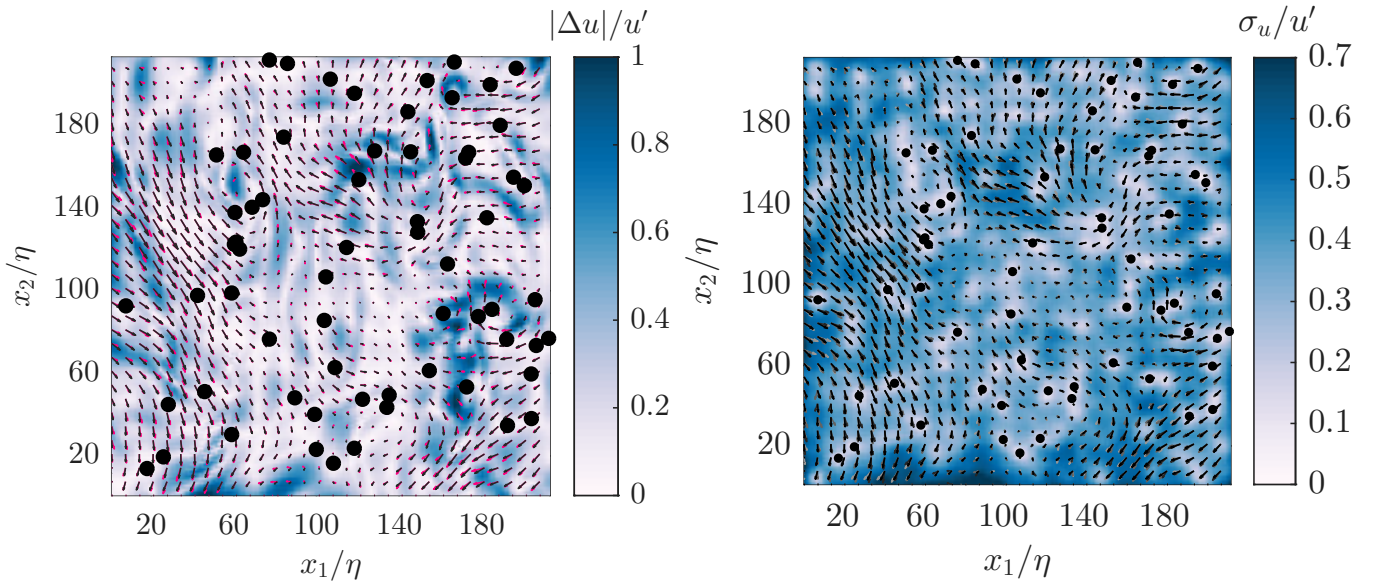


Figure 3. (a) Comparison of ground truth (pink vectors) and reconstructed (black vectors) velocity field based on 4000 particles (black circles) with moderate noise ($\ell_p = 13.6\eta$, $\sigma_u = 0.01u'$). (b) Corresponding prediction of uncertainty in the reconstruction. Black vectors show the most probable flow field (5) whereas grey vectors show three realisations flow field drawn from the posterior distribution.

the optimum. When (5) is used as the initial condition for maximising (8), the solver terminates immediately as it has reached the optimum.

To test the local estimate of uncertainty, we define a standardised z-score for the error $z = \Delta u/\sigma$, where Δu is the measurement error in a velocity component and σ is the local uncertainty estimated by drawing 30 samples from the posterior distribution (7). 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 4b shows the cumulative distribution function (CDF) of z for a noise level of $\sigma_u = 0.01u'$. We observe that the CDF is well approximated by the standard normal distribution and yields a good prediction of the true 95.45% confidence interval; the CDF shows that a $\pm 2.3\sigma$ interval provides this coverage.

Figure 4a shows that, surprisingly, the addition of acceleration information into the regression model (9) does not significantly improve the accuracy of the velocity or vorticity field reconstruction. This result is confirmed in figure 5a, which demonstrates that the spectral coherence between the ground truth and reconstruction is not significantly improved at small scales. Instead, the cutoff wavenumber remains around $\kappa_{max} \approx \pi/\ell_p$, which corresponds to the average Nyquist rate (Lawson, 2026). To examine why, we present the mean square error in the predicted acceleration field, as sampled on tracer positions in figure 5. We observe that, at low seeding concentration, the linear model alone makes a poor prediction of the true acceleration field. This error can be reduced to or below the noise floor of the measurement by including acceleration data. The benefit is most pronounced at low seeding concentrations, where the model has the most to learn by incor-

