SAND2022-5573C

Sandia National Laboratories

Exceptional service in the national interest

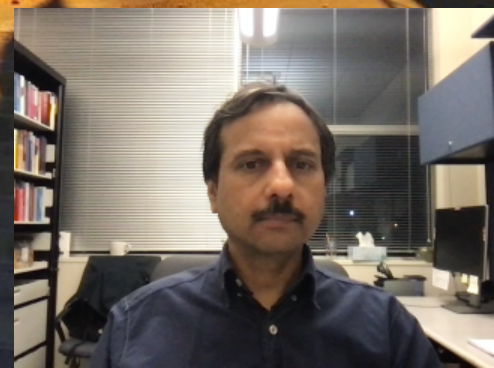
Qualifying Training Datasets for Data-Driven Turbulence Closures

T. Banerjee, J. Ray*, M. Barone & S. Domino

AIAA AVIATION 2022, CHICAGO, IL

SESSION: FD-35

June 30, 2022



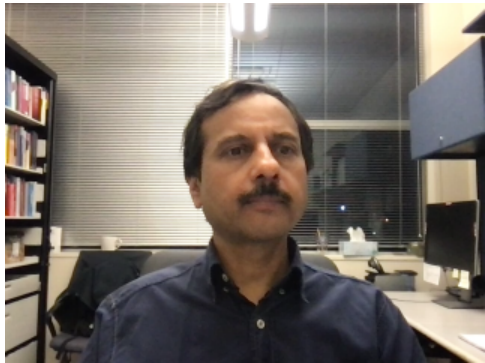
Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia LLC, a wholly owned subsidiary of Honeywell International Inc. for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

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.

Introduction

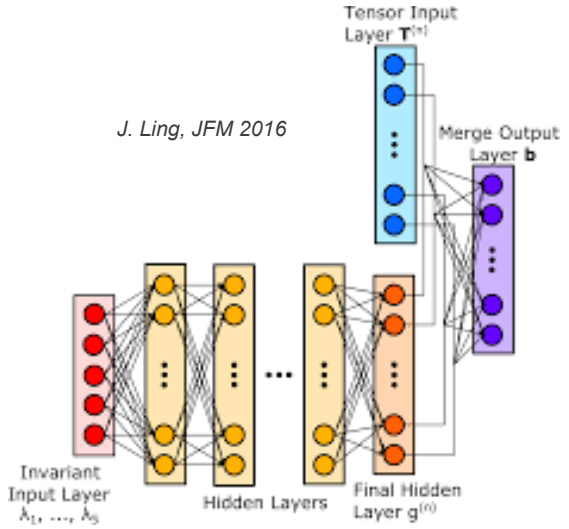
- Aim: Devise a way to qualify training datasets (TD) used for machine-learned (ML) turbulence closures
  - Also, a method to qualify the ML closure trained on the TD
- Datasets & ML models
  - Training data: DNS datasets – 5 channel flows, channel w/ wavy wall, flow around square cylinder
  - Test data: Impinging jet
  - Model: Tensor-basis neural net (TBNN) RANS closure
- Why does this matter? ML closures very inaccurate when extrapolating
  - Need to qualify it – identify when it can & cannot be used
    - Easier to qualify the TD data instead, and then explain the neural network

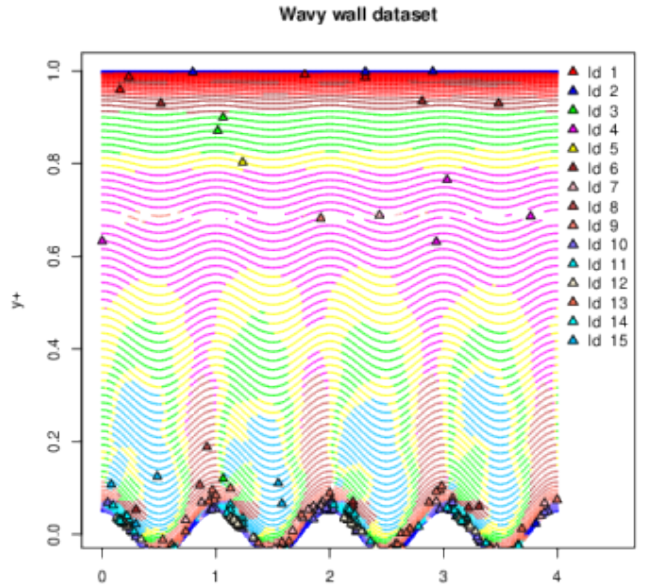
Takeaway: Need a way to assess quality of a ML closure and bound its use



