Feature Selection, Clustering, and Prototype Placement for Turbulence Data Sets

Matthew Barone* and Jaideep Ray† and Stefan Domino‡
Sandia National Laboratories§, Albuquerque, NM 87185

This paper explores unsupervised learning approaches for analysis and categorization of turbulent flow data. Single point statistics from several high-fidelity turbulent flow simulation data sets are classified using a Gaussian mixture model clustering algorithm. Candidate features are proposed, which include barycentric coordinates of the Reynolds stress anisotropy tensor, as well as scalar and angular invariants of the Reynolds stress and mean strain rate tensors. A feature selection algorithm is applied to the data in a sequential fashion, flow by flow, to identify a good feature set and an optimal number of clusters for each data set. The algorithm is first applied to Direct Numerical Simulation data for plane channel flow, and produces clusters that are consistent with turbulent flow theory and empirical results that divide the channel flow into a number of regions (viscous sub-layer, log layer, etc). Clusters are then identified for flow over a wavy-walled channel, flow over a bump in a channel, and flow past a square cylinder. Some clusters are closely identified with the anisotropy state of the turbulence, as indicated by the location within the barycentric map of the Reynolds stress tensor. Other clusters can be connected to physical phenomena, such as boundary layer separation and free shear layers. Exemplar points from the clusters, or prototypes, are then identified using a prototype selection method. These exemplars summarize the dataset by a factor of 10 to 1000. The clustering and prototype selection algorithms provide a foundation for physics-based, semi-automated classification of turbulent flow states and extraction of a subset of data points that can serve as the basis for the development of explainable machine-learned turbulence models.

I. Introduction

Recently, there has been significant research activity in development of fluid turbulence models using machine learning approaches. The models are typically trained using high-fidelity simulation data, from either Direct Numerical Simulation (DNS) or sufficiently resolved Large Eddy Simulation (LES). In machine-learned closure models for the Reynolds-Averaged Navier-Stokes (RANS) equations, the model may predict either the Reynolds stress [1], or a perturbation to a modeled Reynolds stress [2], given some set of input variables or, in the machine learning parlance, *features*.

Feature selection is a well-studied topic in machine learning. In the context of regression models using supervised learning, the aim of feature selection is to identify a set of inputs to a model that lead to optimal outputs. Often, we wish to choose the minimal set of key features that allow for accurate predictions within the domain of interest. In turbulence modeling studies to date, features for machine-learned models are often “hand-selected,” using domain knowledge to guide the selection.

Another context for feature selection is clustering (see Ref. [3], Chp. 14) of unlabeled data using unsupervised learning approaches. Unsupervised learning is the branch of machine learning that seeks to find structure in a dataset without recourse to labeling of data records (also called *examples*) and, thus, without the injection of exogenous information by an analyst. In this context, we seek a set of features that effectively divides turbulence data into distinct clusters. This clustering is performed with the aid of a metric (or *distance function*) defined in terms of flow-features. We wish the clusters to conform, as much as possible, to our human understanding of turbulence, which requires a judicious

*Principal Member of the Technical Staff, Aerosciences Department, AIAA Associate Fellow, `mbarone@sandia.gov`

†Distinguished Member of the Technical Staff, Extreme Scale Data Science and Analytics Department, `jairay@sandia.gov`

‡Principal Member of the Technical Staff, Computational Thermal and Fluid Mechanics Department, `spdomin@sandia.gov`

§Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy’s National Nuclear Security Administration under contract DE-NA0003525. This paper describes objective technical results and analysis. Any subjective views or opinions that might be expressed in the paper do not necessarily represent the ‘views of the U.S. Department of Energy or the United States Government.

selection of flow-features, and definition of the metric. This type of construct can be useful in turbulence modeling for the identification of canonical turbulent states. An example of such a state would be the logarithmic layer of a zero pressure gradient turbulent boundary layer. Similarly, in the analysis of turbulence models, this construct could be used to extract data points that conform to well-studied turbulent states, so that we can verify that the behavior of a data-driven turbulence model reproduces known relationships. This is similar to the commonly applied practice of calibration of turbulence model parameters to reproduce known behavior in canonical turbulent flows. Thus, unsupervised learning methods give us tools for automated classification of a data point (or example) as a member of a particular cluster of points with similar characteristics of the turbulent state. These tools can be useful for determining the completeness of a training data set; that is, whether the training data are sufficiently rich to produce models that are predictive for a given application. If a *test* example does not "belong" to a previously identified cluster, it is an indication that we may be extrapolating to a new region of physics not covered by the training data set.

Note that since clustering is performed based on a metric, the segregation of examples into subsets may not be clear-cut, and it is quite possible to find outliers that could plausibly be assigned to other clusters. This can considerably complicate the calibration or validation of turbulence models, as the training/validation data would be inconsistent with the turbulence/physics being studied. Thus we seek a subset of examples that avoid outliers, are *representative* of the cluster and can be tested, via theory, if the clusters map to canonical turbulent flows. We call these representative examples *prototypes*, or *exemplars*, and describe a method to identify them from a clustered dataset.

In the present study, we target the problem of clustering of turbulent flow datasets, followed by the selection of examples that could serve as prototypes for clusters (called *prototype placement*). As our first task, we apply clustering and feature selection algorithms to turbulent flow data to explore the feasibility and effectiveness of these algorithms. The candidate features are comprised of statistics that are typically available from DNS or LES data sets; most of the quantities are derived from invariants of the Reynolds-averaged strain rate, rotation rate, and turbulent stress tensors. We begin with a study of turbulent channel flow, since this is a relatively simple canonical flow which is well characterized and has already been classified into distinct regions by theoretical analysis and empirical observations. We apply a feature selection algorithm that wraps around a Gaussian mixture model (GMM) clustering algorithm to identify a feature set that effectively clusters the data, and that can be reconciled with our existing domain knowledge of the regions of turbulent channel flow. We then consider several other two-dimensional turbulent flows, re-applying the clustering algorithm to these flows in sequential fashion to find new clusters that the channel flow data did not identify. We then attempt to reconcile the results of the clustering with our understanding of turbulent flow physics.

In our second task, we formulate prototype placement as a set-cover problem [4] and test the efficacy of a solution algorithm for its applicability to turbulent datasets. We define figures of merit that measure the quality of a set of prototypes, allowing us to trade-off simplicity of representation (i.e., a small number of prototypes) against the fidelity of representation. These studies allow us to answer the following fundamental questions:

1) Is unsupervised learning feasible for clustering of turbulent flowfield data?
2) To what extent are the clustering results consistent with our existing understanding of turbulent flow?
3) Can a small set of representative examples be drawn from a cluster for further analysis? Is there a principled way of deciding on the size of the set of prototypes?

Turbulent datasets are large and unwieldy, often containing distinct regions with different turbulent processes. This work begins to lay a foundation for semi-automated isolation and classification of turbulent flow states, as well as the extraction of a representative subset of data records that could be subjected to further analysis in a tractable manner. These can serve as a basis for the development of explainable machine-learned turbulence models.

II. Clustering and Feature Selection Algorithms

The aim of the present work is to identify, from a set of candidate flow state features, a subset of features that successfully clusters turbulent flow states. While many metrics have been devised to compare the performance of clustering algorithms, defining success in a clustering application can be challenging. Here, we loosely define success as automated partitioning of the data into flow-field regions that reconcile with our human understanding of turbulent flow physics. We will rely heavily on the plane channel flow case to select parameters for the clustering and feature selection algorithms since, for this flow, where the single-point statistics vary only in a single coordinate direction, we have a good concept of how the flow should be divided into regions based on empirically supported theory.

To perform this unsupervised learning task, we use the Feature Subset Selection (FSS) algorithm described in Dy and Brodley[5]. The FSS algorithm is an example of a wrapper approach for unsupervised learning, where a search for optimal feature subsets is wrapped around a clustering algorithm. There are three tasks involved in the wrapper approach: the feature search, the clustering algorithm, and the feature subset evaluation [5].

For the search, we use a sequential forward search that starts with a small number of features (or no features) and adds one feature at a time. The feature that is added maximizes the evaluation criterion. This leads to a maximum complexity of the search that is $O(d^2)$ for d features, whereas an exhaustive search would need to evaluate 2^d possible feature subsets. The sequential search stops when adding more features no longer improves the evaluation criterion.

The clustering algorithm used here is the soft Gaussian mixture model algorithm. The GMM algorithm approximates the probability distribution function of the data using a finite mixture of multi-variate Gaussian distributions. The parameters of the Gaussian mixture are calculating by a maximum likelihood fit, using the Expectation-Maximization algorithm. We used the Matlab function `fitgmdist()` for this calculation. The kmeans++ algorithm [6] is used to seed the initial Gaussian component means. The problem of finding local optima is mitigated by replicating the fitting calculation forty times, and selecting the best fit from the replicates.

A data point can belong to more than one cluster, with a probability assigned to each point/cluster pair. The number of clusters, k, is an input parameter for the GMM algorithm. For each candidate feature subset considered, a sweep over number of clusters from 2 to k_{max} is performed in order to identify an optimal number of clusters. In this work we set $k_{max} = 9$. The metric used to evaluate the suitability of number of clusters is Bayes' Information Criterion (BIC). For our turbulence data sets, we observed that the BIC tended to favor large numbers of clusters (ten or more). Its value often increased rapidly with k initially, then more gradually as k became larger. We adopted a selection criterion that the BIC for $k + 1$ clusters, denoted $BIC(k + 1)$, must be greater than $BIC(k) + 2\sigma_k$, where σ_k is the standard deviation of BIC values over the ensemble of forty GMM trials using k clusters. In this way, we demand that the benefit of adding an additional cluster be clear relative to the variation in results over the trials for one less cluster.

We follow Dy and Brodley [5] and use the scatter-separability metric as the feature subset evaluation criterion. This metric is larger when the distances between samples within a cluster are small (low scatter) and when the cluster means are far apart (separated). The metric is also invariant to linear transformations of the features. Cross-projection is used to remove the bias of the scatter-separability metric towards larger feature sets, when comparing two feature sets of different size [5].