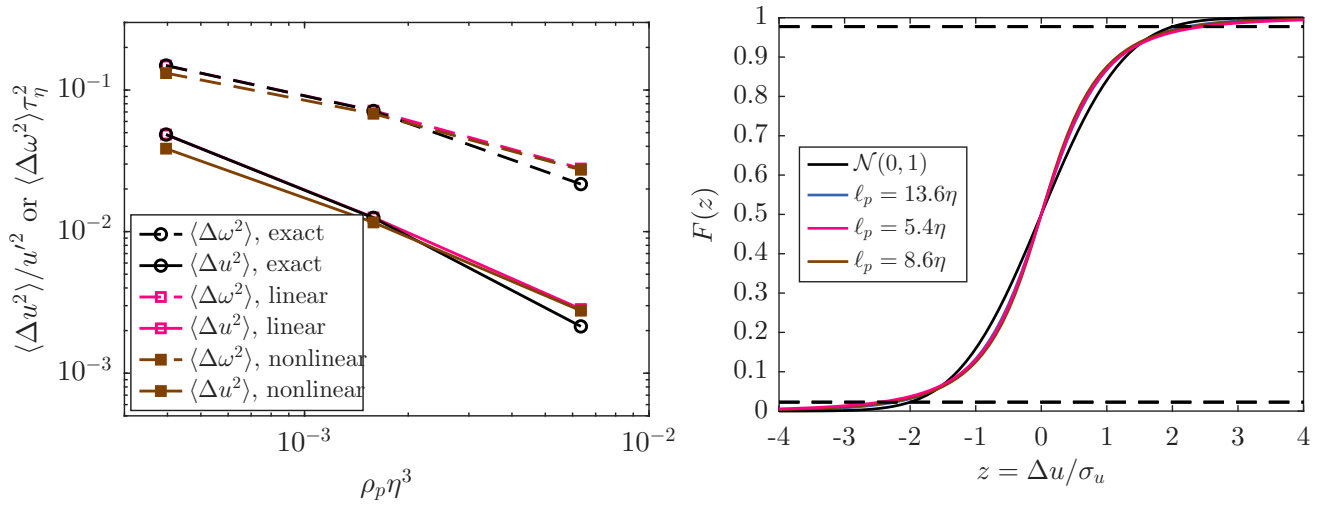


Figure 4. (a) Comparison of component-wise mean square velocity and vorticity error, as a function of seeding concentration, for (○) exact calculation using (5) and approximate calculation using (□) only velocity (8) and (■) both velocity and acceleration data (9) for 2000 iterations. (b) CDF of the z-score of the velocity error $z = \Delta u / \sigma_u$, compared to unit normal distribution.

porating the acceleration information. At the highest seeding concentration tested, the error in the acceleration prediction from the linear model is already smaller than the measurement uncertainty. As a result, little benefit can be achieved by incorporating the acceleration into the regression.

However, figure 5 also shows that there is a significant improvement in the accuracy of the pressure field reconstruction: even using noisy acceleration data can reduce the mean square error in pressure by up to a factor of six. This is unexpected, because the velocity fields predicted by both models are very similar inside the measurement domain. To explain why this happens, figure 6 shows the full extent of the periodic velocity field predicted by the linear and non-linear models. Although the velocity field within the region supported by the measurements is very similar, the velocity field outside this region differs significantly. This domain expansion effectively allows the model to optimise large-scale flow features to better fit the observed acceleration data, capturing large-scale effects in the pressure field without affecting the quality of the velocity fit.

3.2. 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 §2.4. This dataset is particularly noisy and exhibits strong inhomogeneities in seeding concentration throughout the measurement volume. Furthermore, 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 PIGLET, since the model we have used assumes statistical homogeneity.

