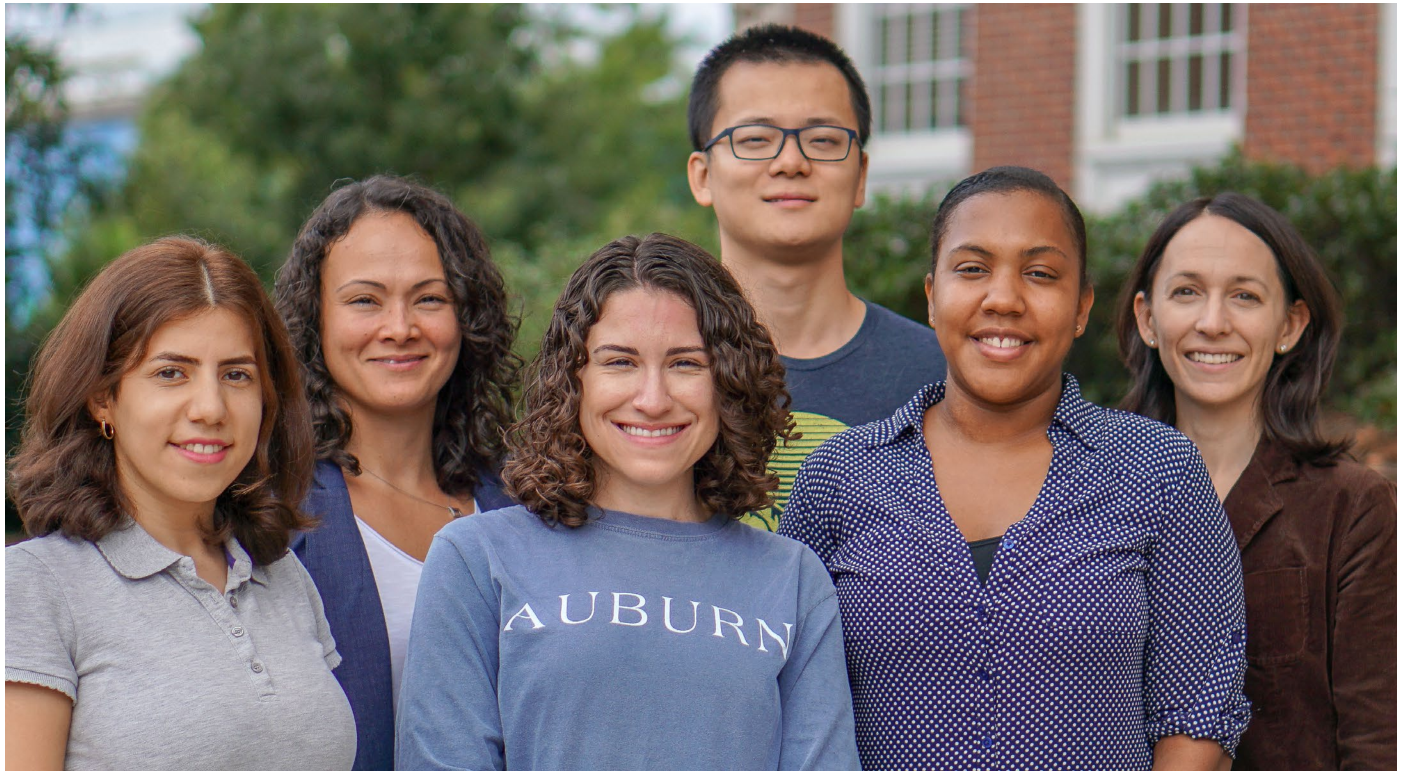
Role of Artificial Intelligence and Machine Learning in Flow Assurance Problems

Selen Cremaschi (Pronouns: she/her)

B. Redd & Susan W. Redd Endowed Eminent Scholar Chair Professor
Graduate Program Chair
Department of Chemical Engineering
Auburn University
selen-cremaschi@auburn.edu

AUBURN UNIVERSITY
SAMUEL GINN COLLEGE OF ENGINEERING

Acknowledgements



JDRF (Improving Lives. Curing Type I Diabetes)
ARMI (Advanced Regenerative Manufacturing Institute)
NSF (National Science Foundation)
DOE (Department of Energy)
ED (Department of Education)
Chevron Technical Center

Cremaschi Group

We develop models and decision support tools for complex systems under uncertainty.