III. Candidate Features for Turbulent Flow

We must begin with some candidate features that are readily available, satisfy Galilean and rotational invariance, and preferably play some role in our current understanding of turbulence. In RANS, we typically have available the mean flow velocity, pressure, temperature, and density fields, the spatial gradients of these quantities, as well as the Reynolds stress as provided by the turbulence model. Ignoring temperature and density variations for the moment, we thus have statistical quantities in the form of two symmetric tensors: the strain rate and Reynolds stress tensors; and three vectors: the mean velocity vector, the vorticity vector, and the pressure gradient vector. The velocity vector and pressure gradient vector depend on the particular reference frame, *i.e.* they are not Galilean invariant, and so are not used to derive features. Invariant features can be constructed from the other quantities, as described in the following sections.

Barycentric Coordinates

The degree and type of anisotropy present in the turbulent stress is described by the barycentric map [7]. The normalized Reynolds stress anisotropy tensor is

$$b_{ij} = \frac{\overline{u_i' u_j'}}{2\bar{k}} - \frac{1}{3}\delta_{ij}. \tag{1}$$

The eigenvalues of the anisotropy tensor are first computed and ordered according to

$$\lambda_1 \geq \lambda_2 \geq \lambda_3 \tag{2}$$

The eigenvalues are then transformed to two coordinates within an equilateral triangle via a linear mapping:

$$C_1 = \lambda_1 - \lambda_2, \quad C_2 = 2\,(\lambda_2 - \lambda_3), \quad C_3 = 3\lambda_3 + 1 \tag{3}$$

$$x_B = C_1 x_1 + C_2 x_2 + C_3 x_3 \tag{4}$$

$$y_B = C_1 y_1 + C_2 y_2 + C_3 y_3 \tag{5}$$

Here, (x_1, y_1) are the two-dimensional coordinates of the "one-component" vertex of the triangle, (x_2, y_2) are the coordinates of the "two-component" vertex, and (x_3, y_3) are the coordinates of the "three-component" vertex. The limiting state coordinates C_1, C_2, C_3 are the weights of each of these three limiting states associated with the triangle vertices.

Typically, the barycentric map is plotted on an equilateral triangle with vertices at, for example, $(0, 0)$, $(1, 0)$, and $(0, \sqrt{3}/2)$. The vertices correspond to physical anisotropy states of turbulence. The right vertex corresponds to one-component turbulence, where turbulent fluctuations act only in one direction. This state is approximately achieved in the buffer layer of turbulent channel flow. The left vertex corresponds to the two-component axi-symmetric limit, where turbulent fluctuations are active in two directions, but not in the third. The top vertex corresponds to the homogeneous limit, where turbulent fluctuations are omnidirectional. Other physical situations can be described by lines in the barycentric map, such as axi-symmetric contraction, axi-symmetric expansion, and plane strain [7].

Angles Between Tensors and Vectors

The strain rate and Reynolds stress tensors are symmetric and, therefore, can be diagonalized to reveal a set of principal axes, defined by the set of orthonormal eigenvectors of the tensor. Thus, each tensor defines a coordinate system, and one can describe their relative orientation by describing the angular coordinates of one tensor in the other tensor's coordinate system. Three angles are required for this purpose. The same can be done to describe the orientation of a vector relative to a tensor, for which two angles are required. These angles can be useful features for classifying a turbulent flow state because they are coordinate-system invariant (at least for the quantities involving velocity gradients), and they are naturally scaled, $O(1)$ quantities.

This development follows the eigensystem ordering conventions of Tao *et al.*[8]. The eigenvalues of the mean strain rate tensor $\overline{S}_{ij}$ are ordered as α_s, β_s, γ_s, where α_s is the most extensive eigenvalue, β_s is the intermediate eigenvalue, and γ_s is the most contracting eigenvalue, such that $\alpha_s \geq \beta_s \geq \gamma_s$. Similarly, eigenvalues of the stress tensor τ are ordered as $\alpha_\tau \geq \beta_\tau \geq \gamma_\tau$. Note that the sign of the stress tensor in the current work is opposite that of Tao *et al.* [8], such that the present τ subscript corresponds to the $-\tau$ subscript in Tao *et al.*[8]. Denote by $\boldsymbol{\alpha_s}$, $\boldsymbol{\beta_s}$, and $\boldsymbol{\gamma_s}$ the unit vectors aligned with the eigenvectors associated with α_s, β_s, and γ_s, respectively. Denote by $\boldsymbol{\alpha_\tau}$, $\boldsymbol{\beta_\tau}$, and $\boldsymbol{\gamma_\tau}$ the unit vectors aligned with the eigenvectors associated with α_τ, β_τ, and γ_τ, respectively.

The relative orientation of the coordinate systems implied by two tensors is described by a collection of three angles : θ, ϕ, and ζ. θ is the angle from $\boldsymbol{\alpha_s}$ to $\boldsymbol{\alpha_\tau}$. ϕ is the angle between the projection of $\boldsymbol{\alpha_\tau}$ in the $\boldsymbol{\beta_s} - \boldsymbol{\gamma_s}$ plane and $\boldsymbol{\beta_s}$. The angles θ and ϕ determine the orientation of $\boldsymbol{\alpha_\tau}$ (and the $\boldsymbol{\beta_\tau} - \boldsymbol{\gamma_\tau}$ plane) in the strain rate tensor coordinate system. ζ is the angle between $\boldsymbol{\gamma_\tau}$ and projection of $\boldsymbol{\gamma_s}$ onto the $\boldsymbol{\beta_\tau} - \boldsymbol{\gamma_\tau}$ plane. Thus, ζ defines a rotation of the stress coordinate system about the $\boldsymbol{\alpha_\tau}$ axis.

When examining probability distributions of the angles between two tensors, one should use the form $P\{cos(\theta), \phi, \zeta\}$. This ensures that the resulting joint distribution is not biased by choice of angular coordinates. In other words, a random white noise velocity field should result in a uniform joint probability density function [8].

The orientation between a vector and one of the tensors can be described by two angular coordinates : θ and ϕ, with the same definitions as before. In this case the joint PDF should be computed from $P\{cos\theta, \phi\}$.

Given the tensors $\overline{S}_{ij}$ and $\overline{\tau}_{ij}$, and mean vorticity vector $\overline{\omega}_i$, we can calculate a number of angles that can be used to describe the flow state. These are listed in Table 1. Note that we do not include vectors such as the mean velocity vector and mean pressure gradient, since these quantities lead to features that violate Galilean invariance, and would call into question the applicability of the results to any general flowfield in an arbitrary inertial reference frame.

Directional Quantities	No. of Angles
$\overline{S}_{ij} : \overline{\tau}_{ij}$	3
$\overline{S}_{ij} : \overline{\omega}_i$	2
$\overline{\tau}_{ij} : \overline{\omega}_i$	2
Total number of angles	7

Table 1 Directional quantity pairs for which relative angles of orientation can be calculated.

Some of these angles can be assigned a physical interpretation. For example, the production term in the enstrophy evolution equation can be written in terms of the alignment of the principal strain directions with the vorticity. The intermediate strain eigenvector is often preferentially aligned with the vorticity, although the degree of alignment can vary depending on instantaneous turbulent structure [9]. Presumably, the alignment of the mean strain intermediate eigenvector and mean vorticity also vary depending on the location within an inhomogeneous turbulent flow.

Other angles contain information relevant for turbulence modeling. For example, the commonly invoked Boussinesq approximation assumes alignment of the turbulent stress with the local strain. Angles between the principal stress and strain directions give a measure of the degree to which this assumption is violated. We hypothesize that these angles may also help differentiate between different turbulent flow states.

Scalar Invariants

A general polynomial expression relating the Reynolds stress anisotropy tensor to the mean rate of strain and rate of rotation tensors is given in Pope [10]. This ten-term expression is

$$\boldsymbol{b} = \sum_{i=1}^{10} g_i(\lambda_1, \lambda_2, \ldots, \lambda_5)\, \boldsymbol{T}_i \tag{6}$$

The coefficients of this expansion, g_i, are functions of scalar invariants of the strain rate and rotation rate tensors.

$$\lambda_1 = \{\boldsymbol{S}^2\}, \quad \lambda_2 = \{\boldsymbol{W}^2\}, \quad \lambda_3 = \{\boldsymbol{S}^3\}, \quad \lambda_4 = \{\boldsymbol{W}^2\boldsymbol{S}\}, \quad \lambda_5 = \{\boldsymbol{W}^2\boldsymbol{S}^2\} \tag{7}$$

Truncated forms of Eq. 6 can be used, which contain a subset of terms. For example, Schmitt and Hirsch [11] used a four-term expansion, retaining only terms $\boldsymbol{T}_i$ that are linear or quadratic in $\boldsymbol{S}$ and $\boldsymbol{W}$. This enabled them to solve for the coefficients in Eq. 6 as algebraic expressions involving both the set of invariants λ_k, as well as additional invariants that involve the anisotropy tensor:

$$\eta_1 = \{\boldsymbol{b}^2\}, \quad \eta_2 = \{\boldsymbol{b}\boldsymbol{S}\}, \quad \eta_3 = \{\boldsymbol{b}\boldsymbol{S}\boldsymbol{W}\}, \quad \eta_4 = \{\boldsymbol{b}\boldsymbol{S}^2\}, \quad \eta_5 = \{\boldsymbol{b}\boldsymbol{W}^2\} \tag{8}$$

The set of five invariants, $\lambda_k, k = 1, \ldots, 5$, in addition to the five invariants, $\eta_k, k = 1, \ldots, 5$, together provide a set of ten scalar quantities that contain information on the local stress-strain relationship in a turbulent flow. The λ_k invariants only describe the mean velocity gradient tensor and, as such, do not directly contain any information about statistics of the turbulence itself. The η_k invariants are formed from combinations of the strain, rotation, and anisotropy tensors, and thus contain information about both the mean velocity gradients as well as the turbulent stress. We do not attach any particular physical significance to any of these invariants, except inasmuch as they describe the local stress-strain relationship in a manner that respects coordinate-system invariance.