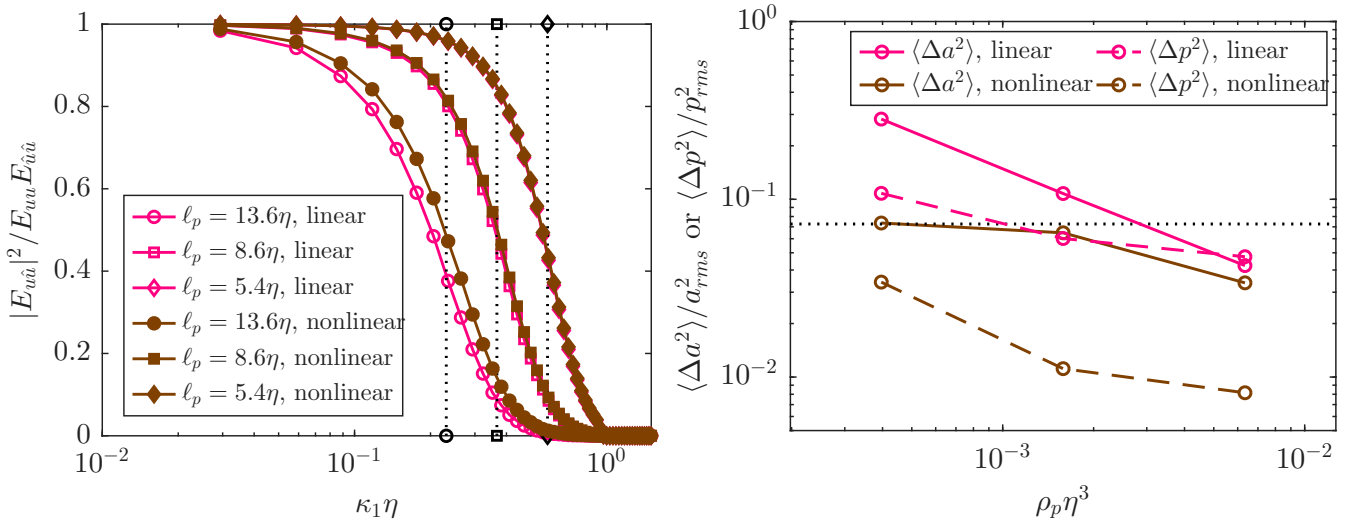


Figure 5. (a) Spectral coherence of velocity fluctuations with ground truth, estimated using (pink) velocity data only or (brown) velocity and acceleration data. Dashed lines show cutoff wavenumber $\kappa_{max} = \pi/\ell_p$. (b) Mean square pressure error and mean square acceleration error (per component, sampled on tracers) as a function of seeding concentration. The dotted line shows the mean square acceleration noise.

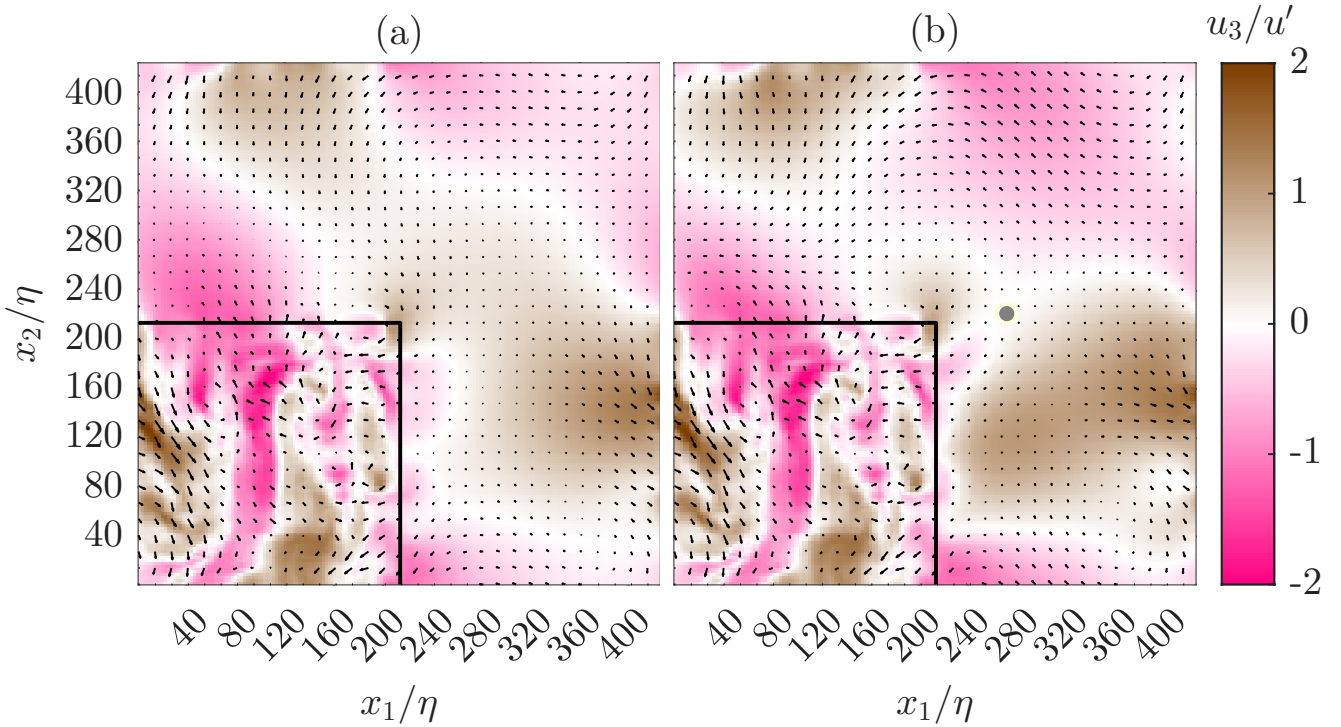


Figure 6. Comparison of velocity field extrapolated beyond the measurement domain boundary (a) with linear regression model only (8) and (b) with nonlinear model (9) for $\ell_p = 5.4\eta$ and $\sigma_u = 0.01u'$

