







RESEARCH ARTICLE

# Drag prediction of rough-wall turbulent flow using data-driven regression

Zhaoyu Shi<sup>1</sup> , Seyed Morteza Habibi Khorasani<sup>1</sup> , Heesoo Shin<sup>2</sup> , Jiasheng Yang<sup>3</sup> , Sangseung Lee<sup>2</sup>  and Shervin Bagheri<sup>1,\*</sup> 

<sup>1</sup>FLOW, Department of Engineering Mechanics, KTH, Stockholm 10044, Sweden

<sup>2</sup>Mechanical Engineering Department, Inha University, Incheon 22212, Republic of Korea

<sup>3</sup>Institute of Fluid Mechanics, Karlsruhe Institute of Technology, Karlsruhe 76131, Germany

\*Corresponding author. E-mail: [shervin@mech.kth.se](mailto:shervin@mech.kth.se)

**Received:** 13 May 2024; **Revised:** 13 September 2024; **Accepted:** 30 October 2024

**Keywords:** Roughness; Drag; Machine learning; Ship hull and aerodynamic design; Drag reduction

## Abstract

Efficient tools for predicting the drag of rough walls in turbulent flows would have a tremendous impact. However, accurate methods for drag prediction rely on experiments or numerical simulations which are costly and time consuming. Data-driven regression methods have the potential to provide a prediction that is accurate and fast. We assess the performance and limitations of linear regression, kernel methods and neural networks for drag prediction using a database of 1000 homogeneous rough surfaces. Model performance is evaluated using the roughness function obtained at a friction Reynolds number  $Re_\tau$  of 500. With two trainable parameters, the kernel method can fully account for nonlinear relations between the roughness function  $\Delta U^+$  and surface statistics (roughness height, effective slope, skewness, etc.). In contrast, linear regression cannot account for nonlinear correlations and displays large errors and high uncertainty. Multilayer perceptron and convolutional neural networks demonstrate performance on par with the kernel method but have orders of magnitude more trainable parameters. For the current database size, the networks' capacity cannot be fully exploited, resulting in reduced generalizability and reliability. Our study provides insight into the appropriateness of different regression models for drag prediction. We also discuss the remaining steps before data-driven methods emerge as useful tools in applications.

## Impact Statement

The accurate estimation of drag in aviation and shipping is of great economic value as it significantly affects energy expenditure and carbon emissions. The long-standing pursuit of a universal correlation between drag and topographical features of roughness has made remarkable progress in recent decades, yet it is still limited by the feasibility of demanding experiments and simulations. There exists no model which is generally applicable to any given rough surface. This positions machine-learning (ML) modelling as a promising cost-effective approach. Therefore, this study seeks to provide more insights into how different ML regression models perform in terms of the trade-off between capturing nonlinearity and training costs. The comprehensive analysis presented in this work aims to offer valuable insights for the future design of ML-based models in the field of drag prediction.

## 1. Introduction

Three-dimensional multi-scale surface irregularities are ubiquitous in industrial applications. The roughness imposes an increased resistance upon an overlying fluid flow, manifested as an increase in the measured drag. The increase in drag causes reduced energy efficiency, especially in turbulent flows. Examples include increased fuel consumption of cargo ships due to fouled hulls, reduced power output of eroded turbines in wind power plants and an increase in the power input required to maintain a constant flow rate in pipelines with non-smooth walls.

An efficient tool that can predict the drag induced by roughness would allow engineers and operators to optimize surface cleaning and treatment. However, as of today, there is no method for drag prediction that is both fast and accurate (Yang *et al.* 2023b). There are accurate techniques that rely on towing tank experiments (Schultz 2004), direct numerical simulations (Thakkar, Busse & Sandham 2016; Forooghi *et al.* 2017; Thakkar, Busse & Sandham 2018) or large eddy simulations (Chung & Mckeen 2010). These methods measure the equivalent sand grain roughness  $k_s$  that can be used to estimate the drag penalty for simple geometries (e.g. pipes) or incorporated into computational fluid dynamics software to evaluate the drag penalty on complex bodies (Andersson *et al.* 2020; De Marchis *et al.* 2020; Kadivar, Tormey & McGranaghan 2023). However, towing tank experiments and computational techniques can be costly and time-consuming. The alternative approach is rough-wall modelling, where the objective is to predict  $k_s$  directly from the roughness topology. As discussed by Yang *et al.* (2023b), models can be divided into correlation-type (Flack & Schultz 2014; Chan *et al.* 2015; Forooghi *et al.* 2017), physics-based (Yang *et al.* 2016) or data-driven regression methods (Jouybari *et al.* 2021; Lee *et al.* 2022; Yang *et al.* 2023a,b; Shin *et al.* 2024; Yang *et al.* 2024). The accuracy, generalizability and complexity of the rough-wall models increase going from correlation-type to physics-based and finally to data-driven regression methods. In particular, while data-driven regression techniques have gained in popularity and show promise, they suffer primarily from lack of data that can be used for training.

Over time, a sufficient amount of relevant roughness data will be accumulated, which can be used to develop efficient regression models for predicting the drag of rough surfaces. Regression models can directly process images or topographical maps of the roughness to predict  $k_s$ , thus replacing experiments and resolved simulations. Recent efforts have focused on relatively complex regression methods. Jouybari *et al.* (2021) adopted a multi-layer perceptron (MLP) and Gaussian processes regression to build a mapping from statistical surface measures to  $k_s$ . Both methods were trained on 45 labelled samples, achieving an accuracy of approximately 10%. Realizing that the database size is the major bottleneck for fully exploiting the advantages of neural networks, Lee *et al.* (2022) and Yang *et al.* (2023a) employed transfer and active learning techniques, respectively. Specifically, Lee *et al.* (2022) trained a MLP model on a small number of high-fidelity numerical simulations of synthetic irregularly rough surfaces to predict the roughness function,  $\Delta U^+$ . However, the model was pre-trained using estimates of drag obtained from empirical correlations for over 10 000 rough surfaces. Yang *et al.* (2023a) used active learning, where the model automatically suggests the surface roughness that should be simulated and added to the database, to most effectively enhance the model performance. In a subsequent study by the same group, Yang *et al.* (2024) examined models of varying complexity and found that those with reduced complexity achieved superior performance when trained and tested on specific roughness types. These findings indicate a nonlinear relationship between surface roughness and induced drag, which can be effectively explored by categorizing roughness types based on their statistical properties. Shin *et al.* (2024) predicted drag using convolutional neural networks (CNNs) based on the raw topographical data of rough surfaces. Additionally, they interpreted their model to discover the drag-inducing mechanisms of roughness structures using a data-driven approach. It should be emphasized that drag prediction is a particularly demanding regression problem since each sample in the database used for training and testing is one direct numerical simulation (DNS) or experiment. Therefore, we are still far from having databases containing of the order of  $10^4$  samples, which is commonly used for developing neural networks.

Given that data-driven models show potential for rough-wall modelling, we assess the performance and limitations of increasingly complex regression methods. More specifically, this work compares

linear regression, a kernel method based on support vector machine, MLP and a convolutional network. We have an order-of-magnitude larger database compared with earlier work (Jouybari *et al.* 2021; Lee *et al.* 2022; Yang *et al.* 2023a,b). Using a GPU-accelerated numerical solver (Costa *et al.* 2021), we developed a DNS database of  $O(10^3)$  samples which includes five types of irregular homogeneous roughness. However, our database is far from complete. There exists many roughness distributions, in particular, patchy inhomogeneous ones, that we do not consider. In addition – as we will demonstrate herein – the size of our database is still small for fully exploiting neural networks. Indeed, the purpose of this study is not to identify an optimal or universal regression technique, since this will depend on the training database and the specific application. Instead, our aim is to understand the advantages and limitations between different regression approaches for drag prediction.

Alongside the actual technique used in regression, the choice of roughness features that constitute the model's input is another important aspect. For homogeneous roughness, the most common approach is to use statistics derived from the roughness height distribution, such as the peak or peak-to-trough height (Flack & Schultz 2014; Forooghi *et al.* 2017), skewness (Jelly & Busse 2018; Busse & Jelly 2023) and effective slopes (Jelly *et al.* 2022), etc. Given that rough surfaces in engineering applications often exhibit heterogeneous, e.g. patchy, structures (Sarakinis & Busse 2023), and anisotropy (Forooghi *et al.* 2017; Jelly *et al.* 2022), using the entire surface topography as input data may be needed to capture these complexities. In this paper, we will discuss different model inputs, including statistical measures and the two-dimensional height distribution of a surface.

This paper is organized into four sections: § 2 describes the generation and statistical properties of the investigated database of rough surfaces. Model training and architecture details are outlined in § 3. The drag prediction results of the modes are presented and discussed in § 4. Finally, a discussion is provided in § 5.

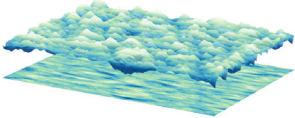
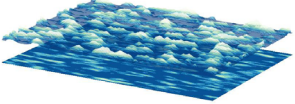
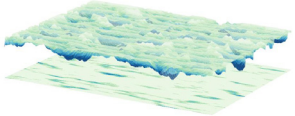
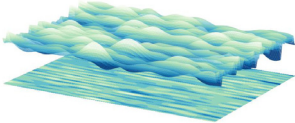
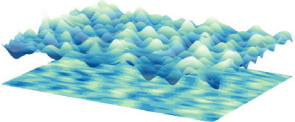
## 2. Problem setting

### 2.1. Generation of irregular rough surfaces

The dataset in this study includes five categories of irregular, statistically homogeneous rough surfaces. The surfaces are represented as a height function (or maps),  $k(x, z)$ , which is a function of streamwise ( $x$ ) and spanwise ( $z$ ) coordinates. Examples of topographies corresponding to each surface category are shown in the first column of table 1, which displays representative height maps along with their projections onto the  $x$ - $z$  plane. Rough surfaces of type  $Sk_0$  were generated using a Fourier-filtering algorithm based on the power spectrum method proposed by Jacobs, Junge & Pastewka (2017). In this study, 271 distinct  $Sk_0$  surfaces were individually produced, each characterized by varying power spectrum amplitudes and random phase shifts in the Fourier modes. By cutting off the heights below the average of random  $Sk_0$  surfaces, we obtained the second type, i.e. the positively skewed roughness ( $Sk_+$ ) with mountainous topography. The negatively skewed surfaces ( $Sk_-$ ) were generated in the opposite manner to  $Sk_+$  and exhibit basins surrounded by flat regions. These three surface types are based on a prescribed skewness and are therefore isotropic. Two other types of anisotropic surfaces were generated using the algorithm from Jelly & Busse (2018). This algorithm generates surfaces by applying linear combinations of Gaussian random matrices using a moving average process. The correlation of discrete surface heights in the wall-parallel direction is governed by a predefined target correlation function. By adjusting key parameters such as the cutoff wavenumber for the circular Fourier filter, as well as the number of streamwise and spanwise points in the correlation function, we were able to create 406 distinct random anisotropic rough surfaces dominated by streamwise- and spanwise-preferential effective slopes. These randomly generated rough surfaces are illustrated by the bottom two rows in table 1, which have streamwise- and spanwise-preferential effective slopes (labelled as  $\lambda_x$  and  $\lambda_z$ , respectively).