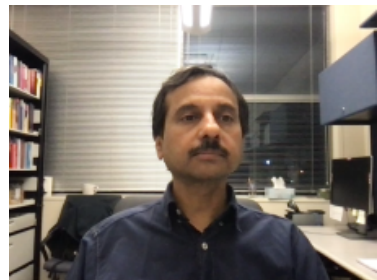
Background

- ML turbulence closures for RANS
  - Predict Reynolds stress or anisotropy, given features / local flow state
  - Random forests & various neural net architectures
- Characterizing training datasets (DOI:10.2514/1.J060919)
  - Physically meaningful partitions: Greedy algorithm to assemble a feature-space for Gaussian Mixture Model clustering of TD
    - Accommodates correlated features from an over-complete dictionary
  - Summarization of TD: Prototype placement, summarizing the TD
    - Supervised learning, with cluster IDs serving as labels
    - Illustrate distribution of TD in feature-space



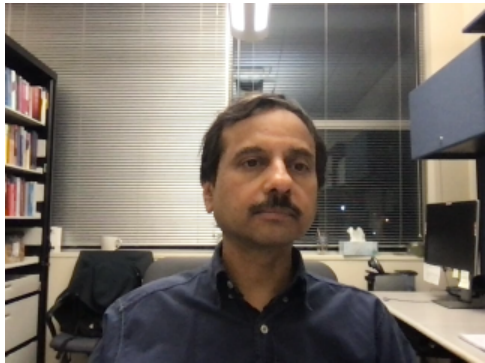


Takeaway: Unsupervised partitioning of TD into homogeneous clusters of turbulent processes

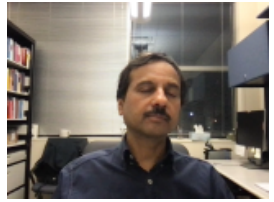


Outstanding questions

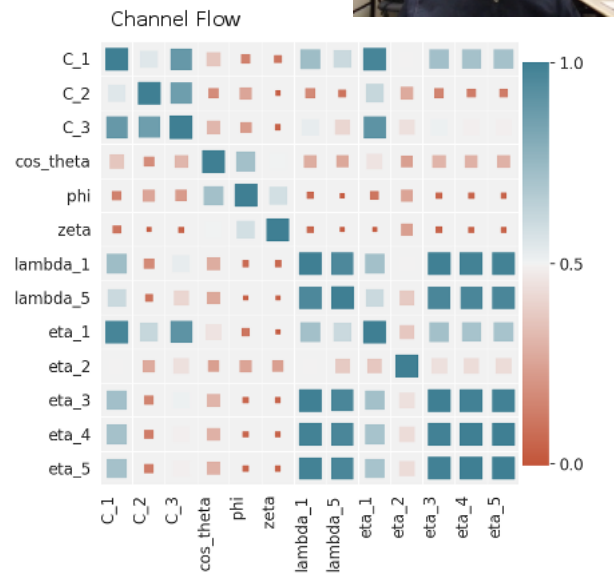
- Qualifying a TD: There exists a feature-space where clustering delivers physically-meaningful clusters
  - Is this feature-space unique?
  - Is it dependent on Gaussian Mixture Model clusters?
  - Can the clusters be used to create a one-class classifier?
    - Can the classifier be used to detect processes (in a ”test” dataset) not in the TD?
- Qualifying a ML closure: Does the closure adhere to turbulence theory?
  - Imbalance in TD will cause ML closure to be biased. Can we detect it from theory?
  - Insufficient TD will lead to a poorly trained NN. Can theory help with detection?

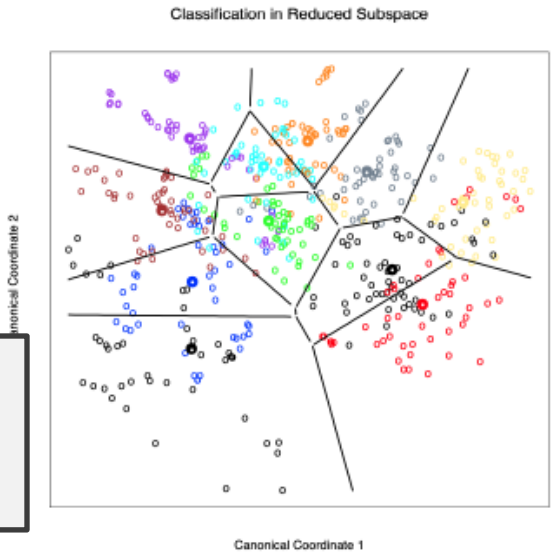


Finding # 1 – Feature-space not unique



- Alternative feature space:
  - Correlation analysis of features' dictionary yields new feature-space
    - Note: Still need a turbulence-informed dictionary of features
  - Classifiers in alternative feature-space are equally predictive of class labels
  - Mapping from class labels to features is rather simple
    - LDA and Random Forest classifiers have similar classification performance
- Alternative clustering methods: In alternative feature-space
  - Spectral clustering in alternative feature-space yields new clusters
    - But they overlap with original clusters 78% - 98% - so not really different