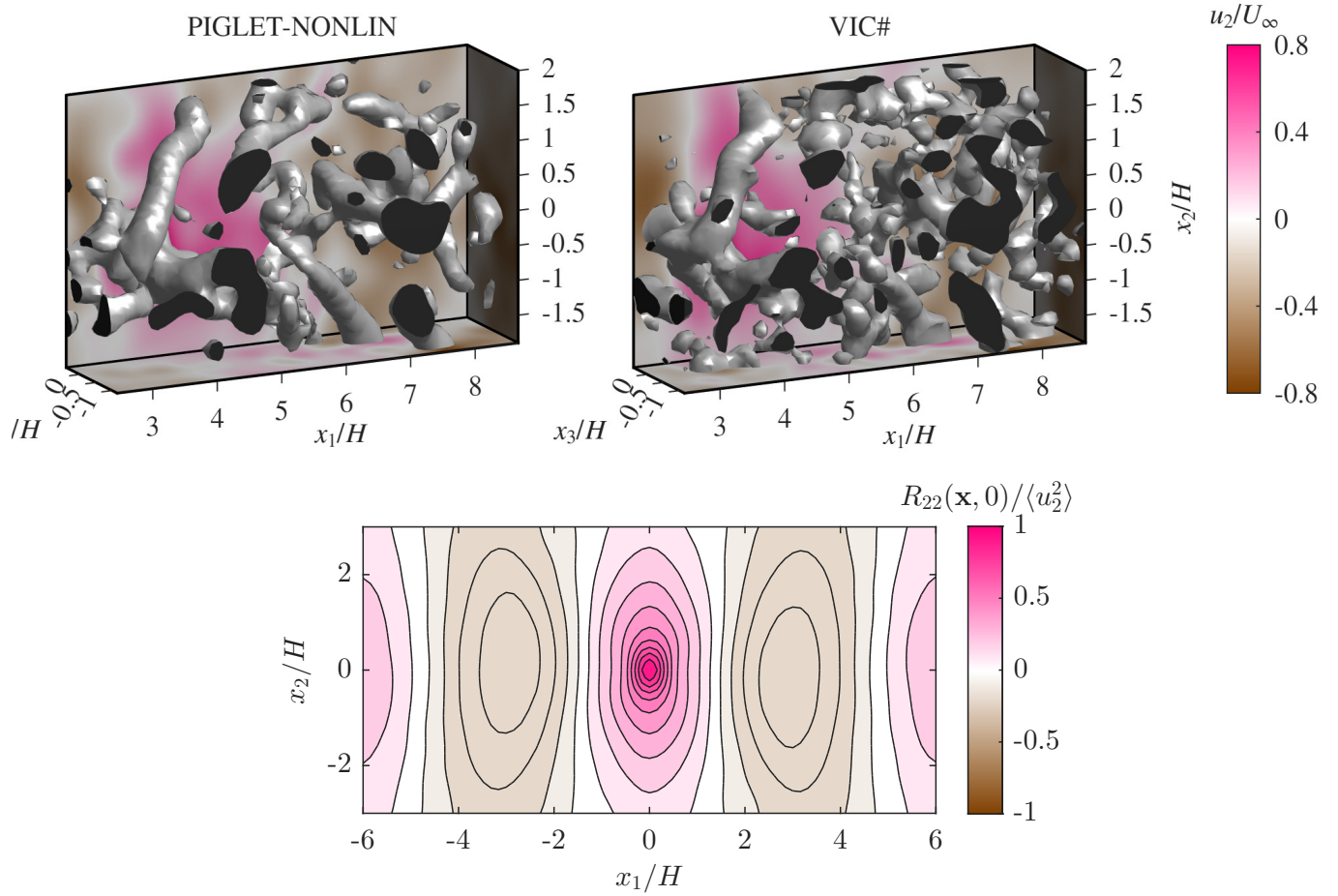


Figure 7. (a) Reconstruction of the flow field based on 1053 particles in the wake of a square prism using the non-linear PIGLET regression (left) and VIC# (right). Isosurfaces show contours of Q-criterion $Q = 0.5U_0^2/H^2$. The colourmap shows the transverse velocity component. (b) Homogeneous model of the spatial autocorrelation of the transverse velocity component learned by PIGLET for this case.