We adopted a number of statistical measures for parameterizing the surface topographies as listed in table 2. The left panel of the table displays the seven parameters that characterize the topographical

**Table 1.** Examples of the five roughness types: the three-dimensional topography of each type and their two-dimensional projections on the  $x$ - $z$  plane are shown in the leftmost column. The number of samples of each type used in this study is given. The sample-averaged skewness and kurtosis  $\langle \rangle$  are provided to demonstrate whether a surface is Gaussian or not in terms of its height distribution. Anisotropy is examined by the mean ratio of effective slopes over the samples in two directions.  $ES$ , effective slope in the  $x$  or  $z$  direction.

Exemplary topography	Type	No.	Properties	Isotropic
	$Sk_0$	271	$\langle Skw \rangle \approx -0.005$ $\langle Ku \rangle \approx 2.98$ Gaussian	yes $\langle ES_x/ES_z \rangle \approx 0.98$
	$Sk_+$	200	$\langle Skw \rangle \approx 1.63$ $\langle Ku \rangle \approx 5.34$ Non-Gaussian	yes $\langle ES_x/ES_z \rangle \approx 0.98$
	$Sk_-$	141	$\langle Skw \rangle \approx -1.62$ $\langle Ku \rangle \approx 5.23$ Non-Gaussian	yes $\langle ES_x/ES_z \rangle \approx 0.99$
	$\lambda_x$	194	$\langle Skw \rangle \approx 0.009$ $\langle Ku \rangle \approx 2.98$ Gaussian	no $\langle ES_x/ES_z \rangle \approx 1.41$
	$\lambda_z$	212	$\langle Skw \rangle \approx 0.005$ $\langle Ku \rangle \approx 2.95$ Gaussian	no $\langle ES_x/ES_z \rangle \approx 0.68$

information of the surface, such as the effective slopes that represent the frontal solidity of the rough surfaces. The range of the parameters for the rough surfaces investigated in this work is listed in [table 5](#) in [Appendix C](#). [Chung et al. \(2021\)](#) provides a comprehensive summary of the physical significance of these statistical parameters. The centre column displays three statistical measures of the topography's height distribution, of which the skewness has been shown to have a notable influence upon turbulent kinetic energy ([Thakkar et al. 2016](#)), shear stress ([Jelly & Busse 2018](#)) and pressure drag ([Busse & Jelly 2023](#)). Following [Jouybari et al. \(2021\)](#), we also use additional parameters formed from pairs of  $ES_x$ ,  $ES_z$ ,  $Skw$  and  $Ku$ . These take into account nonlinear effects in the model input, the significance of which will be discussed later.

## 2.2. Drag measurement

The drag penalty in turbulence from rough walls is commonly represented by the velocity deficit referred to as the roughness function  $\Delta U^+ = \Delta U/u_\tau$  ([Hama 1954](#)), i.e. the friction-scaled downward offset of the mean velocity profile in the logarithmic layer. Here,  $u_\tau \equiv \sqrt{\tau_w/\rho}$  is the friction velocity,  $\tau_w$  is the wall shear stress and  $\rho$  is the fluid density. Note that difference in the skin-friction coefficient  $C_f$  between a smooth and a rough wall (at a matched  $Re_\tau$ ) is equivalent to the roughness function  $\Delta U^+$ .

To determine the drag for each generated surface, DNSs of turbulent channel flow at  $Re_\tau = u_\tau \delta/\nu = 500$  were conducted (here,  $\delta$  is the half-channel height and  $\nu$  is viscosity). Considering the number

**Table 2.** The topographical statistics include ten ‘primary’ parameters and nine ‘pair’ parameters. The main features are divided into the ones bearing physical implications, i.e. crest height  $k_c$ , average height deviation  $R_a$ , effective slopes  $ES_{x,z}$ , porosity  $Po$ , inclinations  $inc_{x,z}$ ; and statistical parameters, i.e. root-mean-square height  $k_{rms}$ , skewness  $Skw$  and kurtosis  $Ku$ .

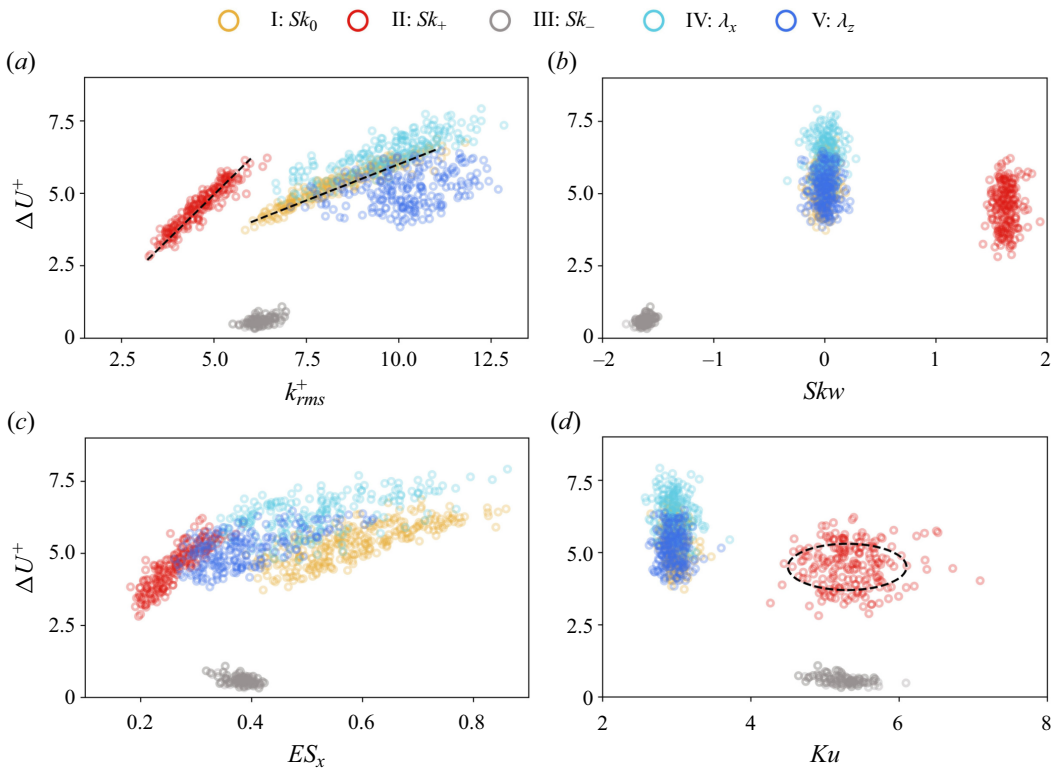
Primary parameters	Pair parameters
$k_c = k_{max} - k_{min}$	$ES_x^2, ES_z^2,$
$R_a = A^{-1} \int_{x,z}  k - k_{avg}  dA$	$ES_x \cdot ES_z,$
$ES_x = A^{-1} \int_{x,z} \left  \frac{\partial k}{\partial x} \right  dA$	$ES_x \cdot Skw,$
$ES_z = A^{-1} \int_{x,z} \left  \frac{\partial k}{\partial z} \right  dA$	$ES_x \cdot Ku,$
$Po = (A \times k_c)^{-1} \int_0^{k_c} A_f dy$	$ES_z \cdot Skw,$
$inc_x = \tan^{-1} \left\{ \frac{1}{2} Skw \left( \frac{\partial k}{\partial x} \right) \right\}$	$ES_z \cdot Ku,$
$inc_z = \tan^{-1} \left\{ \frac{1}{2} Skw \left( \frac{\partial k}{\partial z} \right) \right\}$	$Skw^2, Skw \cdot Ku$

of generated surfaces (1018), the simulations needed to be done in a cost-effective manner. For this reason, we employed the minimal-span channel approach of [Chung \*et al.\* \(2015\)](#) and [MacDonald \*et al.\* \(2017\)](#), which has proven to a successful method for characterizing the hydraulic resistance of rough surfaces under turbulent flow conditions. The minimal-span approach exploits the fact that the flow retardation imposed by the roughness occurs close to it and this effect remains constant away from the roughness, manifesting as a downward shift in the logarithmic region of turbulent velocity profile,  $\Delta U^+$ , and otherwise known as the roughness function ([Clauser 1954](#); [Hama 1954](#)). Therefore, the measure of drag we acquired from the DNS was  $\Delta U^+$ . The size of the minimal channel in this work is  $(L_x, L_y, L_z) = (2.4, 2, 0.8)\delta$ . The simulations were conducted using the open-source code CaNS ([Costa \*et al.\* 2021](#)) which solves the incompressible Navier–Stokes equations on three-dimensional Cartesian grids using second-order central finite differences. In these simulations, periodic boundary conditions were imposed along the streamwise ( $x$ ) and spanwise ( $z$ ) directions, while a Dirichlet boundary condition was applied in the wall-normal ( $y$ ) direction.

To incorporate the generated rough surfaces into the simulations, we augmented CaNS with the volume-penalization immersed-boundary method ([Kajishima \*et al.\* 2001](#); [Breugem, van Dijk & Delfos 2012](#)). Specifics regarding the solver and numerical methods used may be found in the aforementioned references which we omit here to avoid repetition. To ensure, however, that the DNS framework was able to accurately account for the effect of the irregular rough surfaces, a validation was carried out against one of the rough-wall DNS cases of [Jelly & Busse \(2019\)](#). The results of the validation are gathered in [Appendix A](#). The grid resolution in the  $x$ - and  $z$ -directions consisted of 302 and 102 points, respectively, corresponding to grid spacings of  $\Delta x^+ = 4.192$  and  $\Delta z^+ = 4.137$ , where the superscript + denotes scaling by the viscous length scale  $\delta_\nu = \nu/u_\tau$ . In the  $y$ -direction, a hyperbolic tangent stretching function was used for the grid with the smallest grid spacing in wall units,  $y^+$ , approximately 0.5. Velocity data were sampled at regular intervals, with the early stages of the simulations discarded to ensure sufficient convergence of the mean velocity. Details concerning requirements on the domain size and grid resolution that has to be satisfied when performing the minimal-span channel DNS of rough surfaces can be found in [Yang \*et al.\* \(2022\)](#).

### 2.3. Parameter space

The input to the different models is presented by  $\mathbf{x} = (x_1, \dots, x_D)$ , where  $D$  is the number of input variables. For linear regression, support vector regression (SVR) and MLP, we used the primary



**Figure 1.** Scatter distributions of  $\Delta U^+$  and four representative statistics of each type of roughness: (a)  $k_{rms}^+$ , (b)  $Skw$ , (c)  $ES_x$  and (d)  $Ku$ . The dashed straight lines in (a) highlight the linear relationship between  $\Delta U^+$  and  $k_{rms}^+$  for merely zero and positively skewed surfaces while the ‘cluster’ distribution (circle lines) in (d) indicates a nonlinear relationship between  $\Delta U^+$  and  $Ku$  prediction.

and secondary statistical measures in table 2, resulting in an input vector of size  $D = 10$  and  $D = 19$ , respectively. For the CNN model, the roughness height map is used as the input, i.e.  $D = n_z n_x = 102 \times 302$ .