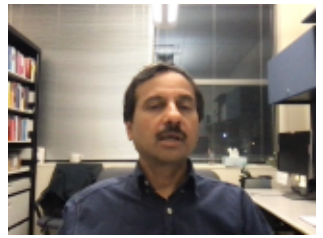




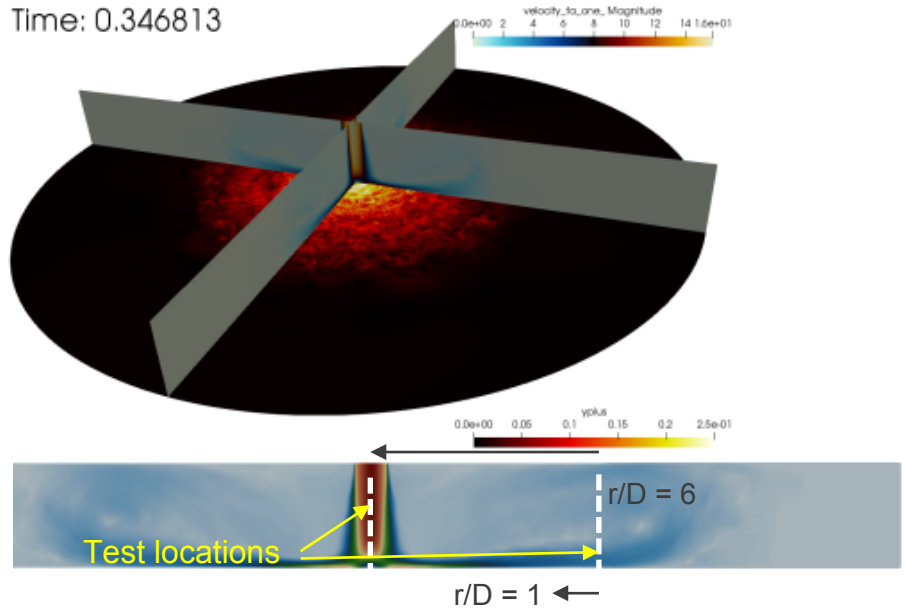
Takeaways:

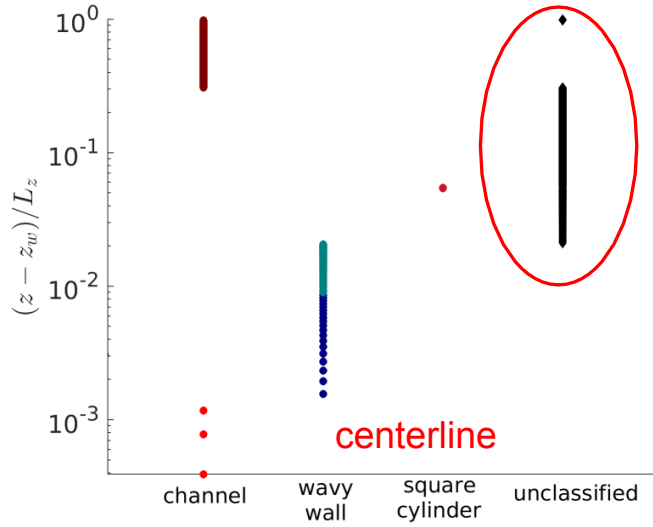
- Feature-space for clustering is not unique
- Clusters are independent of clustering method and feature-space

Finding # 2 – TD qualified with 1-class classifier



- Rationale: Clustering identifies partitions with homogeneous turbulent processes in TD
  - Implication: Given a “test” dataset, intersect with TD and find processes absent from it
    - Qualifies the TD!
- Procedure: Make a 1-class classifier to recognize known turbulent processes /clusters
  - Soft Independent Modeling of Class Analogs (SIMCA) & test
  - Test dataset: Impinging jet
- Outcome: Reliably finds processes absent in TD
  - Manually verified