In their dimensional form, these scalar invariants can vary over orders of magnitude from invariant to invariant within the same flow, or for the same invariant across different flows with varying length and time scales. They are much more useful as descriptors of a turbulent flow state when they are cast in non-dimensional form. The anistropy tensor can be put in non-dimensional form using the turbulence kinetic energy. The strain and rotation rate tensors can be non-dimensionalized using the local turbulence time scale k/ϵ. Usually, if we know the Reynolds stress, we know the turbulence kinetic energy as one half of its trace. We do not always have the turbulence dissipation rate, either because it could not be measured or because it was not calculated and saved in a DNS.

Given the Reynolds stress tensor and strain rate tensor, an estimate of the dissipation rate can be made, within the context of a linear one-term Boussinesq approximation of the stress-strain relationship:

$$\nu_T = C_\mu \frac{k^2}{\epsilon} \tag{9}$$

The eddy viscosity can be taken as the coefficient of the linear term in Eq. 6. In two dimensions, for example, this expression takes the form $\nu_T = \eta_2/\lambda_1$. In three dimensions, however, this expression is much more complicated and can involve degenerate cases [12]. We found that it led to non-smooth eddy viscosity fields when applied to a set of DNS flow-fields. Instead, one can use an estimate of an isotropic linear eddy viscosity [13]:

$$\nu_T = -\frac{k\, S_{ij} b_{ij}}{S_{ij} S_{ij}} \tag{10}$$

This expression led to smooth eddy viscosity fields. Although it suffers from the isotropy assumption, we need not judge its merit by its accuracy in estimating any particular viscosity coefficient; it may still be entirely useful to estimate dissipation rate for non-dimensionalization of turbulence features, without implying any such accuracy.

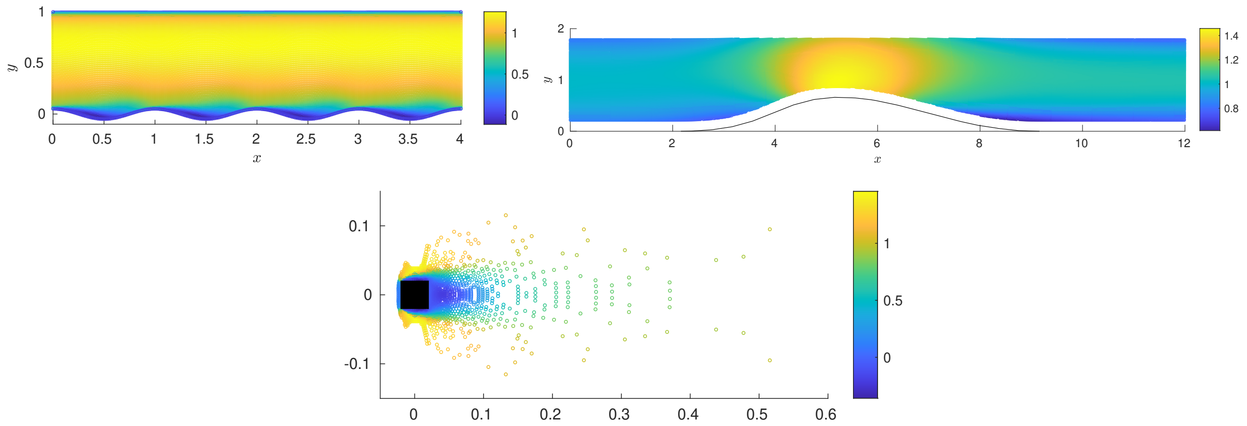

Fig. 1 Mean stream-wise velocity fields at sampled data points for the wavy wall, bump, and square cylinder flows.

Other Candidate Features

Other ad-hoc features may be useful for classification of turbulent states. The ratio of turbulent production to turbulent dissipation rate, $\mathcal{P}/\epsilon$ is often used to characterize turbulent flows. It is approximately equal to unity in the logarithmic layer of an equilibrium turbulent boundary layer, for example, and takes on other values in different regions of turbulent flows.

Summmary of Candidate Features

Table 2 lists the candidate features we have used in the present study. There are a total of twenty-one candidate features: the three limiting state barycentric map coordinates, seven tensor-tensor and tensor-vector angles, ten scalar invariants, and the ratio of turbulent production to dissipation rate. We consider only two-dimensional turbulent flows in the present study, with a homogeneous third dimension. In this case, only one of the angles is non-trivial - $\theta_{S-\tau}$ - and there are then only 15 features.

Feature Type	Features	Number of Features
Barycentric Map Coordinates	C_1, C_2, C_3	3
Tensor and Vector Angles	$\cos\theta_{S-\tau},\ \phi_{S-\tau},\ \zeta_{S-\tau},\ \cos\theta_{S-\omega},\ \phi_{S-\omega},\ \cos\theta_{\tau-\omega},\ \phi_{\tau-\omega}$	7
Scalar Invariants	$\lambda_1,\ \lambda_2,\ \lambda_3,\ \lambda_4,\ \lambda_5,\ \eta_1,\ \eta_2,\ \eta_3,\ \eta_4,\ \eta_5$	10
Ratio of Production to Dissipation	$\mathcal{P}/\epsilon$	1
Complete Set of Candidate Features	-	21

Table 2 The candidate features.

IV. Clustering Results

The present study analyzes data from four simple, two-dimensional flows: plane channel flow at five Reynolds numbers [14], a wavy-walled channel at a bulk Reynolds number of 6850 [15], a bump in a channel [16] at $Re_\tau \approx 600$, and a square cylinder in cross-flow at Reynolds number of 21,400. Figure 1 shows mean streamwise velocity fields for each of the latter three flows; these plots also illustrate the sampling of the fields, with one symbol plotted for each sampled data point.

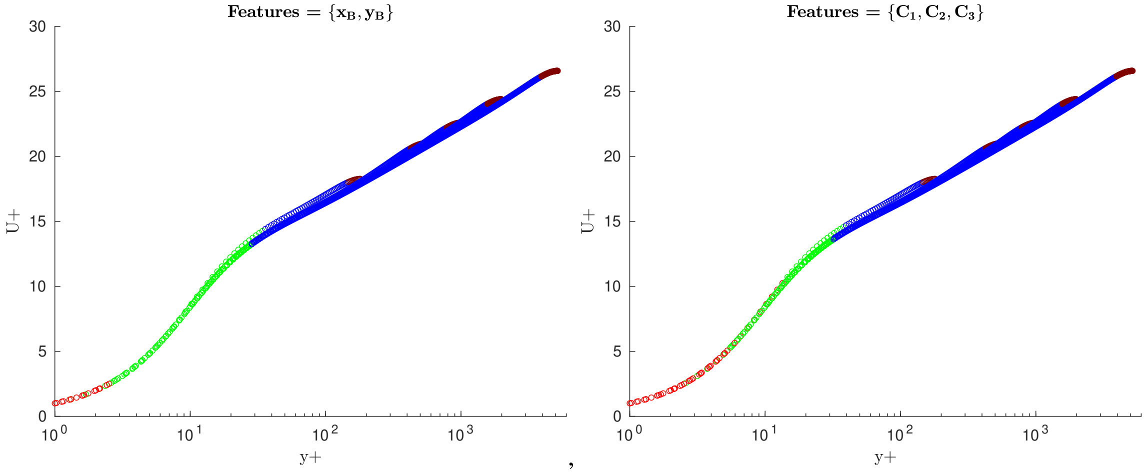

Fig. 2 Clustering results for turbulent channel flow. Left: Cartesian barycentric coordinate feature set. Right: Limiting state barycentric coordinate feature set.

Plane Channel Flow

We first investigate the performance of clustering and feature selection on a single flow: plane turbulent channel flow. Channel flow is one of the most well-studied and well-characterized turbulent flows, therefore serving as a good case with which to judge the machine learning techniques. In other words, the clustering results should be consistent with our existing understanding of this flow. If they are not, then it is likely not worth pursuing the application of these clustering techniques to more complex flow situations.

We apply the clustering algorithm to DNS data for plane channel flow at five different Reynolds numbers: $Re_\tau = 180, 550, 1000, 2000, 5200$ (as described in [14]). All of the proposed candidate features are available in this data set, including the dissipation rate for calculation of ratio of production to dissipation. We make the choice to populate the feature set with an initial set of features consisting of the barycentric coordinates, either in Cartesian form (x_B, y_B) or limiting state form (C_1, C_2, C_3). We first apply the GMM algorithm to cluster the data based only on the barycentric coordinates; the results are shown in Figure 2. The results are presented as mean velocity profiles, with individual data points colored by the most likely cluster for each point. The known regions of turbulent channel flow are easily identified from the mean velocity profiles: the viscous sublayer ($y+ \lessapprox 5$), the buffer layer ($5 \lessapprox y+ \lessapprox 30$), the logarithmic layer ($30 \lessapprox y+,\ y/H \lessapprox 0.15$), and the outer layer ($y/H \gtrapprox 0.15$). Accordingly, we choose initially to set the number of clusters equal to four. The barycentric coordinate feature sets allows for reasonable clustering of the data into four regions that approximately resemble the known regions of turbulent channel flow. However, there is a lack of sharp distinction between the viscous sublayer region and the buffer layer, and the Cartesian coordinates result in a viscous sublayer that ends slightly early, while the limiting state coordinates result in a viscous sublayer that is extends too far away from the wall.

We then ran the FSS algorithm to augment the feature set with further useful features, and to automatically find the optimal number of clusters. We found that we first needed to filter the data to remove anomalous outlier points. Certain tensor angle features were especially problematic. For example, the mean strain rate tensor becomes very small (theoretically zero) at the channel centerline, causing angles between the mean strain rate and Reynolds stress tensors to become ill-defined. We found that removing data points that had features which lay greater than eight standard deviations away from the mean value was sufficient. This resulted in only one percent of the data being eliminated. The optimal feature set identified by the algorithm includes four new features and five clusters; the complete optimal feature set is: $f_{ch} = \{C_1, C_2, C_3, \eta_1, \lambda_1, \eta_4, \eta_3\}$. The resulting clusters are shown in Figure 3. The clustering is remarkably consistent with our prior knowledge of turbulent channel flow regions. Note that no explicit information on distance from the wall, or the wall shear stress used for inner scaling, was provided to the algorithm. The boundaries of the regions are in approximately the correct locations, and the boundaries are "sharp," with little overlap between the clusters.