Before attempting any modelling for drag prediction, simply examining the distribution of input parameters with respect to the output provides insights into the relationship between them. Figure 1 shows the scatter distribution of four representative statistics ( $k_{rms}^+$ ,  $Skw$ ,  $ES_x$ ,  $Ku$ ) with  $\Delta U^+$ . A notable degree of linearity between  $k_{rms}^+$  and  $\Delta U^+$  exists for the  $Sk_0$ - and  $Sk_+$ -surfaces while it is less for the  $\lambda_x$  and  $\lambda_z$  surfaces (figure 1a). Similarly, the effective slopes shown in figure 1(c) show a certain degree of linearity with respect to  $\Delta U^+$ . However, the ‘cluster’ distributions seen for the skewness (figure 1b) and kurtosis (figure 1d) imply a nonlinear relationship that would need to be accounted for in any model for it to be more robust. Note that the negatively skewed roughness yields a much smaller  $\Delta U^+$  compared with other types of roughness. These surfaces are dominated by a pothole-like topography with few stagnation points. As a consequence, the viscous force contributes significantly to the total drag (Busse & Jelly 2023), in contrast to the other surface types where pressure drag is dominant. Finally, we note from figure 1 that, for our surfaces,  $\Delta U^+$  falls into the range of 0.1 to 7.5, thus including both transitionally and fully rough regimes (Jiménez 2004). By including this range of roughness, the models need to learn both viscous and pressure drag components. Note that, while our database covers a continuous range in  $k_{rms}$  and  $ES_x$ , there are gaps in  $Skw$  and  $Ku$ . This is a consequence of our surface generation approach for skewed roughness. A more continuous range of training data for training regression models will be considered in future work. However, we regard the size and span of parameters in the current database of primary importance for our purposes, which is a comparative study between different regression models.

To further quantify the correlation between  $\Delta U^+$  and the input parameters, we show in figure 2 the correlation coefficient  $\rho(x_i, x_j)$ , defined as

$$\rho_{ij} = \frac{\sum (x_i - \langle x_i \rangle)(x_j - \langle x_j \rangle)}{\sqrt{\sum (x_i - \langle x_i \rangle)^2 \sum (x_j - \langle x_j \rangle)^2}}, \quad (2.1)$$

where  $x_i = k_c^+, k_{rms}^+, Skw, \dots, \Delta U^+$ . The matrix in figure 2 visualizes the degree of linearity between the surface parameters (demarcated by the dashed triangle in figure 2a) and also between the surface parameters and  $\Delta U^+$  (bottom row in figure 2a). Coefficient values greater than 0.7 indicate a strong linear correlation between two variables. The matrices of the  $Sk_0$  and  $Sk_+$  surfaces are overall similar, with the roughness height parameters ( $k_c^+, k_{rms}^+, R_a^+$ ) and effective slopes ( $ES_x, ES_z$ ) being strongly linearly correlated to  $\Delta U^+$ . This is in contrast to other surface types which manifest a nonlinear quality with respect to  $\Delta U^+$ . In particular, the  $Sk_-$  surfaces exhibit a low degree of linearity both between parameters and parameter  $\Delta U^+$ . This indicates a more intricate mapping between the surface properties of pitted surfaces and their resulting drag. For all the surface types, the roughness height and effective slopes are the parameters that exhibit a common degree of linear correlation to  $\Delta U^+$ . It is worth noting that, for the surfaces generated by the prescribed skewness (types I, IV, V), a weaker correlation between skewness and  $\Delta U^+$  is observed. As illustrated in figure 1(b), the range of  $Skw$  for each type of surface is limited due to it being a prescribed (and hence controlled) parameter. This limited range precludes the possibility of revealing any relation between  $Skw$  and  $\Delta U^+$ . This also applies to its correlation with other parameters, particularly for Gaussian surfaces.

### 3. Predictive models

For training, we use a sequence of rough surfaces  $\{x_n\}$  together with their corresponding roughness functions  $\{\Delta U_n^+\}$ , where  $n = 1, \dots, N$ . The objective is to find the least complex model that accurately predicts the roughness function  $\Delta \tilde{U}^+$  for a new (i.e. ‘unseen’) rough surface,  $x$

$$\Delta \tilde{U}^+ = f(x). \quad (3.1)$$

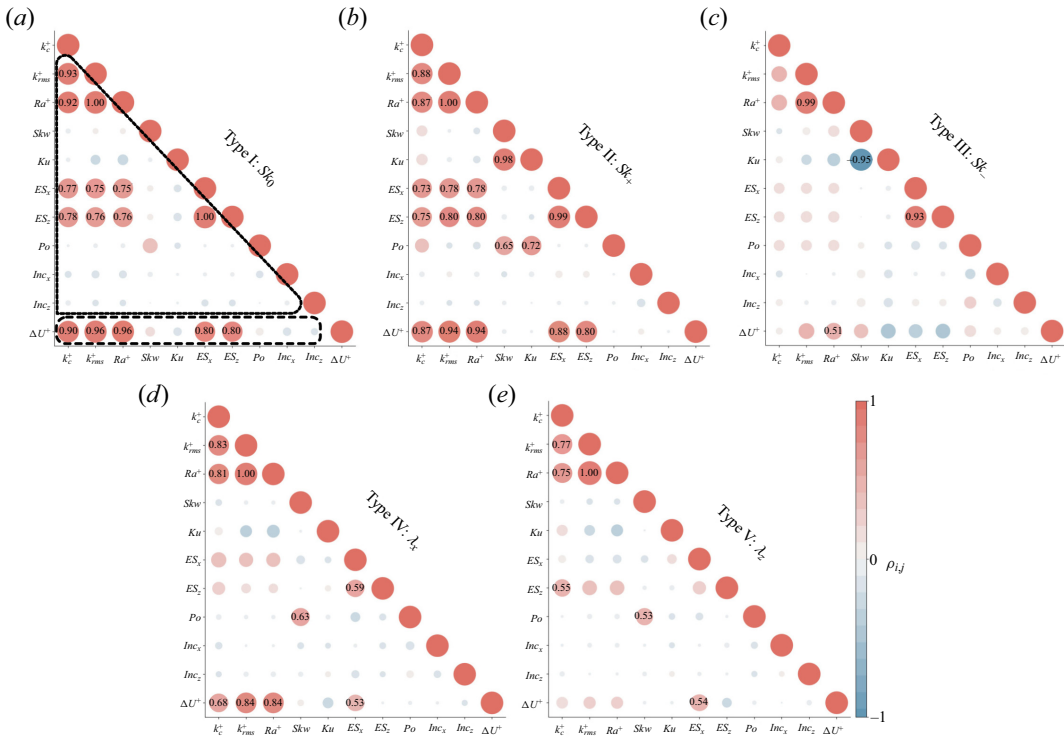
Here,  $f: \mathbb{R}^D \rightarrow \mathbb{R}$  represents different models obtained by solving a regression problem. We adopt the following approaches for creating the models: linear regression (LR), SVR utilizing kernel functions, MLP and CNN. Depending on the model, the inputs are either the statistical parameters listed in table 2 (LR, SVR, MLP) or the height maps bearing the roughness topography (CNN). We used 80 % of the total shuffled roughness data for training and validation with the remaining 20 % used for testing. A random sampling constituting 80 % of the development data is used for training, with each type of roughness comprising an equal fraction of this data. The data partitioning for training and testing is identical for all models.

Figure 3 illustrates the process of the regression modelling. For the neural networks, Bayesian optimization (BO) was used for HP tuning due to the large parameter space. The LR and SVR models were tuned manually. Several measures are used to evaluate the model performance on the test data, including the mean absolute error

$$MAE = \frac{1}{M} \sum_{i=1}^M |\Delta U_i^+ - \Delta \tilde{U}_i^+|, \quad (3.2)$$

and the mean absolute percentage error

$$MAPE = \frac{1}{M} \sum_{i=1}^M \left| \frac{\Delta U_i^+ - \Delta \tilde{U}_i^+}{\Delta U_i^+} \right| \times 100. \quad (3.3)$$



**Figure 2.** Correlation coefficients  $\rho$  of ten primary parameters and  $\Delta U^+$  for each type of roughness. The circles in the bottom row show the linear correlation between  $\Delta U^+$  and the parameters while the rest are the correlations between any two topographical parameters. Larger and darker circles represent stronger linear correlation between two variables. Those with  $|\rho_{ij}| > 0.5$  are annotated.

Here,  $M$  is the number of samples in the test data set,  $\Delta U_i^+$  is the reference drag value obtained from DNS and  $\Delta \tilde{U}_i^+$  is the drag prediction obtained from the regression model (3.1). The above measures provide the absolute and relative accuracy of the regression model. The goodness-of-fit  $R^2$  measure is also reported.

### 3.1. Linear regression

We begin with the LR model, which is the simplest of all models considered in this study. Such a model accounts for the linear correlation between the surface parameters and  $\Delta U^+$ , which were observed in figure 2. The model is defined as

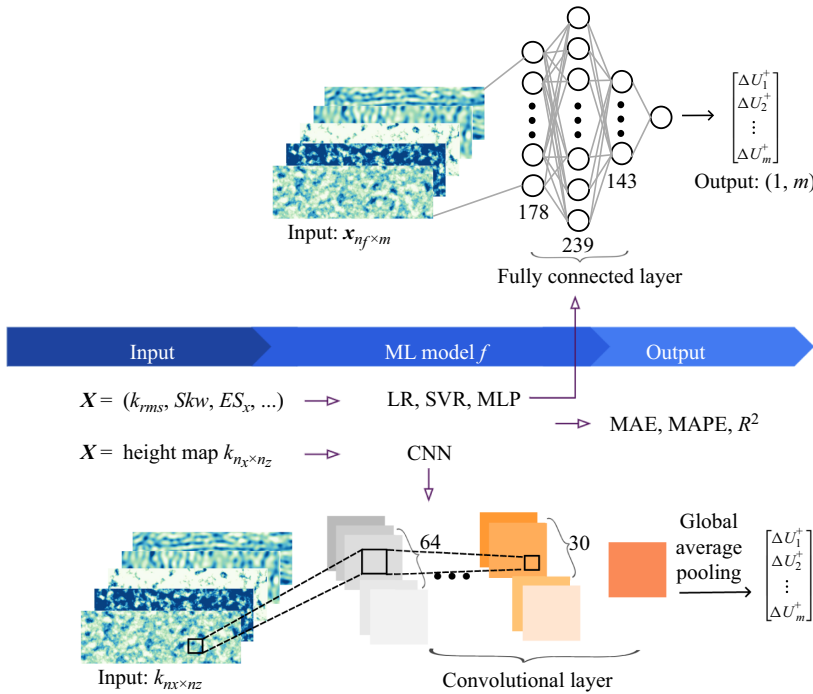
$$\Delta \tilde{U}^+(x, \mathbf{w}) = \mathbf{w}^T \mathbf{x} + b. \tag{3.4}$$

The weights  $\mathbf{w} \in \mathbb{R}^{D \times 1}$  and the bias term  $b$  are found through a least-squares optimization of the model using the training data set

$$E(\mathbf{w}) = \frac{1}{2} \sum_{i=1}^N (\Delta U_i^+ - \Delta \tilde{U}_i^+)^2. \tag{3.5}$$

Figure 4(a) shows the drag prediction using LR on the test data samples. Using the ten primary surface-derived parameters (see table 2), the model has a MAPE of = 7.9%. Figure 4(b) shows the drag prediction obtained when using an extended number of input parameters that includes both primary and pair parameters. The extended-input model reduces the error by 2%, along with a decrease in data scatter





**Figure 3.** Workflow of drag prediction. The four models are evaluated by MAE, MAPE and  $R^2$ . The model architectures of MLP and CNN are illustrated, wherein the hyperparameters (HPs) are determined using Bayesian optimization.