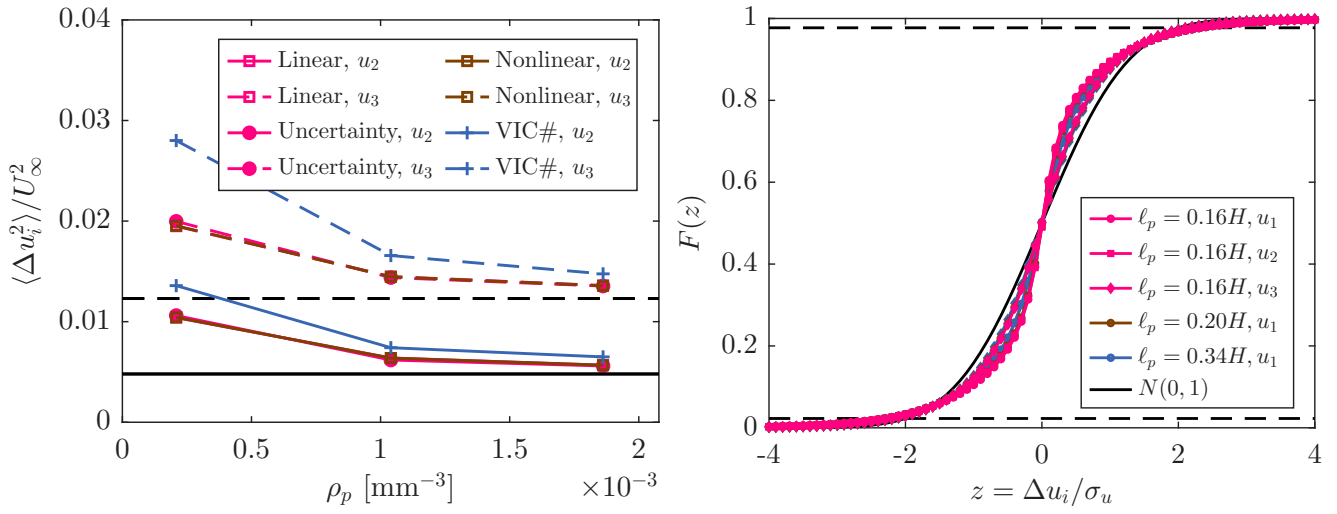


Figure 8. Cross-validation of velocity reconstructions from PIGLET and VIC#. (a) Mean-square cross-validation error for different effective seeding concentration. Solid lines show the transverse component, whereas dashed lines show spanwise component. The streamwise component (not shown) is very similar to the transverse component. (b) CDF of the cross-validation z-score $z = \Delta u / \sigma$.

Figure 7a shows an example of the reconstruction of the transverse velocity component and iso-surfaces of Q-criterion obtained by PIGLET and VIC#. We choose to present the lowest seeding concentration case to demonstrate the performance of both algorithms under the most challenging conditions. Although both reconstructions are similar, there is noticeably more incoherent motion present in the VIC# reconstruction. This is because PIGLET learns the underlying correlation structure of the data. Figure 7b shows the spatial autocorrelation function of the transverse velocity component learned by PIGLET for this case. Upward velocity fluctuations are correlated strongly in the transverse direction, and are anti-correlated with downward velocity fluctuations downstream and upstream of the measurement point. Because PIGLET can learn that velocity fluctuations are correlated over large length scales, the output is more non-local than VIC#, where the locality of the reconstruction is determined by the scale of the radial basis functions used.

To assess the accuracy of velocity prediction we cross-validate velocity field reconstructions by withholding a fraction of the tracks (10, 50 or 90%) used to train and predict velocity fields. These data are then compared to the model's prediction. Figure 8a shows the cross-validation error of the model trained after five epochs. This dataset is particularly noisy: the black lines show the model's estimate of the velocity noise, which is a lower bound for cross-validation error. The model learns that the error is largest for the spanwise component, which corresponds to the component of velocity measured mostly normal to the camera. This is further evidenced by the fact that doubling the effective seeding concentration does not significantly increase the cross-validation error.

Several comparisons are evident in figure 8a. First, the mean-square cross-validation error is well predicted by the mean uncertainty of the regression. Second, the accuracy of reconstruction is not