Upon initial examination, the selection of five clusters seems to be inconsistent with the conventional categorization of four regions of channel flow. However, the buffer region is in actuality simply a transition region between the viscous sublayer and the logarithmic layer, and we do not have much in the way of theory to suggest that the buffer region turbulence has uniform characteristics throughout. The clustering algorithm has split the buffer region into two regions,

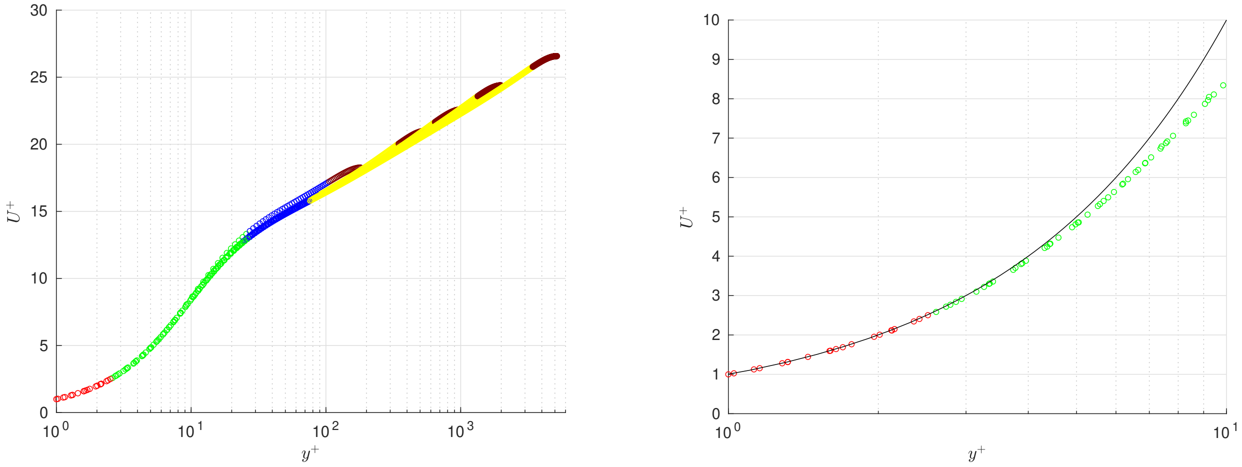

Fig. 3 Clustering results for turbulent channel flow with feature set including the limiting state barycentric coordinates and the invariant η_2. For reference the theoretical profile for the viscous sublayer is shown as a solid black line.

with the boundary at $y^+ \approx 25$. The algorithm has also placed the boundary between the viscous sublayer and the buffer regon at $y^+ \approx 2.5$, which seems to contradict the usual demarcation of $y^+ \approx 5$. Figure 3 shows, however, that the theoretical curve $u^+ = y^+$ for the mean velocity profile is accurate to within three percent at $y^+ = 5$, but begins to depart from the DNS closer to the wall. At $y^+ = 2.5$, the accuracy is close to one percent. Also, the viscous sublayer relationship describes the mean velocity profile and does not directly describe the turbulent state. We are re-assured that there is a distinct cluster associated with the inertial, or logarithmic layer, of the flow. Interestingly, the lowest Reynolds number case, at $Re_\tau = 180$, does not contain any points in this cluster. This is consistent with previous observations that channel flow does not exhibit a clear logarithmic layer at this low Reynolds number, while we also note that entirely precise identification of a logarithmic layer has remained somewhat elusive for DNS results to date.

Overall, the clustering produced by the FSS algorithm is consistent with our existing knowledge of channel flow turbulence, lending some confidence that this technique can group data points by a physical state that is, in some way, meaningful.

Wavy Wall

The next flow we consider is the flow in a plane channel with a wavy bottom wall. Again, as with all the flow cases considered in this paper, the geometry is two-dimensional with a homogeneous lateral dimension. We wish to completely utilize the clusters already obtained with the channel flow data, so we first assign clusters to the wavy wall data points that are well-classified by the channel flow clusters. We use the Mahalanobis distance d_M (see Ref. [3], Chp. 12) to determine how close each wavy wall data point is to each of the channel flow clusters. For the channel flow data, the vast majority of the data points satisfied $d_M < 25$, so we applied this threshold to the wavy wall data. The procedure followed was: 1) calculate the Mahalanobis distance from each wavy wall data point to each channel flow cluster, d_M^k; 2) determine the nearest channel flow cluster based on the minimum distance $\min\limits_k d_M^k$; 3) if the minimum distance satisfies $d_M^k < 25$, then the data point is assigned to the k^{th} channel cluster; 4) apply the FSS algorithm to the remaining unclassified wavy wall data points to identify new wavy wall clusters; 5) classify the remaining wavy wall data points using the new clusters.

Figure 4 shows the wavy wall data points that are well-classified by the channel flow clusters. Virtually the entire top half of the domain is well-classified by the channel flow clusters, as well as many points near, but not immediately adjacent to, the wavy wall surface. Many of the points associated with the region of flow separation downstream of the hump apexes are classified by the green-colored channel cluster that is associated with near-wall buffer layer points. This is due primarily to both regions being characterized by one-component anisotropy (observed also in [17]).

The FSS algorithm was run on the remaining data, again beginning with an initial feature set $\{C_1, C_2, C_3\}$, resulting in an optimal clustering with five clusters and an optimal feature set $f_{ww} = \{C_1, C_2, C_3, \eta_1\}$. The clusters for one periodic segment of the wavy wall domain are shown in Figure 5. There are two near-wall clusters which alternately appear

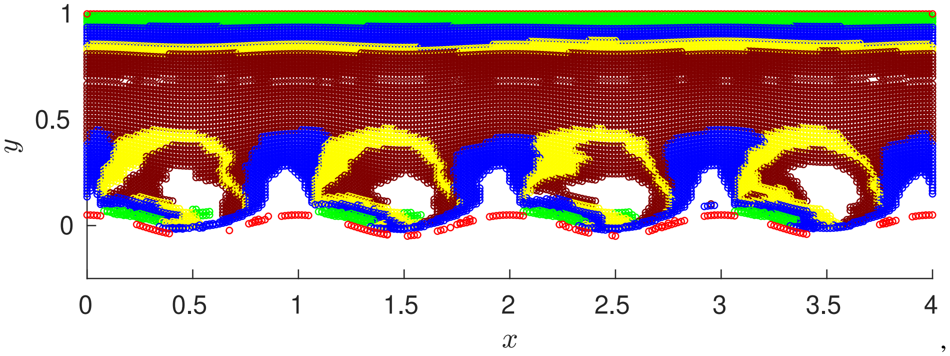

Fig. 4 Wavy wall data that is well-classified by the channel flow clusters, colored by cluster.

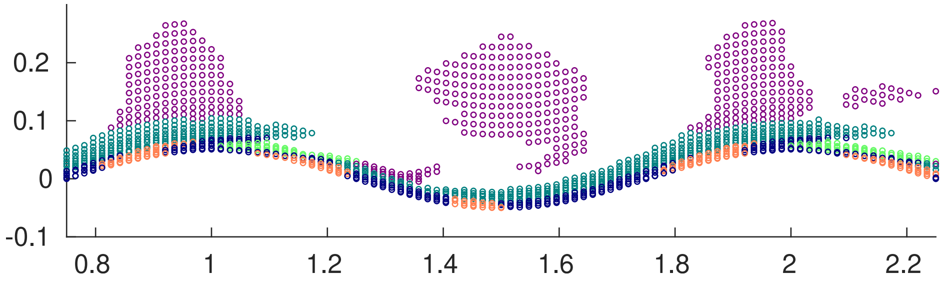

Fig. 5 Wavy wall clusters, showing only one periodic segment of the domain.

along the wall surface direction. Interestingly, these do not correspond to regions of pressure gradient sign. Two other clusters classify points lying just outside this near-wall region, while a fifth cluster classifies points yet further from the wall but still not well-categorized by the channel clusters.

Further visualizations reveal connections between the spatial distributions of the features and their connections to the clusters. For the wavy wall flow, the barycentric map coordinates play an important role. The primacy of the barycentric coordinates in determining the clusters for the wavy wall data is demonstrated in Figure 6. Here, the points are colored by position within the barycentric triangle, following [17]. The clusters from Figure 3 correspond closely to the position within the barycentric map. This is not surprising since only one additional feature, η_1, is employed in the wavy wall clustering. Contours of η_1 are shown in Figure 7. It appears that η_1 is useful in differentiating the "green" cluster from Figure 5, which describes points just downstream of the apex of the wavy surface, where the boundary layer has separated.

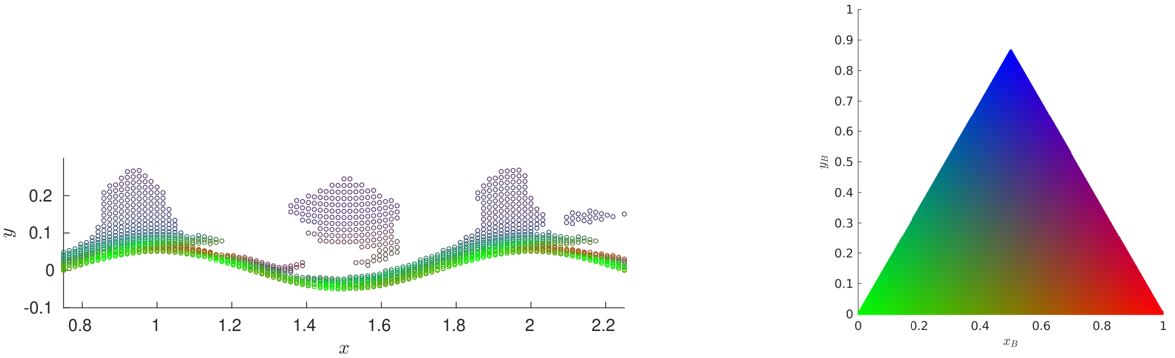

Fig. 6 Wavy wall clustered points, colored by position within the barycentric map.