(improved  $R^2$ ). By including the pair parameters of roughness in the model input, we are incorporating nonlinear effects in the LR model. However, the choice of pair parameters in table 2 is arbitrary and we have chosen them to be similar to those of Jouybari *et al.* (2021).

### 3.2. Support vector regression

To increase the fidelity of the model, we now turn our attention to SVR, which allows for nonlinear regression through the use of kernel functions. Replacing the input vector  $\mathbf{x}$  in (3.4) with a nonlinear mapping  $\bar{\phi}(\mathbf{x})$ , we will have

$$\Delta \tilde{U}^+(\mathbf{x}) = \mathbf{w}^T \bar{\phi}(\mathbf{x}) + b. \tag{3.6}$$

When using kernel functions, the weight vector  $\mathbf{w}$  is given by a linear combination of the expansion basis

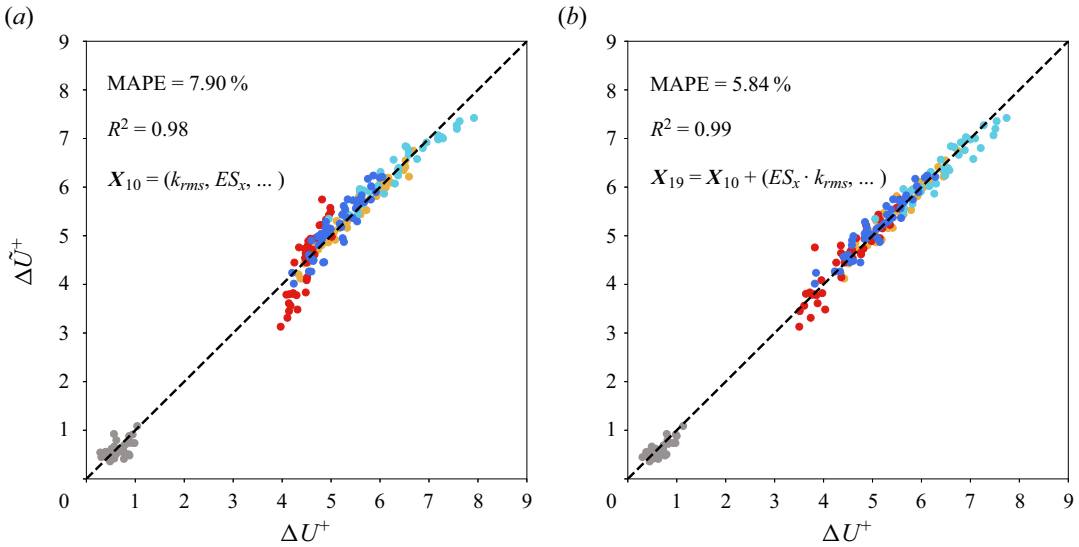
$$\mathbf{w} = \sum_{i=1}^N a_i \bar{\phi}(\mathbf{x}_i). \tag{3.7}$$

Inserting (3.7) into (3.6) results in

$$\Delta \tilde{U}^+(\mathbf{x}) = \sum_{i=1}^N a_i \bar{\phi}(\mathbf{x}_i)^T \bar{\phi}(\mathbf{x}) + b = \sum_{i=1}^N a_i k(\mathbf{x}_i, \mathbf{x}) + b, \tag{3.8}$$

where  $k(\mathbf{x}_i, \mathbf{x})$  is the kernel.

In the model above, the prediction requires  $N$  function evaluations. Since,  $N \in \mathcal{O}(10^3)$  is the number of training samples, kernel evaluations become inefficient for large datasets. Support vector regression



**Figure 4.** The  $\Delta U^+$  predictions of LR versus those from DNS: model using (a) 10 primary statistics and (b) 19 statistics (i.e. including 9 pair-product parameters.)

sparsifies the kernel by including only support vectors in the expansion. To achieve this, instead of a least-squares minimization (3.5), one minimizes the  $\epsilon$ -sensitive cost function, defined as

$$J(\Delta U^+ - \Delta \tilde{U}^+) = \begin{cases} |\Delta \tilde{U}^+ - \Delta U^+| - \epsilon & \text{for } |\Delta \tilde{U}^+ - \Delta U^+| > \epsilon \\ 0 & \text{otherwise} \end{cases} \quad (3.9)$$

This means that only errors larger than  $\epsilon$  contribute to the cost function.

To determine the weights and the bias in (3.6), we minimize the regularized cost function

$$E(\mathbf{w}) = C \sum_{j=1}^N J(\Delta U_j^+ - \Delta \tilde{U}_j^+) + \frac{1}{2} \|\mathbf{w}\|^2. \quad (3.10)$$

The second term is the regularization term that penalizes large weights, i.e. promoting flatness (i.e. achieving a smoother loss value). Note that, by convention, the regularization parameter  $C$  appears in front of the first term. The key aspect of SVR is that, by using (3.9),  $a_j$  values in (3.8) are non-zero only for the training samples either lying on or above the boundary defined by  $\epsilon$ .

The choice of kernel in this work for nonlinear mapping is the radial basis function (RBF)

$$k(\mathbf{x}_i, \mathbf{x}) = \exp(-\gamma \|\mathbf{x}_i - \mathbf{x}\|^2), \quad (3.11)$$

where  $\gamma = 1/(N\sigma^2)$  is the kernel coefficient and  $\sigma^2$  is the variance of the training data. The input data  $\mathbf{x}$  are rescaled by the min–max normalization while the scaling of the target  $\Delta U^+$  is insignificant for prediction. The parameter  $C$  and kernel bandwidth  $\epsilon$  were tuned and the best performance was obtained for values of  $C = 0.1$  and  $\epsilon = 0.01$ . We have presented a simplified formulation of the optimization problem associated with SVR here. We refer to Cortes & Vapnik (1995) and Smola & Bernhard (2002) for the complete formulation of the kernel in the optimization process, including the use of slack variables.

### 3.3. Neural networks

While SVR has far greater capacity and fidelity than LR – due to mapping the input space onto a higher-dimensional space – it still requires the user to choose an appropriate expansion basis  $\bar{\phi}(\mathbf{x})$ . Neural networks can learn  $\bar{\phi}$  from a broad class of functions and form a composition of such functions using hidden layers. Neural networks often require more training data than SVR to generalize well and constitute a non-convex optimization problem. To explore neural networks for drag prediction, we consider MLP and CNN.

#### 3.3.1. Multi-layer perceptron

The MLP model is composed of multiple layers of neurons, where the neurons of two adjacent layers are connected by weights. The inputs are either the 10 primary statistics or the extended set of 19 statistics listed in table 2. The output,  $\Delta\tilde{U}_i^+$ , is composed from the nonlinear transfer functions of each layer. This is what enables an MLP to account for high degrees of nonlinearity. The objective is to identify the weights of a network such that the following loss function is minimized:

$$E(\mathbf{w}) = \frac{1}{2} \sum_{i=1}^N \|\Delta U_i^+ - \Delta\tilde{U}_i^+\|^2 + \frac{\lambda}{2} \mathbf{w}^T \mathbf{w}. \quad (3.12)$$

The loss function is composed of a sum of squared errors term and a regularization term. The weight vector  $\mathbf{w}$  contains values between the neurons of adjacent layers of the network.

We performed BO to determine the HPs of the MLP, including the number of layers, the number of neurons, the learning rate, the regularization term  $\lambda$ , the activation function and the initialization of the weights. The Gaussian process acts as a surrogate model to estimate the model performance and the HPs are updated after each evaluation of the loss function. The acquisition function directs the next search location in the given range of parameter space to find the optimal set of HPs. At each iteration, these HPs are evaluated by training the neural network, where the number of evaluations depends on the input dimension.

Using BO, we developed two architectures. The first one maintains a fixed number of layers with an optimized number of neurons. The second architecture has an optimized number of layers but a fixed number of neurons. Given that each layer learns different information from the previous input, the number of neurons or filters, in theory, should differ at each layer. After conducting a set of comparative trials for both architectures, we adopted the first architecture since it exhibited a slightly lower relative error. The final HPs for the two MLP models are displayed in table 3. Note that, to ensure consistent scaling, the inputs were rescaled by their respective standard deviations.

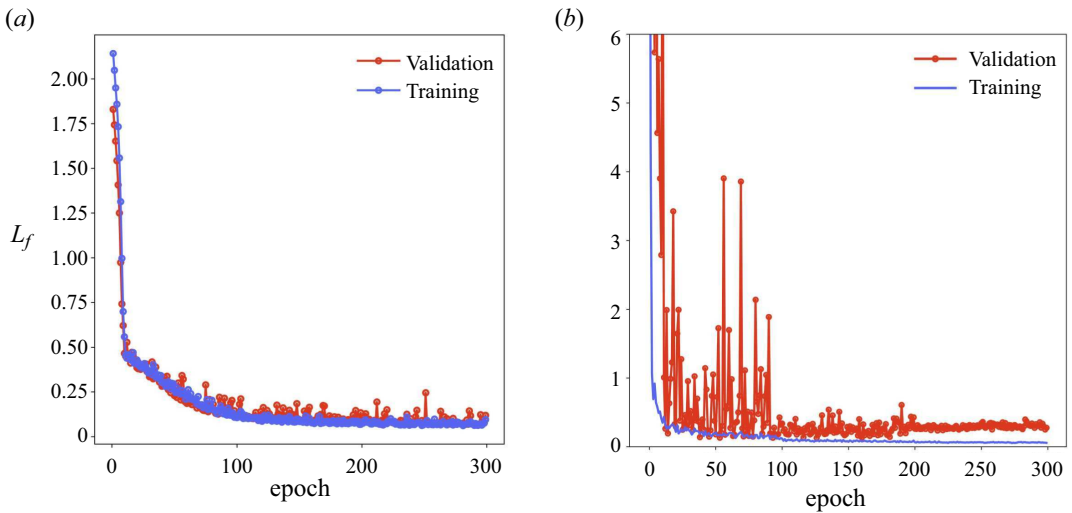
Figure 5(a) shows the training and validation losses for the MLP as a function of the number of epochs (i.e. iterations in the BO optimization process). The rapid decay of the training curve to a plateau after 100 epochs indicates a fast convergence. The validation curve – which represents the loss on a separate dataset not used for training – follows a similar initial decay followed by a plateau. This indicates that the model generalizes to unseen data relatively well.