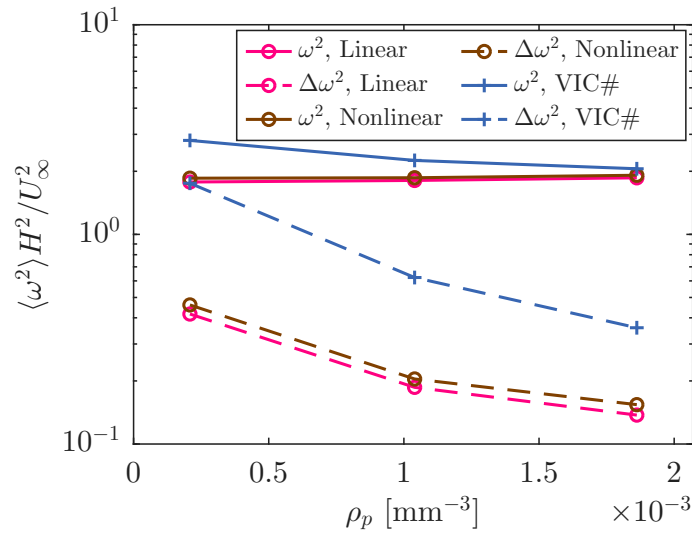


Figure 9. Temporal coherence of vorticity field obtained from PIGLET and VIC#.

significantly improved by the inclusion of acceleration data into the non-linear regression. This is because the acceleration signal is too noisy to add meaningful information; we estimate the signal-to-noise ratio is less than one. Third, to provide a comparison against the industry standard, the same statistics are computed based on VIC# reconstruction. We observe that, in all cases, Gaussian process regression using PIGLET leads to a lower cross-validation error than VIC#.

As a more detailed test of the model's uncertainty prediction, figure 8b shows the CDF of the cross-validation z-score. Some deviations from the standard normal distribution are evident, which are most pronounced for the streamwise and transverse velocity components. However, the modelled uncertainty is generally a good predictor of the true error. Over all velocity components and effective seeding concentrations tested, at least 93.7% of cross-validation data lie within $|z| < 2$.

To quantify the temporal coherence of the fine-scale motion, we performed velocity field reconstructions at two times, 16.5ms apart, using consecutive frames. Figure 9 shows mean square difference in the vorticity field $\Delta\omega(\mathbf{x}, t) = \omega(\mathbf{x}, t + \Delta t) - \omega(\mathbf{x}, t)$ over this interval. We do not expect the vorticity to change significantly over this time, since it is just 0.7% of the vortex shedding period. We observe that the PIGLET reconstruction of the vorticity field is substantially more coherent in time than VIC#. When using PIGLET, the discrepancy between vorticity fields is 2.6 – 4.2 times smaller in absolute terms in comparison to VIC#. The inclusion of the acceleration data marginally worsens the temporal coherence: this is expected, given that the data is very noisy. This coherence does not come at the expense of filtering fine-scale motion: solid lines in figure 9 shows only a marginal increase in mean-square enstrophy as the seeding concentration is increased. This indicates that the reconstruction at the lowest seeding concentration captures most of the small-scale motions present. With VIC#, the enstrophy decreases as the seeding concentration increases, indicating that significant noise in the vorticity field is present at low seeding concentration.

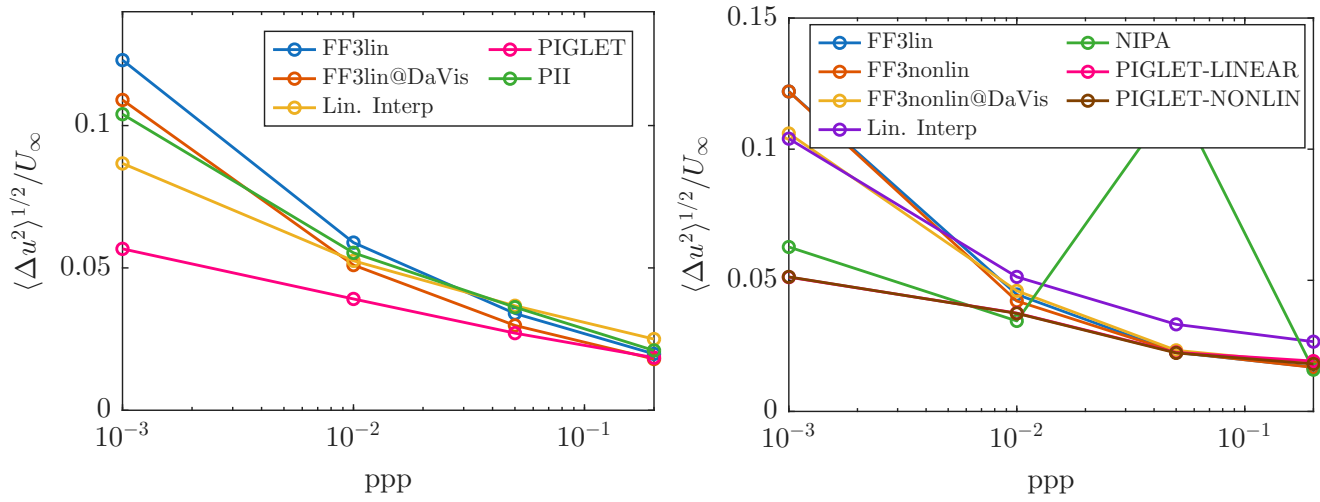


Figure 10. Performance of PIGLET in the 2nd Data Assimilation challenge. RMS velocity prediction error is shown as a function of the seeding concentration. (a) two-pulse case and (b) four-pulse case. The other algorithms are FlowFit 3 (Godbersen et al., 2024) as implemented by DLR and LaVision, Physics Informed Interpolation (Sperotto et al., 2024), Neural-Implicit Particle Advection (Zhou & Grauer, 2023) and linear interpolation.