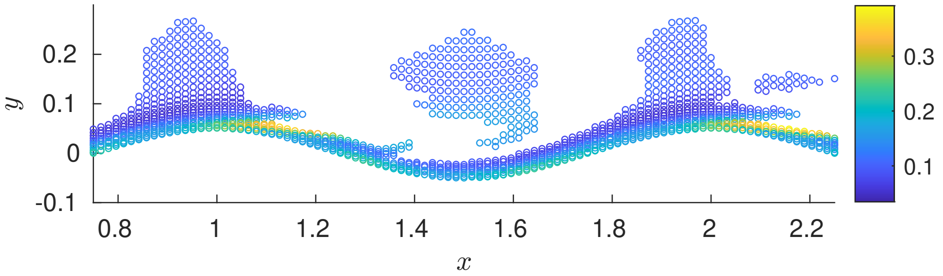

Fig. 7 Contours of scalar invariant η_1 for the wavy wall clustered points.

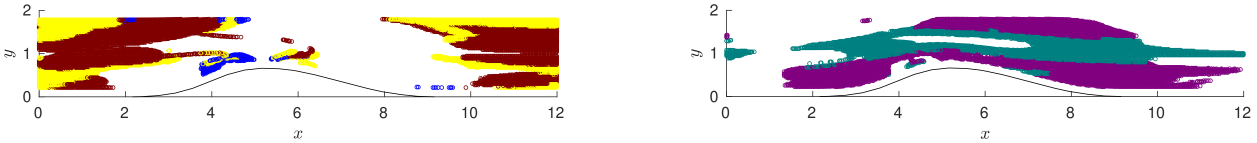

Fig. 8 Left: Bump in channel data that is well-classified by the channel flow clusters, colored by cluster. Right: Bump in channel data that is well-classified by the wavy wall clusters, colored by cluster.

Bump in Channel

The third flow considered is a two-dimensional bump in a channel. The data set for this flow did not include data near solid surfaces; the minimum distance from the closest point to a wall was about twenty percent of the channel half-width. The channel flow clusters were able to classify 46.3 % of the bump-in-channel, while the wavy wall clusters classified a further 49.8 %, leaving only 3.9 % of the data points left to be clustered. Figure 8 shows the channel flow clusters and wavy wall clusters for the bump flow, with the cluster colors consistent with previous figures. We might expect this situation, given the similarity in geometric configuration between the wavy wall and bump flows. Nonetheless, it is re-assuring that the previous clusters were good fits for much of the bump flow-field. The remaining points were run through the FSS algorithm, with an additional five clusters found as optimal, and the best feature set was $f_b = \{C_1, C_2, C_3, \cos\theta_{S-\tau}, \lambda_3\}$. Figure 9 shows these five clusters, with member points lying in a region above the bump but relatively distant from the surface. There is no obvious physical significance one can attach to these clusters. It is notable that the bump feature set is the first set to include the stress-strain tensor angle as well as one of the λ invariants.

Square Cylinder

The fourth flow considered is a square cylinder in cross-flow at a Reynolds number of 21,000 based on cylinder width*. Despite the significantly different flow topology relative to the previous three cases, a large number of points are well-classified by the existing clusters identified from those cases. In fact, the Mahalanobis distance threshold used for the square cylinder case was lowered from 25 to 15 to allow for an adequate number of data points to identify new clusters. With this threshold, 15 % of the square cylinder points were classified by channel flow clusters, while 63 % were classifed by the wavy wall points. These points are visualized in Figure 10. Only three points were classified by the bump clusters. The remaining points were grouped into three clusters by the FSS algorithm, with optimal feature set

*The LES data set used here is unpublished

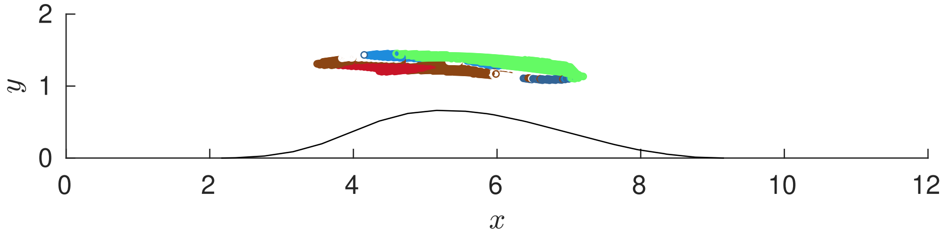

Fig. 9 Bump in channel data clusters.

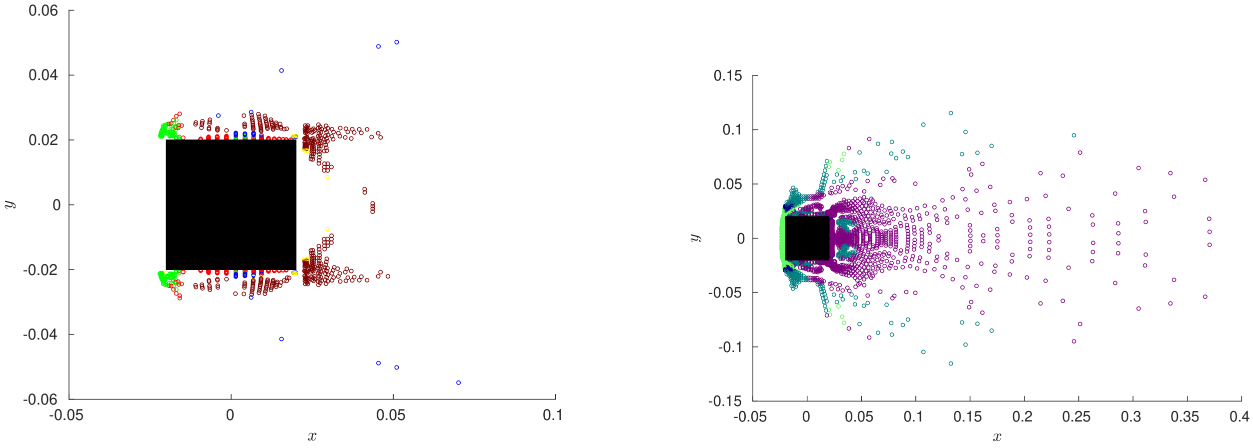

Fig. 10 Left: Square cylinder data that is well-classified by the channel flow clusters, colored by cluster. Right: Square cylinder in channel data that is well-classified by the wavy wall clusters, colored by cluster.

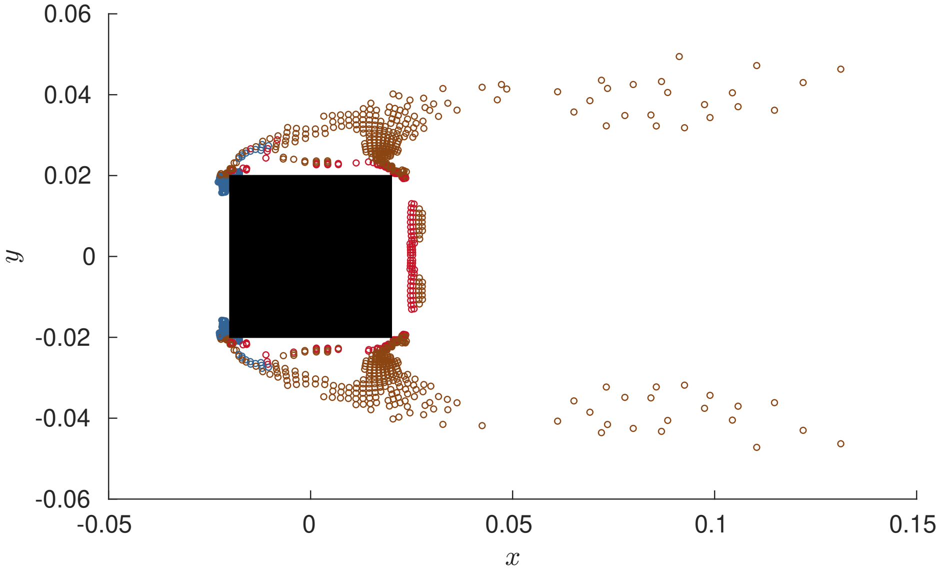

Fig. 11 Square cylinder data clusters.

$f = \{C_1, C_2, C_3, \theta_{S-\tau}\}$.

The square cylinder clusters are shown in Figure 11. One cluster is comprised of points mainly located around the forward corners of the cylinder, a high strain-rate region where the flow accelerates around the corner; note that the flow in this region has likely not fully transitioned to turbulence. A second cluster appears to be associated with the mean position of the shear layers above and below the cylinder, while the third cluster contains points in between the first two clusters.

V. Prototype Placement

Prototype placement (also known as selection of exemplars) is the selection of a subset of examples from a dataset that can adequately summarize it. It yields a tractably small dataset that can be used to interpret and understand a larger dataset. Prototype placement can be performed in an unlabeled dataset or in a dataset where the examples have been "colored", i.e., where they have been labeled/categorized via clustering, as in our case (see Sec. IV). Below, we describe some prototype placement approaches and apply them to our turbulence datasets, to extract a subset. The distribution of the prototypes provides an approximate measure of the variation of the features in space.

A. Prototype Placement

An example in a dataset is a n-dimensional vector ϕ. There are three different prototype placement scenarios, each with its own set of algorithmic solutions, *viz.* when features are not unlabeled but can be projected to a low-dimensional space, when they are unlabeled and intrinsically high-dimensional and lastly, when they are high-dimensional but can be labeled (or "colored"). If unlabled ϕ can be arranged in a matrix Φ where they form the columns, and if ϕ lie in a low-dimensional subspace, then the problem of isolating prototypes reduces to finding the smallest set of columns that yields the best-conditioned sub-matrix [18–23]. This can be performed using the Rank Revealing QR factorization of Φ [22] or via randomized, greedy algorithms [20, 21, 23, 24].

If ϕ are high-dimensional, unlabeled and do not lie on a low-dimensional manifold, then it is advantageous to define a similarity $s(i,k)$ between examples i and k and use it instead. Note that $s(i,k)$ need not be symmetric i.e., $s(i,k) \neq s(k,i)$ in general. Affinity Propagation [25] is an iterative algorithm that extracts prototypes using $s(i,k)$. It augments $s(i,k)$ with "responsibility" $r(i,k)$, an assessment of how well i may serve as k's prototype, given other potential prototypes of i, and "availability" $a(i,k)$, the ability of i to serve as k's prototype, given other examples that it "covers". Expressions to update $r(i,k)$ and $a(i,k)$ over the iterations are provided. These two constructs allow the placement of N_{proto} prototypes based on initial values of r and a. This method does not require multiple (random) initializations, like K-means [26] and K-mediods [27] do; in addition, it does not require one to guess the number of prototypes. In Ref. [28], the authors describe an algorithm that might be more suitable if one has a set of potential prototypes $\mathbb{X}$ for the dataset $\mathbb{Y}$. The algorithm reduces to finding a sparse subset of $\mathbb{X}$ based on the dissimilarities defined between the constituents of $\mathbb{X}$ and $\mathbb{Y}$. This is obviously more efficient if $\mathbb{X}$ is much smaller than $\mathbb{Y}$.