#### 3.3.2. Convolutional neural network

This regression model is a network with convolutional layers, i.e. a set of filters (or kernels) that are convoluted with the layer's input data. One key feature is that it has sparse connectivity between the neurons, allowing for the processing of very high-dimensional input data. In our case, the input is a two-dimensional function representing the height of the surface roughness. The objective of the CNN is to identify weights to minimize the loss function (3.12). We followed the same procedure used for the MLP to determine the architecture, i.e. the HPs were obtained using Bayesian optimization. The number of blocks, filters, kernel size, learning rate, activation function and weight initialization of the CNN are reported in table 3.

**Table 3.** The Bayesian-optimized HPs in  $MLP_{10}$ ,  $MLP_{19}$  and CNN that are used for prediction in this work. ReLU, rectified linear unit.

Model	Number of layers/blocks	Number of neurons/filters	Filter sizes	Learning rate	$\lambda_2$	Batch sizes	Activation function	Initialization
$MLP_{10}$	3	(256, 109, 256)	N/A	$6 \times 10^{-3}$	$2.2 \times 10^{-4}$	(3, 2)	leaky ReLU	Glorot uniform
$MLP_{19}$	3	(178, 239, 143)	N/A	$1.3 \times 10^{-4}$	$1.3 \times 10^{-4}$	(3, 9)	leaky ReLU	Glorot uniform
CNN	5	(64, 37, 64, 44, 30)	(3, 6, 7, 8, 3)	$7 \times 10^{-5}$	$1 \times 10^{-5}$	(17, 16)	leaky ReLU	Glorot uniform



**Figure 5.** Loss curves of training and validation in the Bayesian-optimized (a)  $MLP_{10}$  and (b) CNN with learning rate reschedule. Early stopping was employed during the neural network within the BO process to mitigate overfitting and expedite training.

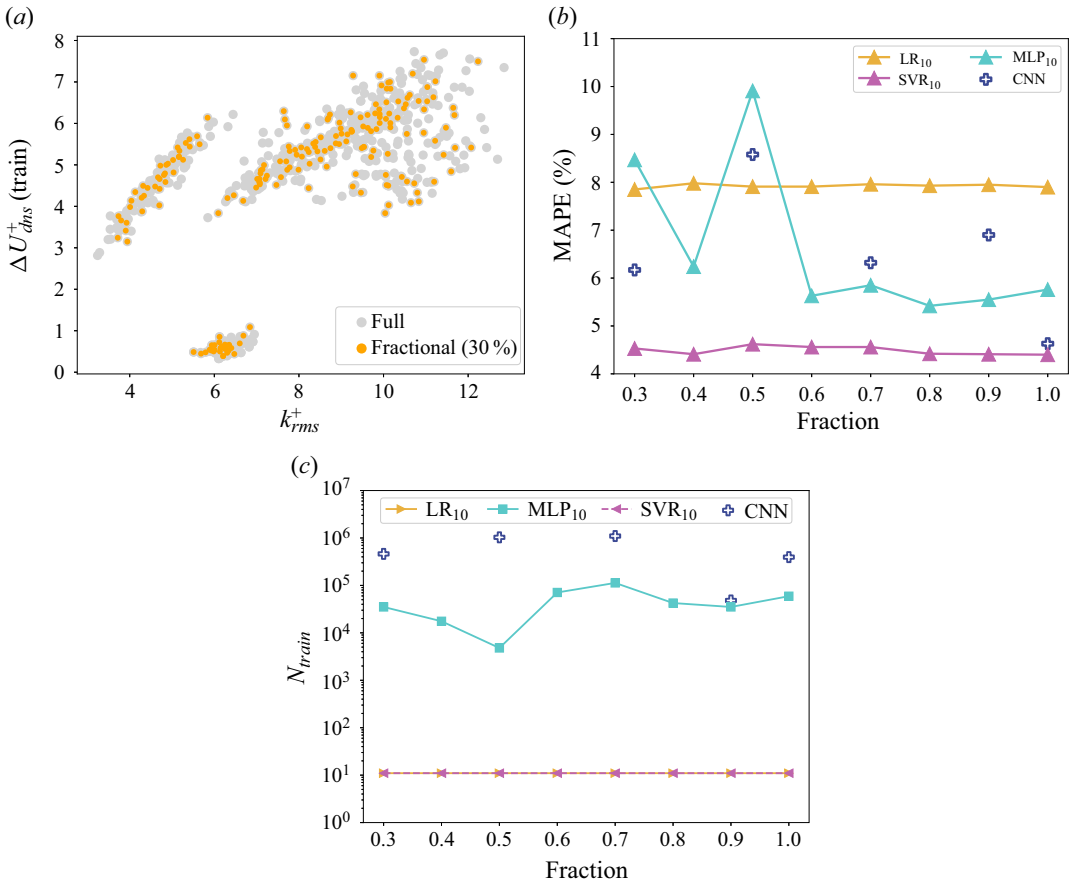
Figure 5(b) shows the training and validation losses for the CNN model. While the loss function of the training demonstrates a fast convergence, the corresponding validation curve shows large oscillations for the first 100 epochs, indicating overfitting of the unseen data. To improve CNN convergence, we implemented a learning rate schedule that reduces the rate by 0.1 every 100 epochs starting from epoch 100. Note that many other architectures are potentially more suitable for drag prediction. Our choice is, however, sufficient for comparative purposes.

### 3.3.3. Sensitivity analysis of training size

The size of training data is critical for truly exploiting the advantages of neural networks. While our dataset with over 1000 samples is, to the best of our knowledge, the largest such collection for rough-wall turbulence, it is still relatively small compared with what is commonly used for training neural networks in other applications. Therefore, we conducted a sensitivity analysis of the sample size for the training process. To ensure an even representation across different surface categories in parameter space, training samples are uniformly downsampled by the same proportion, as illustrated in figure 6(a). The depth of the new neural networks (NNs) trained using varying data fractions was kept the same as the initial MLP and CNN architectures, while the number of network units were optimized using BO.

Figure 6(b) presents the relative prediction errors (MAPE) of identical test data using models trained with varying data fractions. The SVR achieves the lowest error and exhibits high robustness, as its predictions remain consistent for all training data fractions. The prediction by linear regression is also not affected by the data size but consistently yields the highest error among the models compared. As expected, the performance of NNs depends on the size of the training data. The MLP model converges with a 60% fraction of the entire data, while the CNN model does not exhibit a clear convergence trend. Despite that, the best CNN, trained using the full dataset, achieves a low error of 4.6%, which is comparable to SVR. Therefore, the CNN model has not yet achieved adequate generalizability to be employed for unseen data.

Figure 6(c) shows the variation of the number of trainable parameters ( $N_{train}$ ) for the models with different training data fractions. Unlike LR and SVR, where  $N_{train}$  is fixed, the NN models exhibit a non-monotonic trend. The MLP seems to stabilize around an  $N_{train}$  of the order of  $10^5$  after reaching a fraction of 70%, with roughly an order of magnitude fewer trainable parameters on average than the CNN. This value is an approximation of the optimal model capacity for learning the underlying



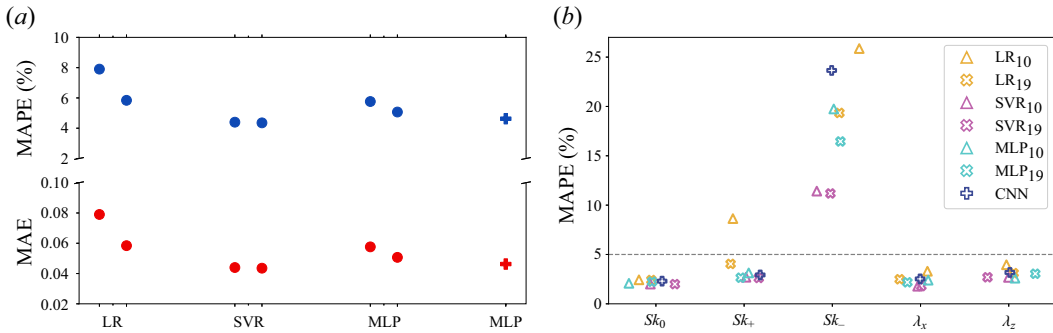
**Figure 6.** (a) The sample coverage in  $\Delta U^+ - k_{rms}^+$  space at the fraction of 30%. The reduced training samples consistently cover the full parameter space. (b) The MAPE of inference obtained from LR<sub>10</sub>, MLP<sub>10</sub>, SVR<sub>10</sub> and CNN at different sample fractions. (c) The variation in the number of trainable parameters in each model at different sample fractions.

mapping. In contrast, CNN models experience a significant drop by an order of magnitude at a fraction of 90%, followed by an increase. This behaviour indicates an overfitting for CNN with the current volume of data, as was reflected by the loss in figure 5(b). The model evaluations in § 4 use predictions from models which were trained on the entire dataset.

## 4. Drag prediction performance

### 4.1. Comparison of regression models

We now compare the performance of the four types of regression models. We trained LR, SVR and MLP using two sets of inputs: ten primary statistics and 19 pair statistics, as listed in table 2. We refer to these models as LR<sub>10</sub>, LR<sub>19</sub>, etc. The absolute error (MAE in (3.2)) and relative error (MAPE in (3.3)) obtained from the seven evaluated models trained on the entire roughness dataset are shown in figure 7(a). The LR trained using ten primary statistics displays the largest error in predicting  $\Delta U^+$ . As previously observed in figure 4, by incorporating the nine additional pair parameters, some degree of nonlinearity becomes incorporated into the LR model and its error becomes reduced.



**Figure 7.** (a) Values of MAPE (%) (blue) and MAE (red) obtained from all models trained by the hybrid data. Left and right symbols correspond to 10 and 19 input parameters. All maximum errors correspond to negatively skewed surfaces (type:  $Sk_-$ ). (b) Applying the trained model by the full dataset on each type of surface thus the corresponding mean errors. The data are slightly displaced on the horizontal axis for the same type of roughness to increase clarity.

The SVR emerges as the optimal predictive model with an error of 4.4 %, the smallest of all models. The use of the extended input ( $SVR_{19}$ ) does not improve prediction compared with the  $SVR_{10}$  model. This suggests that the chosen kernel function effectively captures the nonlinearity embedded within the input space. Moving on to the performance of MLP, we make two observations. First,  $MLP_{19}$  yields a MAE of around 5 %, which is slightly larger than the SVR model. Second, the input size has a small influence on the prediction performance, with  $MLP_{19}$  performing slightly better than  $MLP_{10}$ . Although the network has a near-optimal performance, these observations imply that it has not fully captured the nonlinearity in the mapping from the inputs to the output. Presumably, a different network and/or larger database are required to reach the same level of performance as the SVR model.