3.3. Data Assimilation Challenge

We compare the performance of PIGLET against other entries to Case 1 of the second data assimilation challenge in Figure 10. We consider the two pulse and four pulse data assimilation cases only, since these correspond to the type of data that PIGLET is able to assimilate. In all cases, PIGLET outperforms or performs comparably to state-of-the-art algorithms in terms of velocity prediction error. Notably, the incorporation of acceleration data in the four-pulse case does not meaningfully improve prediction accuracy. The evidence is twofold. First, the smallest achievable velocity interpolation error is comparable between two-pulse and four-pulse cases. Second, both PIGLET and Flow Fit 3 (Godbersen et al., 2024) have linear and non-linear variants which incorporate acceleration data. Both variants perform similarly.

4. Conclusion

We have introduced PIGLET, an open-source framework utilising physics-informed Gaussian process regression to reconstruct turbulent flow fields from LPT and PTV data. This paper extends our previous work (Lawson, 2026) by formulating a finite-dimensional Gaussian process model for homogeneous turbulent flows, which can incorporate Navier-Stokes dynamics in addition to statistical symmetries, incompressibility and boundary conditions. By formulating the GPR problem as a Bayesian inference problem, we mitigate the conventional computational bottleneck, permitting the assimilation of $\mathcal{O}(100,000)$ particle tracks. Another significant advantage of this formulation is that we can use the Matheron update rule to efficiently draw samples of the most likely flow fields

which are consistent with the data, which can be used to quantify uncertainty in post-processed outputs.

We have tested PIGLET with synthetic data from homogeneous isotropic turbulence and experimental data from a square prism wake. Tests in HIT show that the two-point correlation function and 3D energy spectrum can be recovered accurately even from sparse and noisy particle tracks. Tests in the square prism wake show that PIGLET can learn the spatial coherence of characteristic von-Kármán vortex shedding in the wake of the prism. In both cases, sample estimates of measurement uncertainty are in excellent agreement with the error measured by cross-validation and comparisons against ground truth. We also find that PIGLET out-performs VIC# in terms of velocity cross-validation error and the temporal coherence of the vorticity field. We have benchmarked the performance of PIGLET against other state-of-the-art algorithms in the 2nd Data Assimilation Challenge. PIGLET outperforms or performs comparably to state-of-the-art algorithms when assimilating velocity fields from two-pulse or four-pulse PTV data.

We have investigated the incorporation of Navier-Stokes dynamics using material acceleration data. Interestingly, incorporating acceleration information does not significantly improve the accuracy of velocity or vorticity predictions. This finding is consistent with other algorithms submitted to the 2nd Data Assimilation Challenge. However, it substantially reduces the mean square error of the predicted pressure field. For noisy experimental data, the low signal-to-noise ratio of acceleration data currently limits its utility.

Data availability

PIGLET is open source. The code and datasets used to create this manuscript are available to download from <https://github.com/jml70/PIGLET>.

Acknowledgements

This work was supported by EPSRC grant number EP/Y036409/1. We 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.

References

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer New York.