Ref. [29] describes an algorithm by which prototypes can be placed in a colored dataset and is the method used in our study. The method is meant for problems where (1) ϕ can be represented by a point in a high-dimensional continuous feature-space and (2) the ϕ can be "colored" or labeled by their cluster ID (signifying a particular type of turbulence that they represent). One starts with the assumption that all ϕ could potentially be prototypes. One chooses an example ϕ_i, of class l, and grows a sphere of radius ϵ to cover as many ϕ of the same class and as few ϕ of a different class. ϕ_i is then added to the set of prototypes for class l. ϕ_i may leave many ϕ uncovered by its sphere, and the next prototype is chosen in a greedy manner with penalty λ to cover, as far as possible, the uncovered ϕ. The algorithm stops when the greedy search can no longer find another prototype (of any class) that would maximize the quality of the set cover. The two user-defined inputs, ϵ and λ, control the number of prototypes N_{proto} that are placed and the quality of the cover/summarization that the prototypes achieve. A set of prototypes can incur three types of errors which ultimately define the quality of the coverage:

1) *Uncovered features:* A set of N_{proto} prototypes, with spheres of radius ϵ can leave a fraction u of examples uncovered. Optimal prototype selection will minimize u.
2) *Impure spheres:* A sphere of radius ϵ, centered on example ϕ_i, of class l, could cover examples of a different class, leading to its "impurity" $\bar{i}$ (expressed as a ratio of examples from classes other than l that the sphere covers). As impurity is undesirable, the sum $b = u + \bar{i}$ is a measure of the quality of the selected prototypes.
3) *Misclassification rate:* A set of prototypes, along with ϵ, can serve as a nearest-neighbor classifier of the data, and forms a second measure of the quality of a set of prototypes. The performance of this classifier can be quantified as its misclassification or error rate m.

B. Prototype Placement Procedure

The process of placing prototypes essentially reduces to determining a value for (ϵ, λ) that delivers a desired quality and level of summarization (as quantified by (b, m, N_{proto})) of a dataset. The user specifies a desired quality level (b^*, m^*) and one searches for the "optimal" (ϵ^*, λ^*) that can deliver it. Upper and lower bounds are specified in the $\epsilon - \lambda$ space and we draw 200 samples of (ϵ, λ) from it in a space-filling manner. The search for (ϵ^*, λ^*) is conducted via seven-fold cross-validation. Computations are performed in the R statistical framework using the package protoclass [30] for prototype placement and randtoolbox [31] for random sampling.

The first step in prototype placement involves balancing the dataset, since the sizes of the clusters vary widely. Equal numbers of examples of each class are drawn from the dataset to constitute a new dataset for the search. The features are centered (each component of ϕ has its mean subtracted from it) and scaled (each component of ϕ is divided by the standard deviation, so that the values vary, approximately, between -3 and 3). This dataset is then randomly divided into seven folds. We iterate through the 200 (ϵ, λ) samples. For a given (ϵ, λ)-pair, we designate one of the folds as the "testing" fold whereas the rest are the "learning" folds. Prototypes are placed, using protoclass, in the "learning" folds and (b, N_{proto}) computed from the placement. The prototypes, with their ϵ, are next used as a nearest-neighbor classifier

Cluster	Re_τ = 180	Re_τ = 550	Re_τ = 1000	Re_τ = 2000	Re_τ = 5200
Outer layer (black)	26 (4)	45 (3)	59 (2)	82 (1)	174 (1)
Log-layer (magenta)	0 (0)	73 (1)	129 (1)	232 (1)	525 (0)
Buffer layer, upper part (red)	35 (3)	35 (3)	29 (2)	28 (1)	33(1)
Buffer layer, lower part (blue)	23 (11)	24 (13)	24 (8)	26 (5)	28 (3)
Viscous sublayer (brown)	5 (3)	15 (3)	15 (3)	16 (3)	8 (1)
N_{proto}	21	23	16	11	9
$(u^*, \bar{i}^*)$	(0%, 0%)	(0.5%, 4.7%)	(0.4, 6.2%)	(1.3%, 5.5%)	(2%, 6%)

Table 3 Cluster sizes (number of examples) and number of prototypes (in parentheses) for the 5 plane channel flow problems considered in this study. All flows have been segregated into 6 clusters, whose sizes are tabulated; the figures in the parentheses are the number of prototypes placed in the cluster. The last row of the table provides the fraction of examples left uncovered by the prototypes and the average impurity of the spheres grown at the prototypes.

to classify the data in the "testing" fold to compute m. We iterate through the seven folds to compute seven different (b, N_{proto}, m) values; these are averaged to serve as the performance figures for the (ϵ, λ)-pair. 200 such performance figures-of-merit are computed, and (b, m) plotted as a function of N_{proto}. The (ϵ, λ)-pair that yields (b, m) closest to (b^*, m^*) is designated the "optimal" (ϵ^*, λ^*) result.

The final prototype placement is performed by configuring `protoclass` with (ϵ^*, λ^*) and applying to the full dataset; $(u^*, \bar{i}^*)$ are computed as a measure of the quality of final set of prototypes.

C. Channel Flow Results

As a first step, we apply prototype selection to the plane channel flow results in Sec. IV for $\mathrm{Re}_\tau = 180, 550, 1000, 2000$ and 5200. Here we set $\phi = f_{ch}$ and perform prototype placement one Re_τ at a time. Figure 12 (left) shows the computation of (b, m, N_{proto}) via 7-fold cross-validation (CV), as described in Sec. V.B, performed for the $\mathrm{Re}_\tau = 1000$ dataset. We plot the "bad coverage" error b and the misclassification error m as a function of N_{proto}, the number of prototypes chosen, as we iterate through the (ϵ, λ) samples. The horizontal line shows the desired values (b^*, m^*) leading to a (ϵ^*, λ^*) that results in $N_{proto} = 16$. Similar analyses were run for all the other channel-flow cases, to calculate dataset-specific (ϵ^*, λ^*) configurations.

In Figure 12 (right), we plot the mean velocity profiles from the channel flow cases, colored by the cluster ID. The open circles denote the prototypes, placed for all the velocity profiles. As is clear, the number of prototypes changes from flow to flow, and their locations on the velocity profile are not the same. In Table 3 we tabulate the cluster sizes and the number of prototypes chosen in each cluster, as a measure of the summarization achieved by the prototypes. $(u^*, \bar{i}^*)$ are also stated, as a measure of the quality of the summarization by the prototypes. We see that geometrically large regions/clusters with many grid points need not necessarily have many prototypes. This is because spheres are grown (when computing prototypes) after ϕ has been centered and scaled i.e., the spheres, mapped back to physical space are highly skewed and irregular. In addition, the quality of the summarization by the prototypes is quite variable, e.g., for $\mathrm{Re}_\tau = 5200$, no prototype was found in the log-layer (see Table 3). Further, apart from the low Re_τ case, all the flows have approximately the same number of prototypes placed in them. The summarization is also seen to become more efficient with Re_τ. This could be due, in part, to the improved separation of the log layer from the inner and outer layers as the Reynolds number increases. For each case with a log layer present, a single prototype is placed, consistent with the self-similar nature of the turbulence in this region.

D. Complex Flow Results

We now apply the placement of prototypes to somewhat more complex flows, viz., flow over a wavy wall. The feature set $\phi = f_{ww}$ is used for placement of prototypes. Figure 13 (left) plots the cluster sizes and the number of prototypes placed in those cluster. We see that there are 10 clusters whose sizes vary over a factor of 10. In Figure 13

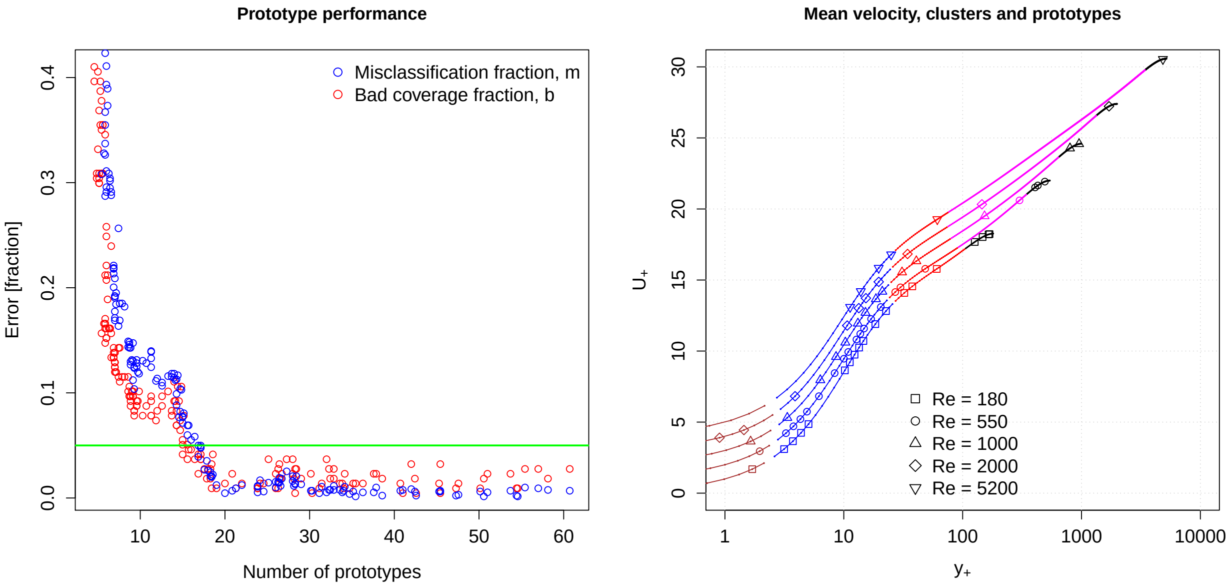

Fig. 12 **Left: (b, m) plotted against N_{proto} for plane channel flow $\mathrm{Re}_\tau = 1000$ case as we search through 200 (ϵ, λ) samples for those close to the desired value (plotted in green). We see that approximately 16 prototypes can summarize this dataset. Right: Plots of the mean velocity profile, colored by their cluster, for $\mathrm{Re}_\tau = 180, 550, 1000, 2000$ and 5200. Prototypes are plotted with open symbols and have the same color as the cluster they summarize. For $\mathrm{Re}_\tau = 5200$ no prototypes were found in the log-layer (magenta points); in addition, the prototype in the viscous sublayer (brown points) is positioned at $y_+ < 1$ and is not visible in the plot. Note: The velocity profiles have been shifted vertically to make them legible. The various parts of a turbulent boundary layer can be read off using the horizontal axis (y^+).**