Finally, we observe that the best-performing CNN achieves comparable results to SVR. However, it requires significantly longer training time due to its vast number of trainable parameters ( $O(10^5)$ ). As mentioned in the previous section, to develop a generalizable CNN model a larger dataset is required. The local spatial topographical information in this specific case does not offer discernible advantage for solely predicting a single scalar value ( $\Delta U^+$ ). However, CNN is inherently capable of learning hierarchical representations from grid-like data, which could potentially be advantageous when considering patchy, inhomogeneously distributed roughness where a statistical parameterization becomes non-trivial.

#### 4.2. Model performance for different surface categories

In addition to evaluating the prediction accuracy using test data from the full roughness database, further insight into the models is gained by assessing how accurately they predict different roughness types. As shown in figure 7(b), all predictive models demonstrate a comparable level of accuracy in predicting the Gaussian surfaces, i.e.  $Sk_0$ ,  $\lambda_x$  and  $\lambda_z$ . The average error of all models, including  $LR_{10}$ , is around 2 %–3 %. The largest errors are found for the negatively skewed roughnesses, which have a pit-dominated topography. Note that MAPE is normalized with  $\Delta U^+$ , resulting in large errors for small values of  $\Delta U^+$ , which is the case for  $Sk_-$ . For example, applying  $LR_{10}$  to the  $Sk_-$  test data gives  $\langle \Delta U^+ \rangle = 0.60$  and  $\langle \Delta \tilde{U}^+ \rangle = 0.64$ , resulting in a large  $MAPE = 25$  %. This is despite the difference between the DNS and predicted drag being relatively small. Considering that MAPE is a sensitive measure for small target values, we can confirm that the SVR models notably outperform the NNs, where the latter have an error of  $\sim 11$  %.

In summary, we observe that SVR is consistently the most robust (figure 6) and efficient (figure 7) model in predicting the additional drag induced by homogeneously distributed irregular roughness.

While the NN models have comparable performance, they lack robustness when trained on our dataset consisting of  $\mathcal{O}(10^3)$  samples.

### 4.3. Key features for SVR prediction

We observed that SVR model's prediction performance did not benefit from extending the input with the pair parameters. In contrast, we have observed that the nonlinearity introduced by using the extended input improves the prediction performance of both the LR and MLP models, which indicates that these models are unable to fully learn the nonlinearity inherent in the data. In the following section, we attempt to provide insight into the capabilities of SVR.

#### 4.3.1. Choice and interpretation of kernels

The SVR uses a kernel to map a low-dimensional input vector to a high-dimensional space where the relationship between the inputs and the output (e.g.  $\Delta U^+$ ) can be mapped linearly. This feature allows SVR to implicitly take into account the pair parameters in table 2, even though the actual input,  $\mathbf{x}$ , used in the model (3.6) only contains the primary parameters. In Appendix B, we present a simple example using the kernel

$$k(\mathbf{x}_i, \mathbf{x}) = (1 + \mathbf{x}_i^T \mathbf{x})^2, \quad (4.1)$$

to illustrate how a nonlinear relation between primary surface statistics and  $\Delta U^+$  is transformed to a linear relation between an extended-input vector and  $\Delta U^+$ .

Leveraging kernels removes the arbitrariness that exists when manually selecting different combinations of the primary statistics. In fact, by constructing the kernel using radial basis functions, not only the pair parameters but an infinite number of products of the primary statistics are taken into account. To see how this works, consider the RBF kernel of (3.11) and assume that  $N = 1$ ,  $D = 1$  and  $\gamma = 1$ , i.e.

$$k(x_1, x) = e^{-(x_1-x)^2} = e^{-x_1^2-x^2} \sum_{j=0}^{\infty} (x_1 x)^j. \quad (4.2)$$

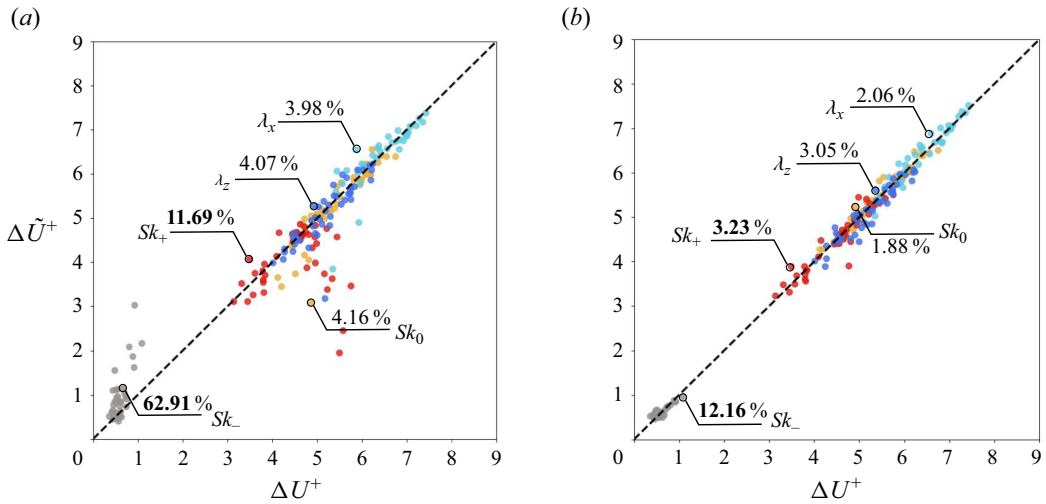
We observe that the last term is an infinite sum containing products of  $x$ . More generally, the Gaussian kernel can be interpreted as a measure of similarity between the input vector of a new surface,  $\mathbf{x}$ , and those of each surface in the training data set,  $\mathbf{x}_i$ . If  $\mathbf{x}$  is close to  $\mathbf{x}_i$  (in a Euclidean sense), then the corresponding term in the expansion (3.6) is large.

#### 4.3.2. Minimal SVR input space

We now move on to identify the smallest set of primary statistics that is needed in the input vector  $\mathbf{x}$  for accurate drag prediction using SVR. There is no need to consider products of primary parameters, since SVR implicitly takes into account all such nonlinearities through the RBF kernel. Previous works (see e.g. Chung *et al.* 2021) have highlighted that measures of height, effective slope and skewness are necessary to capture the drag increase from homogeneously distributed roughness. Indeed, it is likely that inputs in  $SVR_{10}$  and  $SVR_{19}$  contain redundancy since, for example, the first three parameters ( $k_c^+$ ,  $k_{rms}^+$ ,  $R_a^+$ ) all represent the roughness height. The high correlations between these features, as seen in the triangle area demarcated in figure 2, further support this notion.

We first focus on the case where the input is  $\mathbf{x} = (k_{rms}^+, ES_x, ES_z)$  and comprises only the vectors of the three parameters. A new SVR model,  $SVR_3$ , is trained on all types of roughness using these three parameters, followed by inference on the same testing data employed earlier. Figure 8(a) shows the diagonal spread of  $\Delta U^+$  (obtained from DNS) and  $\Delta \hat{U}^+$  (predicted) for each category. A reasonably good agreement for Gaussian surfaces ( $Sk_0$ ,  $\lambda_x$ , and  $\lambda_z$ ) is observed with a mean error of  $\sim 4\%$ . However, for  $Sk_+$  surfaces, the prediction exhibits notable deviations from the DNS data, which cause the mean error to become  $\sim 12\%$ , in contrast to the significantly lower errors of  $\sim 2.7\%$  obtained from  $SVR_{10}$  and





**Figure 8.** The scatter distribution of  $\Delta U^+$  vs  $\Delta \tilde{U}^+$  obtained from the new SVR. The model is trained with the reduced input space involving (a)  $k_{rms}^+$ ,  $ES_x$ ,  $ES_z$  (b) and additional  $Skw$ . The error reduction is observed for  $Sk_-$  and  $Sk_+$ -roughness, marked in bold.

$SVR_{19}$ . Such outliers are also observed for  $Sk_-$ . Furthermore, we tried to replace  $k_{rms}^+$  with  $k_c^+$  and  $R_a^+$  for SVR training. While using the crest height,  $k_c^+$ , still gave a model with reasonable prediction accuracy, it was still inferior compared with models which used  $k_{rms}^+$  and  $R_a^+$ . This finding suggests that  $k_{rms}^+$  and  $R_a^+$  provide a more comprehensive quantification of the surface terrain than  $k_c^+$ .

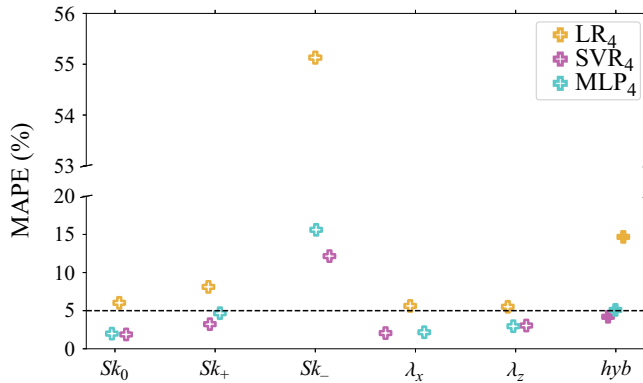
Measures of roughness height and effective slopes, or any nonlinear combination of them, are apparently insufficient to represent the skewed surfaces. We therefore trained model  $SVR_4$  where the input was  $\mathbf{x} = (k_{rms}^+, ES_x, ES_z, Skw)$ . Figure 8(b) shows a significant error reduction for  $Sk_-$  and  $Sk_+$ , demonstrating the essential role in rectifying all deviant samples for non-Gaussian surfaces. This result aligns with the observations of previous studies showing that skewness plays an important role in drag prediction (Flack *et al.* 2016; Forooghi *et al.* 2017; De Marchis *et al.* 2020; Busse & Jelly 2023).

#### 4.4. Model performance using minimal input space

Our findings suggest that skewness, in combination with  $k_{rms}^+$  and effective slopes in both directions, can satisfactorily predict  $\Delta U^+$ . To investigate the influence of model complexity on key features, we trained LR and MLP models using only these four parameters. Figure 9 compares MAPE for each roughness type and all roughness types combined (denoted as *hyb*). Here,  $LR_4$  achieves a MAPE of  $\sim 15\%$  for all-surface prediction, which is twice the error obtained by  $LR_{10}$ . In contrast,  $SVR_4$  and  $MLP_4$  maintain low errors of around 5%, which is comparable to the performance of  $SVR_{10}$  and  $MLP_{19}$ . For Gaussian surfaces,  $LR_4$  can still yield a reasonably accurate prediction, while its performance deteriorates for negatively skewed surfaces, resulting in  $MAPE = 55\%$ .

## 5. Discussion

In this work, we assessed data-driven regression models of different complexity to provide insight into the most appropriate modelling choice for rough-wall turbulence drag prediction. To achieve this, we generated 1018 surface samples comprising five categories of homogeneous roughness (Gaussian/non-Gaussian, isotropic/anisotropic). To train and test the models, DNSs at  $Re_\tau = 500$  were used to generate corresponding drag values, both in the transitionally and fully rough regimes. The database is sufficiently large to allow for comparison of different regression models, including linear models, SVR and NNs.