- Brunton, S. L., Noack, B. R., & Koumoutsakos, P. (2019). Machine learning for fluid mechanics. *Annu. Rev. Fluid Mech.* 2020, 52, 477-508. Retrieved from <https://doi.org/10.1146/annurev-fluid-010719> doi: 10.1146/annurev-fluid-010719
- Cadambi Padmanaban, U., Ganapathisubramani, B., & Symon, S. (2026, Jan). Three-dimensional variational data assimilation of separated flows using time-averaged experimental data. *Phys. Rev. Fluids*, 11, 014904. Retrieved from <https://link.aps.org/doi/10.1103/q7sd-q3dn> doi: 10.1103/q7sd-q3dn
- Cakir, B. O., Saiz, G. G., Sciacchitano, A., & van Oudheusden, B. (2022, 12). Dense interpolations of lpt data in the presence of generic solid objects. *Measurement Science and Technology*, 33. doi: 10.1088/1361-6501/ac8ec7
- Casa, L. D., & Krueger, P. S. (2013). Radial basis function interpolation of unstructured, three-dimensional, volumetric particle tracking velocimetry data. *Measurement Science and Technology*, 24. doi: 10.1088/0957-0233/24/6/065304
- Gesemann, S., Huhn, F., Schanz, D., & Schröder, A. (2016). From noisy particle tracks to velocity, acceleration and pressure fields using b-splines and penalties. In (p. 4-7).
- Godbersen, P., Gesemann, S., Schanz, D., & Schröder, A. (2024). Flowfit3: Efficient data assimilation of lpt measurements. In *21th international symposium on application of laser and imaging techniques to fluid mechanics*.
- Jeon, Y. J. ., Schneiders, J. F. G. ., Müller, M. ., Michaelis, D. ., Wieneke, B., Jeon, Y. J., ... Michaelis, D. (2018). 4d flow field reconstruction from particle tracks by vic+ with additional constraints and multigrid approximation. In *18th international symposium on flow visualization*. Retrieved from <https://doi.org/10.3929/ethz-b-000279199> doi: 10.3929/ethz-b-000279199
- Lalescu, C. C., Bramas, B., Rampp, M., & Wilczek, M. (2022, 9). An efficient particle tracking algorithm for large-scale parallel pseudo-spectral simulations of turbulence. *Computer Physics Communications*, 278. doi: 10.1016/j.cpc.2022.108406
- Lawson, J. M. (2026, Jan). Physics informed gaussian process regression for particle tracking data assimilation. *Phys. Rev. Fluids*, -. Retrieved from <https://link.aps.org/doi/10.1103/zvm4-wtkq> doi: 10.1103/zvm4-wtkq
- Mons, V., Marquet, O., Leclaire, B., Cornic, P., & Champagnat, F. (2022, 12). Dense velocity, pressure and eulerian acceleration fields from single-instant scattered velocities through navier-stokes-based data assimilation. *Measurement Science and Technology*, 33. doi: 10.1088/1361-6501/ac8dac

- Pirozzoli, S., Modesti, D., Orlandi, P., & Grasso, F. (2018). Turbulence and secondary motions in square duct flow. *Journal of Fluid Mechanics*, 840, 631–655. doi: 10.1017/jfm.2018.66
- Pope, S. B. (2000). *Turbulent flows*. Cambridge University Press. doi: 10.1088/1468-5248/1/1/702
- Ratz, M., & Mendez, M. A. (2024, 9). A meshless and binless approach to compute statistics in 3d ensemble ptv. *Experiments in Fluids*, 65. doi: 10.1007/s00348-024-03878-x
- Schneiders, J. F., & Scarano, F. (2016, 9). Dense velocity reconstruction from tomographic ptv with material derivatives. *Experiments in Fluids*, 57. doi: 10.1007/s00348-016-2225-6
- Schröder, A., & Schanz, D. (2023). Annual review of fluid mechanics 3d lagrangian particle tracking in fluid mechanics. *Annu. Rev. Fluid Mech.* 2023, 55, 511-540. Retrieved from <https://doi.org/10.1146/annurev-fluid-031822> doi: 10.1146/annurev-fluid-031822
- Sciacchitano, A., Leclaire, B., & Schröder, A. (2021, 8). Main results of the first data assimilation challenge. In *14th international symposium on particle image velocimetry – ispio2021*.
- Shin, H., & Schröder, A. (2025). *Robust flow field reconstruction using pinn for 3d lagrangian particle tracking*.
- Sperotto, P., Pieraccini, S., & Mendez, M. A. (2022, 9). A meshless method to compute pressure fields from image velocimetry. *Measurement Science and Technology*, 33. doi: 10.1088/1361-6501/ac70a9
- Sperotto, P., Watz, B., & Hess, D. (2024, 8). Meshless track assimilation (mta) of 3d ptv data. *Measurement Science and Technology*, 35. doi: 10.1088/1361-6501/ad3f36
- Steinfurth, B., Hassanein, A., Doan, N. A. K., & Scarano, F. (2024, 09). Physics-informed neural networks for dense reconstruction of vortex rings from particle tracking velocimetry. *Physics of Fluids*, 36(9), 095110. Retrieved from <https://doi.org/10.1063/5.0212585> doi: 10.1063/5.0212585
- Wilson, J. T., Borovitskiy, V., Terenin, A., Mostowsky, P., & Deisenroth, M. P. (2020, 11). Path-wise conditioning of gaussian processes. *Journal of Machine Learning Research*. Retrieved from <http://arxiv.org/abs/2011.04026>
- Zhou, K., & Grauer, S. J. (2023). Flow reconstruction and particle characterization from inertial lagrangian tracks. *arXiv preprint arXiv:2311.09076*.