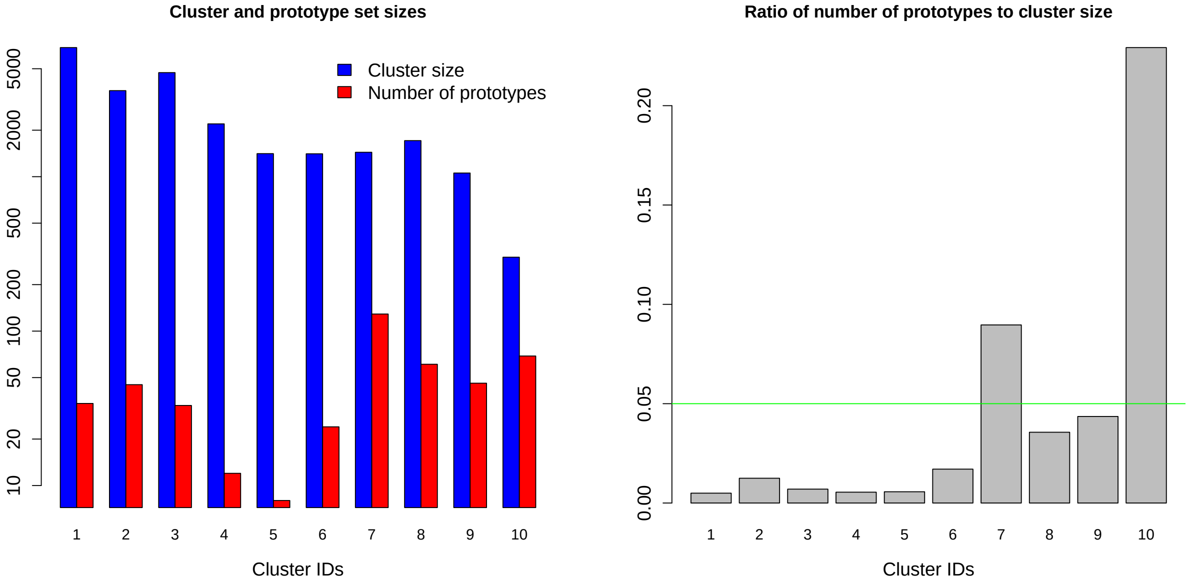

Fig. 13 Left: Comparison of the size of the 10 clusters in the wavy-wall dataset and the prototypes placed in them. Right: Ratio of the number of prototypes and the cluster sizes. Except for the smallest cluster (cluster ID 10), we obtain a reduction of 10X – 20X with the prototypes.

(right) we plot the ratio of the number of prototypes and the cluster size, as a measure of the summarization obtained by prototypes; it is clear that for all but the smallest cluster, prototypes form 5% or less of the cluster. The dataset has 461 prototypes, with $u = 20\%$ and $\bar{i} = 9.7\%$.

In Figure 14 (left) we plot the flow-over-a-wavy-wall dataset colored by the cluster ID. The prototypes are plotted with symbols of the same color. Not surprisingly, most of the cluster and prototypes are found near the wall where high gradients of the turbulent statsitics exist. The periodic structure of the wall cause periodicity of the blue and red clusters. These distinct clusters have the same color because in the ϕ-space they occupy a contiguous region (they are identical, from a turbulent processes point of view). The splitting of a contiguous region in feature space, as seen in this figure, is one of the reasons for using prototypes to summarize a turbulent dataset, rather than cluster centers. Prototypes, being examples drawn from the dataset, can be mapped between ϕ-space and physical space trivially, which simplifies their fluid-dynamical interpretation. In contrast, while averaging the geometrical locations of, say, the points in the blue cluster is easy (to yield the geometrical coordinates of the cluster center), mapping it back to physical space places it in the red cluster. In Figure 14 (right), we stretch the vertical axis with a logarithmic transform to expose the near-wall region. Some of the thinner clusters are more visible in this figure, as are the prototypes placed in them.

Finally, we apply the prototype placement algorithm on a dataset from the simulation of flow over a bump. The feature set used is $\phi = f_b$. As seen in Figure 15 (left), there are 11 clusters, whose sizes vary by two orders of magnitude (cluster 16 versus cluster 6). In addition, many of the physics (cluaters) seen in the previous flows (channel flow and wavy wall) are not seen here; cluster IDs 3, 5, 8, 9 and 10 are missing. The ratio of the number of prototypes placed in a cluster to the cluster size are plotted with gray bars; the prototypes vary between 10% to 0.1% of the cluster. The quality of the cover the prototypes provide is given by $(u, \bar{i}) = (18\%, 15\%)$; $N_{proto} = 139$.

Figure 15 (right) plots the clustering on either side of the bump and the placement of the prototypes. Turbulent flows in regions with favorable and unfavorable pressure gradients are clearly seen, as well at the region above the tip of the bump, where the flow quickly changes character. We see that the prototypes are *not* uniformly distributed in physical space, indicating severe contortions of the clusters as they mapped between physical and ϕ-space where clustering and prototype-pacement is performed.

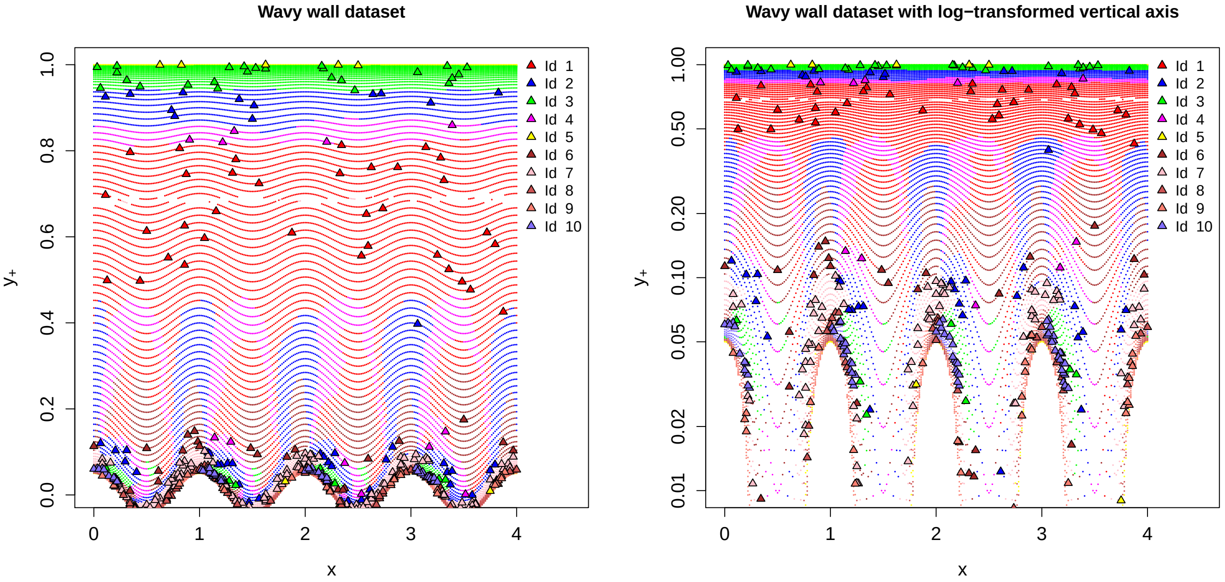

Fig. 14 Left: The flow-over-a-wavy-wall dataset with points clustered by their cluster ID. Prototypes are plotted with symbols. Right: The same figure but with the vertical axis log-transformed to illustrate the clustering (and prototype placement) in the near-wall region.

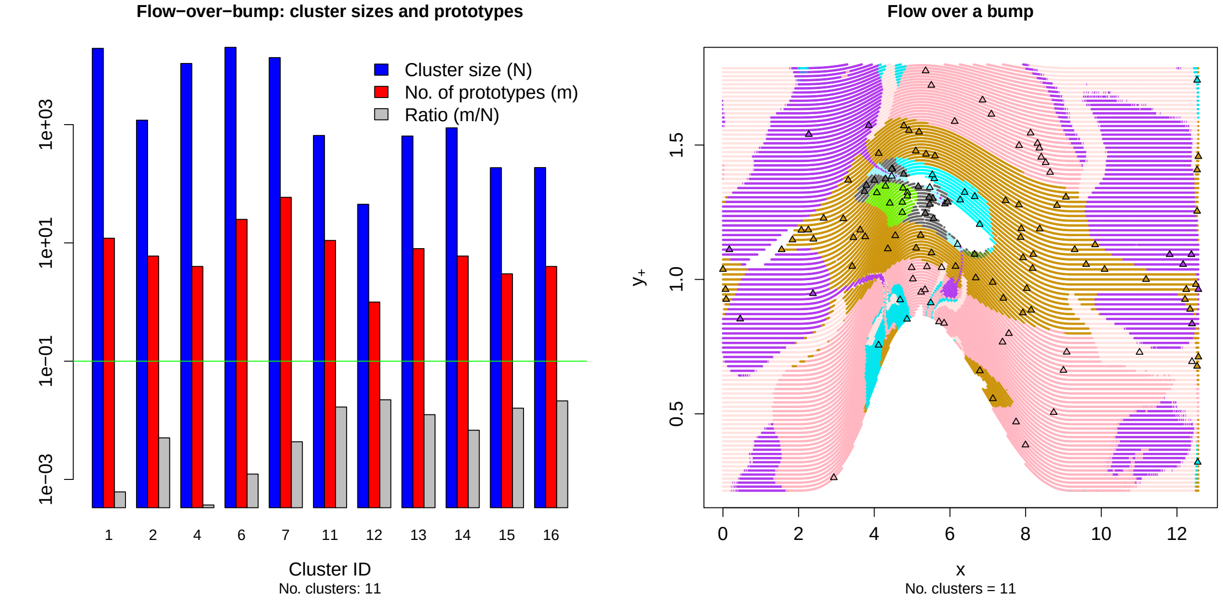

Fig. 15 Left: Cluster sizes, prototypes and the degree of summarization obtained. We see that there are 10X – 100X fewer prototypes than examples. Right: The flow-field segregated into clusters. There are 11 clusters in all.