**Figure 9.** The average errors of each roughness category (empty circles) obtained from the three models ( $LR_4$ ,  $SVR_4$ ,  $MLP_4$ ) trained by the reduced input  $\mathbf{x} = (k_{rms}^+, ES_x, ES_z, Skw)$ . The solid crosses represent the mean error of all surfaces.

However, one should bear in mind that there exists many types of roughness distributions (e.g. patchy) and flow configurations (e.g. pressure gradients) that we have not considered.

For datasets of size  $\mathcal{O}(10^3)$  or smaller, kernel-based SVR serves as a very competitive tool for drag prediction – in particular when only a few statistical measures of the surface are available. We have shown how kernels transform nonlinear mapping between the surface statistics ( $k_{rms}^+$ ,  $ES_x$ ,  $ES_z$ ,  $Skw$ , ...) and  $\Delta U^+$  into a linear one via a so-called feature map function  $\bar{\phi}$ . Specifically, using SVR with RBFs produces an efficient prediction model. For example, SVR predictions for the Gaussian ( $Sk \approx 0$ ) and peak-featuring ( $Sk \gtrsim 0$ ) surfaces had mean errors of  $\sim 3\%$ . We also demonstrated that an SVR model that uses four surface measures ( $Skw$ ,  $k_{rms}$ ,  $ES_x$  and  $ES_z$ ) provides nearly as good performance as one using 10 or 19 measures.

We demonstrated that LR methods can be expected to deliver a mean error of around 10% for homogeneous rough surfaces, which may be sufficient for some applications. However, the model uncertainty is high, with errors reaching 25% for negatively skewed surfaces falling into in the transitionally rough regime. The model captures the linear correlations in the data, which can be significant, for example, between  $k_{rms}^+$  and  $\Delta U^+$  for Gaussian-type roughness. However, linear regression cannot capture nonlinear correlations in the data, for example, between  $Skw$  and  $\Delta U^+$ . Nevertheless, a linear mapping between surface characteristics and drag accounts for a significant amount of physical information for most surface categories.

Neural networks can be regarded as an extension of kernel-based methods, because the basis functions  $\bar{\phi}(\mathbf{x})$  (and thus the kernel) can, alongside the weights  $w$ , be tuned during the training. While the prediction accuracy of MLP is on par with SVR, it still depends on the input data size. This can be explained by the fact that the optimization problem for NN is not convex. The SVR, in contrast, is a convex optimization problem and will therefore find the global minimum of the objective function (3.9). In addition, the number of trainable parameters of MLP is three orders of magnitude larger than LR and SVR. It therefore seems that – for our dataset – MLP does not offer any clear advantage over SVR. The true advantage of MLP emerges for much larger training datasets, allowing for the many trainable parameters to be adequately tuned. In theory, a two-layer MLP can approximate any continuous function on a compact input domain if the network has sufficiently many hidden units. The SVR, on other hand, becomes rapidly inefficient for large training datasets, since the number of expansion terms in the model is of the same order of magnitude as the number of training samples.

We also investigated convolutional networks for drag prediction using the roughness height distribution of a surface as the input data. We did not design a CNN that is optimal for drag prediction, as the purpose of our study was merely to understand the advantages and limitations between different

regression approaches. We found that CNN performs satisfactorily, but the model does not generalize well, which is not surprising as the number of trainable parameters is very large compared with the size of the dataset. The CNNs, however, have features that make them particularly interesting for inhomogeneous rough surfaces, i.e. roughness with features that vary spatially on a length scale comparable to the system scale (e.g. pipe radius, boundary layer thickness). The reason is that mapping the full surface topography to drag can be considered as a pattern recognition problem. The CNN is well suited for such problems, allowing invariances to be built into the architecture. For example, when feeding the roughness height of a surface into a model, one may expect that the output should be independent of the exact position of a particular roughness element with respect to other elements. Such translational invariance can be built into the structure of a CNN. Another reason why CNNs are appropriate is because it is highly non-trivial to characterize inhomogeneous roughness using statistical measures, motivating the direct use of the surface height distribution.

A key takeaway from this study is that kernel-based SVRs possess high-enough fidelity for drag prediction, at least for the foreseeable future where very large databases would be unavailable. In time – and perhaps through a collective community effort – it is likely that a sufficient amount of relevant roughness data will be accumulated to facilitate the development of efficient prediction models. We anticipate that different regression models, some of them studied here, will be suitable for different applications.

**Supplementary material.** Raw data are available from the author upon reasonable request.

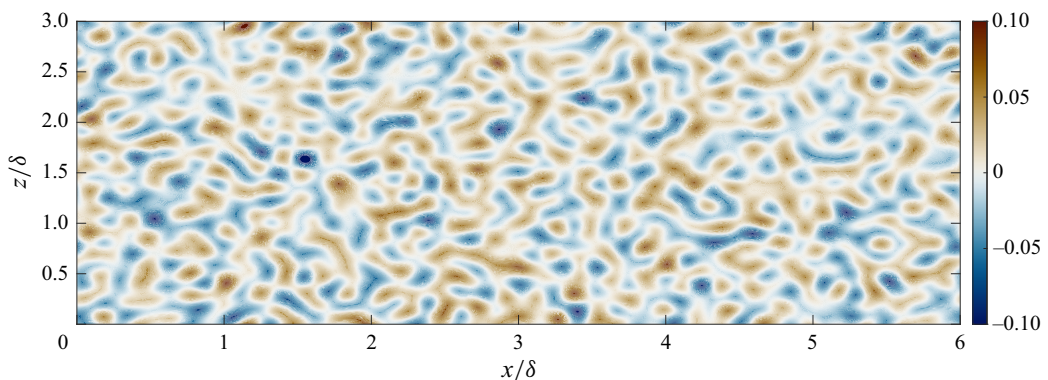
**Funding statement.** Z.S., S.M.H.K. and S.B. gratefully acknowledge funding provided by the Swedish Energy Agency and computing time acquired through the National Academic Infrastructure for Supercomputing in Sweden (NAISS). H.S. and S.L. acknowledge the funding received through the Inha university research grant and the National Research Foundation of Korea grant funded by the Korean government (NRF-2022R1F1A1066547).

**Declaration of interests.** The authors declare no conflict of interest.

## Appendix A. Validation of DNS solver for turbulent channel flow over irregular roughness

We chose the irregular rough surface of Jelly & Busse (2019) for validation purposes which belongs to a DNS dataset made publicly available by the authors. The height map of the surface is shown in figure 10.

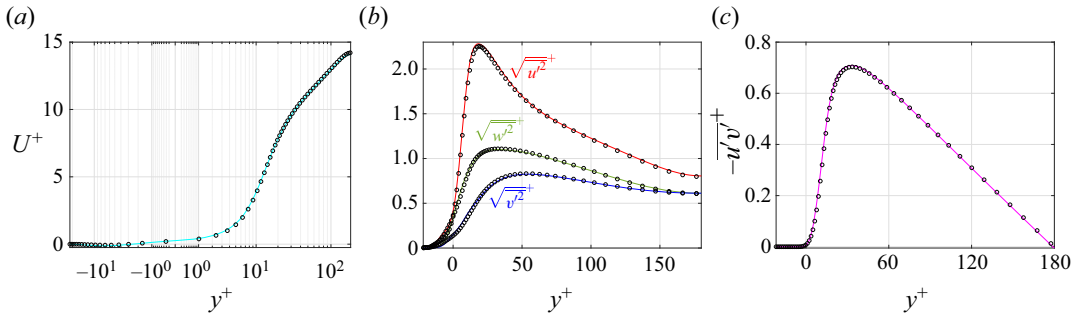
To minimize uncertainties in trying to reproduce the results of Jelly & Busse (2019), we endeavoured to match their DNS parameters as closely as possible. Therefore, the grid resolution was chosen to be the same as theirs and the stretched wall-normal grid was also taken from their dataset. These and other relevant parameters are listed in table 4.



**Figure 10.** Height map of the irregular rough surface from Jelly & Busse (2019). Colour bar values indicate the surface height relative to the mean reference plane, i.e.  $k/\delta$ .

**Table 4.** Simulation parameters for the validation case of the irregular rough surface of figure 10: friction Reynolds number ( $Re_\tau$ ); domain lengths along streamwise ( $L_x$ ), wall-normal ( $L_y$ ) and spanwise ( $L_z$ ) directions; viscous-scaled streamwise grid spacing ( $\Delta x^+$ ), minimum wall-normal grid spacing ( $\Delta y_{min}^+$ ); maximum wall-normal grid spacing ( $\Delta y_{max}^+$ ); spanwise grid spacing ( $\Delta z^+$ ). Additional details and descriptions can be found in Jelly & Busse (2019).

$Re_\tau$	$L_x/\delta$	$L_y/\delta$	$L_z/\delta$	$\Delta x^+$	$\Delta y_{min}^+$	$\Delta y_{max}^+$	$\Delta z^+$
180	6.0	3.0	2.0	2.81	0.67	5.00	2.81



**Figure 11.** Validation results: (a) mean velocity; (b) root-mean-square velocity fluctuations; (c) Reynolds shear stress. Lines are the results using CaNS and symbols are the data from Jelly & Busse (2019).

Figure 11 compares the time- and plane-averaged mean flow and fluctuations obtained using CaNS to those of Jelly & Busse (2019). The agreement is good across all of the quantities considered, thereby validating the DNS methodology used to obtain the  $\Delta U^+$  values for the rough surfaces of this study.

**Appendix B. Kernel methods: an example**

Consider an input vector  $\mathbf{x} = (x_1, x_2) \in \mathbf{R}^2$  and the kernel

$$k(\mathbf{x}_i, \mathbf{x}) = (1 + \mathbf{x}_i^T \mathbf{x})^2, \tag{B1}$$

where  $\mathbf{x}_i = (x_{i,1}, x_{i,2})$  corresponds to the statistics of the  $i$ th surface in the training dataset of size  $N$ . The corresponding kernel-based model (3.8) can be written as

$$\Delta \tilde{U}^+(\mathbf{x}) = \sum_{i=1}^N a_i (1 + \mathbf{x}_i^T \mathbf{x})^2 = \mathbf{w}^T \hat{\mathbf{x}}, \tag{B2}$$

with  $\hat{\mathbf{x}} = (1, \sqrt{2}x_1, \sqrt{2}x_2, x_1^2, x_2^2, \sqrt{2}x_1x_2) \in \mathbf{R}^6$ . Specifically, if the two-dimensional input vector to the kernel method is  $\mathbf{x} = (ES_x, Skw)$  then there is an equivalent LR model with a six-dimensional input vector  $\hat{\mathbf{x}} = (1, ES_x, Skw, ES_x^2, Skw^2, ES_x \cdot Skw)$  in the transformed space. This example illustrates how the pair parameters included explicitly in the input vector of the LR model (3.4) are implicitly taken into account using a kernel method.