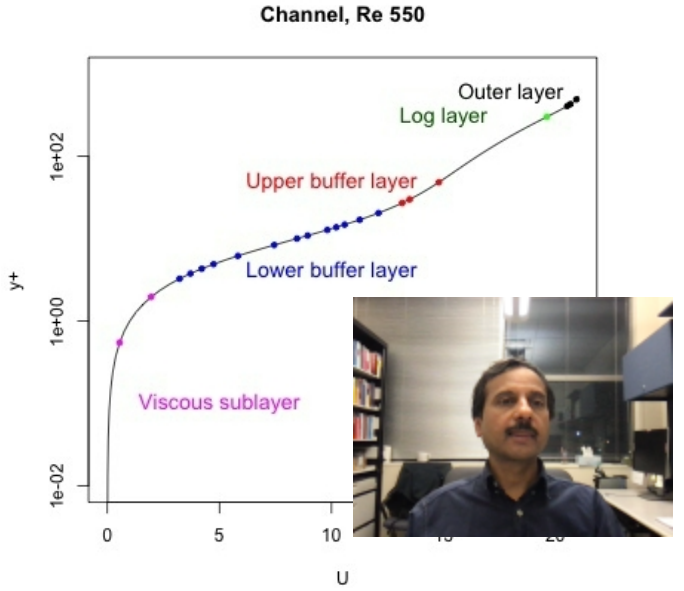


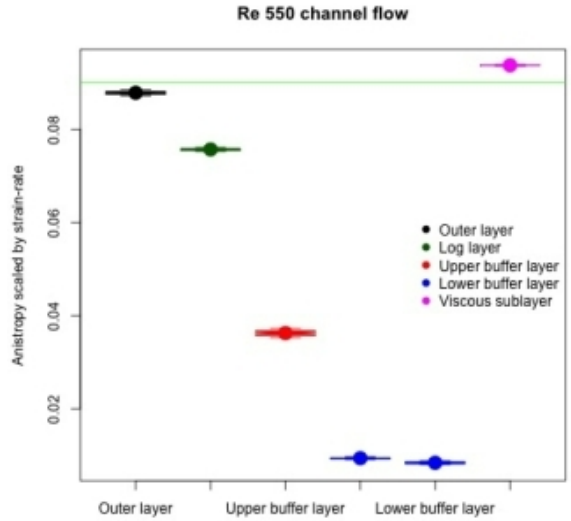
Takeaway: Physically meaningful clusters + 1-class classifier = Qualifiable Training Dataset

Finding # 3 – LIME explanations of ML closures

- A TBNN closure should resemble a linear eddy viscosity model
  - At least when trained on channel flow training data
- Use LIME to find the functional form of TBNN closure
  - At prototypes chosen from a turbulent boundary layer (TBL)
- Use TBNN evaluations near a prototype to fit a Generalized Linear Mixed-effects Model
  - Reveals the functional dependence on TBNN inputs
- Findings with 2 separate TDs:
  - TBL regions well represented in the TD resemble a LEVM
  - For insufficient TD, TBNN functional form is non-physical

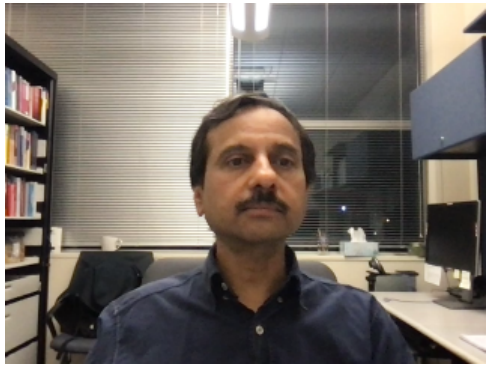
Takeaway: LIME can expose the functional form of a black-box closure for comparison with closed-form expressions





Summary

- A training dataset can be clustered into partitions with homogeneous turbulent process
  - The feature-space can be assembled from a theory-informed dictionary
  - The feature-space is not unique
  - Cluster IDs can become class labels
- A labeled TD can be used to identify absent processes and thus qualify it
  - and any model trained with it
  - Requires one to make a 1-class classifier (SIMCA, 1-class SVMs etc.)
- A black-box neural net closure can reduced to its functional form via LIME
  - Best done at prototypes taken from its TD
  - Allows comparison with closed-form expressions & identification of model shortcomings
- More info:
  - M. Barone et al, "Feature Selection, Clustering, and Prototype Placement for Turbulence Data Sets", *AIAA Journal,* 2022. DOI:10.2514/1.J060919
  - T. Banerjee et al, “Qualifying Training Datasets for Data-Driven Turbulence Closures”, *AIAA Aviation* 2022.



Acknowledgement

This work was supported by the Laboratory Directed Research and Development program at Sandia National Laboratories. 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. 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.