VI. Conclusions

In this paper we have explored the premise that turbulent states can be succesfully classified, or categorized, using unsupervised machine learning techniques. We have proposed a number of candidate features based on typically available (at least in simulations) single-point turbulence statistics. We applied feature selection and clustering techniques to the "featurized" data in a sequential, flowfield-by-flowfield fashion, resulting in a classification of the data into a number of clusters. The clusters can be regarded as a collection of data points that are similar in the feature space. In some cases, with channel flow providing the best example, we can identify the different clusters with our existing theoretically and empirically derived understanding of turbulent flows. Furthermore, we have demonstrated how exemplar data points, or prototypes, can be automatically selected from the clusters. The set of exemplars, which summarizes the flow dataset, is $10\times - 1000\times$ smaller that the dataset it represents.

To close, we mention some of the use cases for clustering and prototype selection of turbulence data. Our own motivation is to use the resulting prototypes to facilitate the creation of explainable neural network turbulence models. The prototypes give a relatively small number of data points, which can be considered representative of a certain set of turbulent flow physics. We can probe the behavior of the neural network in the vicinity of the prototypes, and test whether the neural network predictions are consistent with desired model behavior for the identified physical situation. There are other potential use cases for our classification, one of which is determination of model extrapolation beyond a training dataset. Again we are considering the case where a machine-learned turbulence model has been trained using a training data set, and then required to make a prediction at a new point. If the new point clearly belongs to one of the clusters identified in the training data, then the model will likely make a valid prediction. If the new point is not well-classified by one of the clusters in the training data, this would indicate the model may not give a valid prediction, and new training data are required. Further studies are certainly required to explore the full utility of the approach, using a greater variety of turbulence data and including three-dimensional flow-fields. We conclude by noting that the present approach is not necessarily limited to single-point, time-mean features but could, in principle, be applied to any invariant statistical quantities one may choose to define a turbulent state.

References

[1] Ling, J., Kurzawski, A., and Templeton, J., "Reynolds averaged turbulence modelling using deep neural networks with embedded invariance," *J. Fluid Mech.*, Vol. 807, 2016, pp. 155–166.

[2] Wang, J.-X., Wu, J.-L., and Xiao, H., "Physics-informed machine learning approach for reconstructing Reynolds stress modeling discrepancies based on DNS data," *Phys. Rev. Fluids*, Vol. 2, 2017.

[3] Hastie, T., Tibshirani, R., and Friedman, J., *The Elements of Statistical Learning*, 2^{nd} ed., Springer, 233 Sprint Street, New York, NY, 2008.

[4] Wikipedia contributors, "Set cover problem — Wikipedia, The Free Encyclopedia," `https://en.wikipedia.org/w/index.php?title=Set_cover_problem&oldid=977797135`, 2020. [Online; accessed 31-October-2020].

[5] Dy, J. G., and Brodley, C. E., "Feature Selection for Unsupervised Learning," *J. of Machine Learning Research*, Vol. 5, 2004, pp. 845–889.

[6] Arthur, D., and Vassilvitskii, S., "K-means++: The Advantages of Careful Seeding," *SODA '07: Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, 2007, p. 1027–1035.

[7] Banerjee, S., Krahl, R., Durst, F., and Zenger, C., "Presentation of anisotropy properties of turbulence, invariants versus eigenvalue approaches," *J. of Turbulence*, Vol. 8, No. 32, 2007.

[8] Tao, B., Katz, J., and Meneveau, C., "Statistical geometry of subgrid-scale stresses determined from holographic particle image velocimetry measurements," *J. Fluid Mech.*, Vol. 457, 2002, pp. 35–78.

[9] Buchner, A.-J., Lozano-Durán, A., Kitsios, V., Atkinson, C., and Soria, J., "Local topology via the invariants of the velocity gradient tensor within vortex clusters and intense Reynolds stress structures in turbulent channel flow," *2nd Multiflow Summer School on Turbulence, J. of Physics: Conference Series*, Vol. 708, 2016.

[10] Pope, S. B., "A more general effective-viscosity hypothesis," *J. Fluid Mech.*, Vol. 72, No. 2, 1975, pp. 331–340.

[11] Schmitt, F., and Hirsch, C., "Experimental study of the constitutive equation for an axisymmetric complex turbulent flow," *Z. Angew. Math. Mech.*, Vol. 80, 2000, pp. 815–825.

[12] Jongen, T., and Gatski, T. B., "General explicit algebraic stress relations and best approximation for three-dimensional flows," *Int. J. Engineering Science*, Vol. 36, 1998, pp. 739–763.

[13] Ling, J., and Templeton, J., "Evaluation of machine learning algorithms for prediction of regions of high Reynolds averaged Navier Stokes uncertainty," *Phys. Fluids*, Vol. 27, 2015.

[14] Lee, M. K., and Moser, R. D., "Direct numerical simulation of turbulent channel flow up to $Re_\tau \approx 5200$," *J. Fluid Mech.*, Vol. 774, 2015, pp. 395–415.

[15] Gorlé, C., Emory, M., Larsson, J., and Iaccarino, G., "Epistemic uncertainty quantification for RANS modeling of the flow over a wavy wall," Center for Turbulence Research Annual Research Briefs, 2012.

[16] Marquillie, M., Ehrenstein, U., and Laval, J. P., "Instability of streaks in wall turbulence with adverse pressure gradient," *J. Fluid Mech.*, Vol. 681, 2011, pp. 205–240.

[17] Emory, M., and Iaccarino, G., "Visualizing turbulence anisotropy in the spatial domain with componentality contours," Center for Turbulence Research Annual Research Briefs, 2014.

[18] Elhamifar, E., Sapiro, G., and Vidal, R., "See all by looking at a few: Sparse modeling for finding representative objects," *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 1600–1607. 10.1109/CVPR.2012.6247852.

[19] Esser, E., Moller, M., Osher, S., Sapiro, G., and Xin, J., "A Convex Model for Nonnegative Matrix Factorization and Dimensionality Reduction on Physical Space," *IEEE Transactions on Image Processing*, Vol. 21, No. 7, 2012, pp. 3239–3252. 10.1109/TIP.2012.2190081.

[20] Tropp, J. A., "Column Subset Selection, Matrix Factorization, and Eigenvalue Optimization," *Proceedings of the 2009 Annual ACM-SIAM Symposium on Discrete Algorithms*, 2009, pp. 978–986. 10.1137/1.9781611973068.106, URL https://epubs.siam.org/doi/abs/10.1137/1.9781611973068.106.

[21] Boutsidis, C., Mahoney, M. W., and Drineas, P., "An Improved Approximation Algorithm for the Column Subset Selection Problem," *Proceedings of the 2009 Annual ACM-SIAM Symposium on Discrete Algorithms*, 2009, pp. 968–977. 10.1137/1.9781611973068.105, URL https://epubs.siam.org/doi/abs/10.1137/1.9781611973068.105.

[22] Chan, T. F., "Rank revealing QR factorizations," *Linear Algebra and its Applications*, Vol. 88-89, 1987, pp. 67 – 82. https://doi.org/10.1016/0024-3795(87)90103-0, URL http://www.sciencedirect.com/science/article/pii/0024379587901030.

[23] Balzano, L., Nowak, R., and Bajwa, W., "Column subset selection with missing data," *NIPS Workshop Low-Rank Methods Large-Scale Mach. Learn.*, 2010. URL http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.295.9188&rep=rep1&type=pdf.

[24] Bien, J., Xu, Y., and Mahoney, M. W., "CUR from a sparse optimization viewpoint," *Proc. Adv. Neural Inf. Process. Syst.*, 2010. URL http://papers.nips.cc/paper/3890-cur-from-a-sparse-optimization-viewpoint.

[25] Frey, B. J., and Dueck, D., "Clustering by Passing Messages Between Data Points," *Science*, Vol. 315, No. 5814, 2007, pp. 972–976. 10.1126/science.1136800, URL https://science.sciencemag.org/content/315/5814/972.

[26] Duda, R., Hart, P., and Stork, D., *Pattern Classification*, Wiley-Interscience, Hoboken, NJ, USA, 2004.

[27] Kaufman, L., and Rousseeuw, P., *Statistical Data Analysis Based on the L1 Norm and Related Methods*, North-Holland, Amsterdam, The Netherlands, 1987. URL https://www.researchgate.net/profile/Peter_Rousseeuw/publication/243777819_Clustering_by_Means_of_Medoids/links/00b7d531493fad342c000000.pdf.

[28] Elhamifar, E., Sapiro, G., and Sastry, S. S., "Dissimilarity-Based Sparse Subset Selection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 38, No. 11, 2016, pp. 2182–2197. 10.1109/TPAMI.2015.2511748.

[29] Bien, J., and Tibshirani, R., "Prototype selection for interpretable classification," *Annals of Applied Statistics*, Vol. 5, No. 4, 2011, pp. 2403–2424. 10.1214/11-AOAS495, URL https://doi.org/10.1214/11-AOAS495.

[30] Bien, J., and Tibshirani, R., *protoclass: Interpretable classification with prototypes*, 2013. URL https://CRAN.R-project.org/package=protoclass, r package version 1.0.

[31] Christophe, D., and Petr, S., *randtoolbox: Generating and Testing Random Numbers*, 2019. R package version 1.30.0.