Moreover, the weights  $\mathbf{w} = (w_0, \dots, w_5)$  in (B2) are given by

$$w_0 = \sum_{i=1}^N a_i, \quad w_1 = \sum_{i=1}^N \sqrt{2}a_i x_{i,1}, \quad w_2 = \sum_{i=1}^N \sqrt{2}a_i x_{i,2}, \quad (\text{B3a-c})$$

$$w_3 = \sum_{i=1}^N a_i x_{i,1}^2, \quad w_4 = \sum_{i=1}^N a_i x_{i,2}^2, \quad w_5 = \sum_{i=1}^N \sqrt{2}a_i x_{i,1} x_{i,2}, \quad (\text{B4a-c})$$

which demonstrates that they depend on a linear combination of all the training data with non-zero expansion coefficients  $a_i$ . These coefficients depend on the cost function and can be obtained by using an adjoint formulation of the optimization problem. For a cost function composed of the sum of squares of the error with a regularization term, one can derive (see. e.g. Bishop 2007) the explicit dependence of  $a_i$  on the training data  $\mathbf{x}_i$  and the corresponding ground truth  $\Delta U_i^+$ . It becomes clear that large  $w_i$  indicates a high sensitivity to the corresponding input in training data set. For example,  $w_3 > w_2$  means that – in the training data set – a variation of  $ES_x^2$  results in a larger change in  $\Delta U^+$  compared with  $Skw^2$ .

### Appendix C. Parameters of rough surface database

**Table 5.** The range of friction-scaled primary parameters of five types of roughness.

	$Sk_0$	$Sk_+$	$Sk_-$	$\lambda_x$	$\lambda_z$
$k_c^+$	(43.84, 81.23)	(18.16, 44.40)	(34.29, 45.69)	(45.61, 81.10)	(49.40, 81.49)
$k_{rms}^+$	(5.85, 11.80)	(3.24, 6.44)	(5.51, 6.95)	(6.67, 12.85)	(7.08, 12.70)
$R_a^+$	(4.64, 9.46)	(2.58, 5.05)	(4.35, 5.49)	(5.32, 10.36)	(5.61, 10.12)
$Skw$	(-0.21, 0.19)	(1.42, 1.94)	(-1.78, -1.50)	(-0.34, 0.28)	(-0.22, 0.23)
$Ku$	(2.65, 3.47)	(4.26, 7.09)	(4.64, 6.09)	(2.57, 3.72)	(2.63, 3.60)
$ES_x$	(0.39, 0.86)	(0.18, 0.39)	(0.32, 0.42)	(0.33, 0.86)	(0.26, 0.62)
$ES_z$	(0.40, 0.88)	(0.19, 0.40)	(0.32, 0.43)	(0.24, 0.79)	(0.36, 0.80)
$Po$	(0.39, 0.57)	(0.87, 0.92)	(0.004, 0.112)	(0.37, 0.58)	(0.42, 0.59)
$inc_x$	(-0.11, 0.19)	(-0.11, 0.11)	(-0.07, 0.08)	(-0.41, 0.48)	(-0.90, 0.81)
$inc_z$	(-0.13, 0.18)	(-0.20, 0.22)	(-0.09, 0.11)	(-1.00, 0.84)	(-0.74, 0.46)

### References

- ANDERSSON, J., OLIVEIRA, D.R., YEGINBAYEVA, I., LEER-ANDERSEN, M. & BENSOW, R.E. 2020 Review and comparison of methods to model ship hull roughness. *Appl. Ocean Res.* **99**, 102119.
- BISHOP, C.M. 2007 *Pattern Recognition and Machine Learning (Information Science and Statistics)*, 1st edn. Springer.
- BRUEGEM, W.-P., VAN DIJK, V. & DELFOS, R. 2012 An efficient immersed boundary method based on penalized direct forcing for simulating flows through real porous media. In *ASME Volume 1: Symposia, Parts A and B Rio Grande, Puerto Rico, USA, 8-12 July*, pp. 1407–1416.
- BUSSE, A. & JELLY, T.O. 2023 Effect of high skewness and kurtosis on turbulent channel flow over irregular rough walls. *J. Turbul.* **24**, 57–81.
- CHAN, L., MACDONALD, M., CHUNG, D., HUTCHINS, N. & OOOI, A. 2015 A systematic investigation of roughness height and wavelength in turbulent pipe flow in the transitionally rough regime. *J. Fluid Mech.* **771**, 743–777.
- CHUNG, D., CHAN, L., MACDONALD, M., HUTCHINS, N. & OOI, A. 2015 A fast direct numerical simulation method for characterising hydraulic roughness. *J. Fluid Mech.* **773**, 418–431.
- CHUNG, D. & MCKEON, B.J. 2010 Large-eddy simulation of large-scale structures in long channel flow. *J. Fluid Mech.* **661**, 341–364.
- CHUNG, D., NICHOLAS, H., MICHAEL, P.S. & KAREN, A.F. 2021 Predicting the drag of rough surfaces. *Annu. Rev. Fluid Mech.* **53** (1), 439–471.

- CLAUSER, F.H. 1954 Turbulent boundary layers in adverse pressure gradients. *J. Aeronaut. Sci.* **21** (2), 91–108.
- CORTES, C. & VAPNIK, V. 1995 Support-vector networks. *Mach. Learn.* **20**, 273–297.
- COSTA, P., PHILLIPS, E., BRANDT, L. & FATICA, M. 2021 GPU acceleration of cans for massively-parallel direct numerical simulations of canonical fluid flows. *Comput. Maths Applics.* **81**, 502–511.
- DE MARCHIS, M., SACCONI, M., MILICI, D. & NAPOLI, E. 2020 Large eddy simulations of rough turbulent channel flows bounded by irregular roughness: advances toward a universal roughness correlation. *Flow Turbul. Combust.* **105**, 627–648.
- FLACK, K.A. & SCHULTZ, M.P. 2014 Roughness effects on wall-bounded turbulent flows. *Phys. Fluids* **16**, 101305.
- FLACK, K.A., SCHULTZ, M.P., BARROS, J.M. & KIM, Y.C. 2016 Skin-friction behavior in the transitionally-rough regime. *Trans. ASME J. Fluids Engng* **61**(A), 21–30.
- FOROOGHI, P., STROH, A., MAGAGNATO, F., JAKIRLIĆ, S. & FROHNAPFEL, B. 2017 Toward a universal roughness correlation. *Trans. ASME J. Fluids Engng* **139**, 121201.
- HAMA, F.R. 1954 Boundary layer characteristics for smooth and rough surfaces. *Trans. Soc. Nav. Archit. Mar. Engrs* **62**, 333–358.
- JACOBS, T.D.B., JUNGE, T. & PASTEWKA, L. 2017 Quantitative characterization of surface topography using spectral analysis. *Surf. Topogr.: Metrol. Prop.* **5**, 013001.
- JELLY, T.O. & BUSSE, A. 2018 Reynolds and dispersive shear stress contributions above highly skewed roughness. *J. Fluid Mech.* **852**, 710–724.
- JELLY, T.O. & BUSSE, A. 2019 Reynolds number dependence of Reynolds and dispersive stresses in turbulent channel flow past irregular near-gaussian roughness. *Intl J. Heat Fluid Flow* **80**, 108485.
- JELLY, T.O., RAMANIAND, A., NUGROHO, B., HUTCHINS, N. & BUSSE, A. 2022 Impact of spanwise effective slope upon rough-wall turbulent channel flow. *J. Fluid Mech.* **951**, A1.
- JIMÉNEZ, J. 2004 Turbulent flows over rough walls. *Annu. Rev. Fluid Mech.* **36**, 173–196.
- JOUYBARI, M.A., YUAN, J., BRERETON, J.G. & MURILLO, M.S. 2021 Data-driven prediction of the equivalent sand-grain height in rough-wall turbulent flows. *J. Fluid Mech.* **912**, A8.
- KADIVAR, M., TORMEY, D. & MCGRANAGHAN, G. 2023 A comparison of rans models used for cfd prediction of turbulent flow and heat transfer in rough and smooth channels. *Intl J. Thermofluids* **20**, 100399.
- KAJISHIMA, T., TAKIGUCHI, S., HAMASAKI, H. & MIYAKE, Y. 2001 Turbulence structure of particle-laden flow in a vertical plane channel due to vortex shedding. *JSME Intl J. Ser. B* **44** (4), 526–535.
- LEE, S., YANG, J., FOROOGHI, P., STROH, A. & BAGHERI, S. 2022 Predicting drag on rough surfaces by transfer learning of empirical correlations. *J. Fluid Mech.* **933**, A18.
- MACDONALD, M., CHUNG, D., HUTCHINS, N., CHAN, L., OOI, A. & GARCÍA-MAYORAL, R. 2017 The minimal-span channel for rough-wall turbulent flows. *J. Fluid Mech.* **816**, 5–42.
- SARAKINOS, S. & BUSSE, A. 2023 Reynolds number dependency of wall-bounded turbulence over a surface partially covered by barnacle clusters. *Flow Turbul. Combust.* **112**, 85–103.
- SCHULTZ, M.P. 2004 Frictional resistance of antifouling coating systems. *Trans. ASME J. Fluids Engng* **126**, 1039–1047.
- SHIN, H., KHORASANI, S.M.H., SHI, Z., YANG, J., BAGHERI, S. & LEE, S. 2024 Data-driven discovery of drag-inducing elements on a rough surface through convolutional neural networks. *Phys. Fluids* (in press).
- SMOLA, A.J. & BERNHARD, S. 2002 Introduction to support vector machines. In *Kernel Methods in Computational Biology*, pp. 21–45. Springer.
- THAKKAR, M., BUSSE, A. & SANDHAM, N. 2016 Surface correlations of hydrodynamic drag for transitionally rough engineering surfaces. *J. Turbul.* **18** (2), 138–169.
- THAKKAR, M., BUSSE, A. & SANDHAM, N.D. 2018 Direct numerical simulation of turbulent channel flow over a surrogate for nikuradse-type roughness. *J. Fluid Mech.* **837**, R1.
- YANG, J., STROH, A., CHUNG, D. & FOROOGHI, P. 2022 Direct numerical simulation-based characterization of pseudo-random roughness in minimal channels. *J. Fluid Mech.* **941**, A47.
- YANG, J., STROH, A., LEE, S., BAGHERI, S., FROHNAPFEL, B. & FOROOGHI, P. 2023a Prediction of equivalent sand-grain size and identification of drag-relevant scales of roughness – a data-driven approach. *J. Fluid Mech.* **975**, A34.
- YANG, J., STROH, A., LEE, S., BAGHERI, S., FROHNAPFEL, B. & FOROOGHI, P. 2024 Assessment of roughness characterization methods for data-driven predictions. *Flow Turbul. Combust.* **113** (2), 275–292.
- YANG, X.I.A., SADIQUE, J., MITTAL, R. & MENEVEAU, C. 2016 Exponential roughness layer and analytical model for turbulent boundary layer flow over rectangular-prism roughness elements. *J. Fluid Mech.* **789**, 127–165.
- YANG, X.I.A., ZHANG, W., YUAN, J. & KUNZ, R.F. 2023b In search of a universal rough wall model. *Trans. ASME J. Fluids Engng* **145** (10), 101302.