Tools and Methods

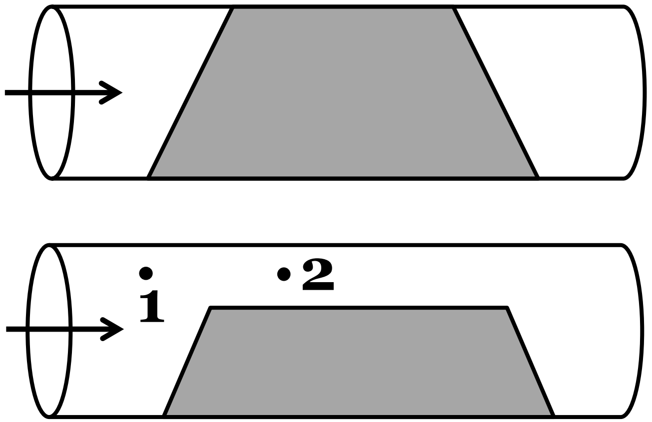
- Optimization under uncertainty
- Data analytics, machine learning, data-driven modeling, hybrid modeling
- Uncertainty quantification and propagation

Application Areas

- Oil and gas production
- Infrastructure and supply chain resilience
- Biofuels and bio-chemicals production
- CO_2 capture and sequestration
- Pharmaceutical R&D pipeline management
- Cancer care system
- Tissue manufacturing

Increasing confidence in model estimates

Fluid velocity for sand transport



$$v_2 > v_1 \Rightarrow P_1 > P_2$$

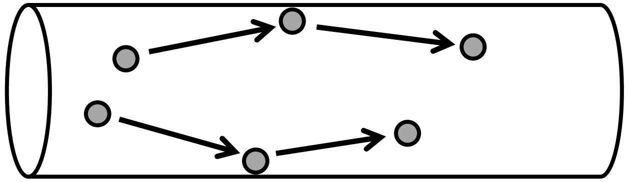
Avoid

Blockages

Pressure losses

Pigging

Erosion in conduits



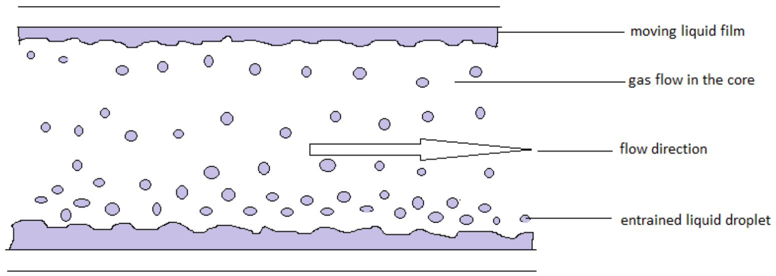
Avoid

Environment impact

Lost production

Repair costs

Liquid entrainment fraction

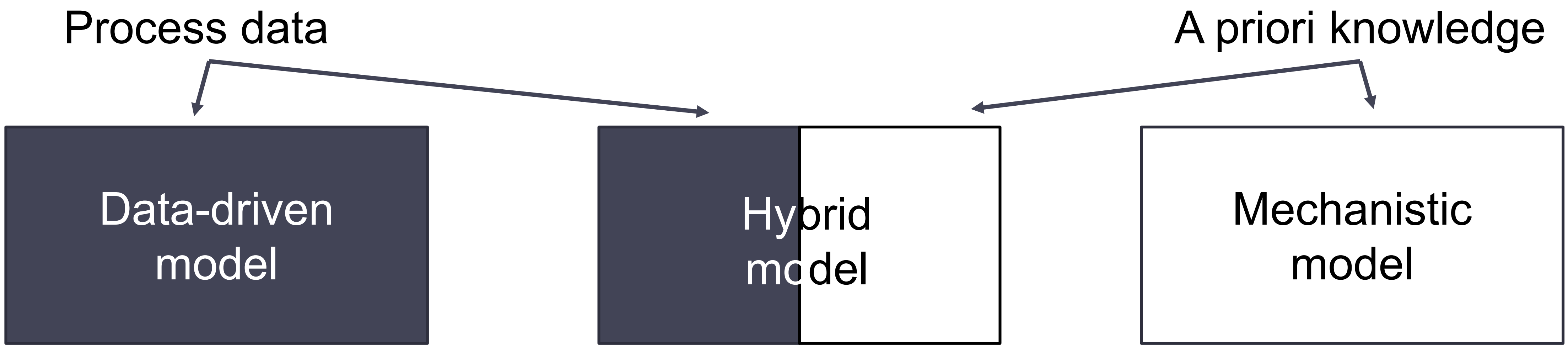


$$f_E = \frac{\text{Liquid mass flow rate in the gas phase}}{\text{Total liquid mass flow rate}}$$

Avoid

Under-sizing separation facilities and downstream equipment

Hybrid model[1,2]



Input-output data inference
No need for process knowledge
Valid within the range of data

Process mechanism is known
Laws of nature are applied
Idealized by assumptions

1. von Stosch, et al. (2014). Hybrid semi-parametric modeling in process systems engineering: Past, present and future. *Comput. Chem. Eng.*, 60, pp 86-101
2. Bradley et al. (2022). Perspectives on the integration between first-principles and data-driven modeling. Comput. Chem. Eng., 166, 107898.

Our hybrid modeling efforts in flow assurance

Data clustering

Group data into similar sets
Identify best models for each set

Uncertainty quantification

Error analysis
Propagate input and data uncertainty

Discrepancy modeling

Machine learning
for capturing mismatch between
observation and first-principle model

Model refinement

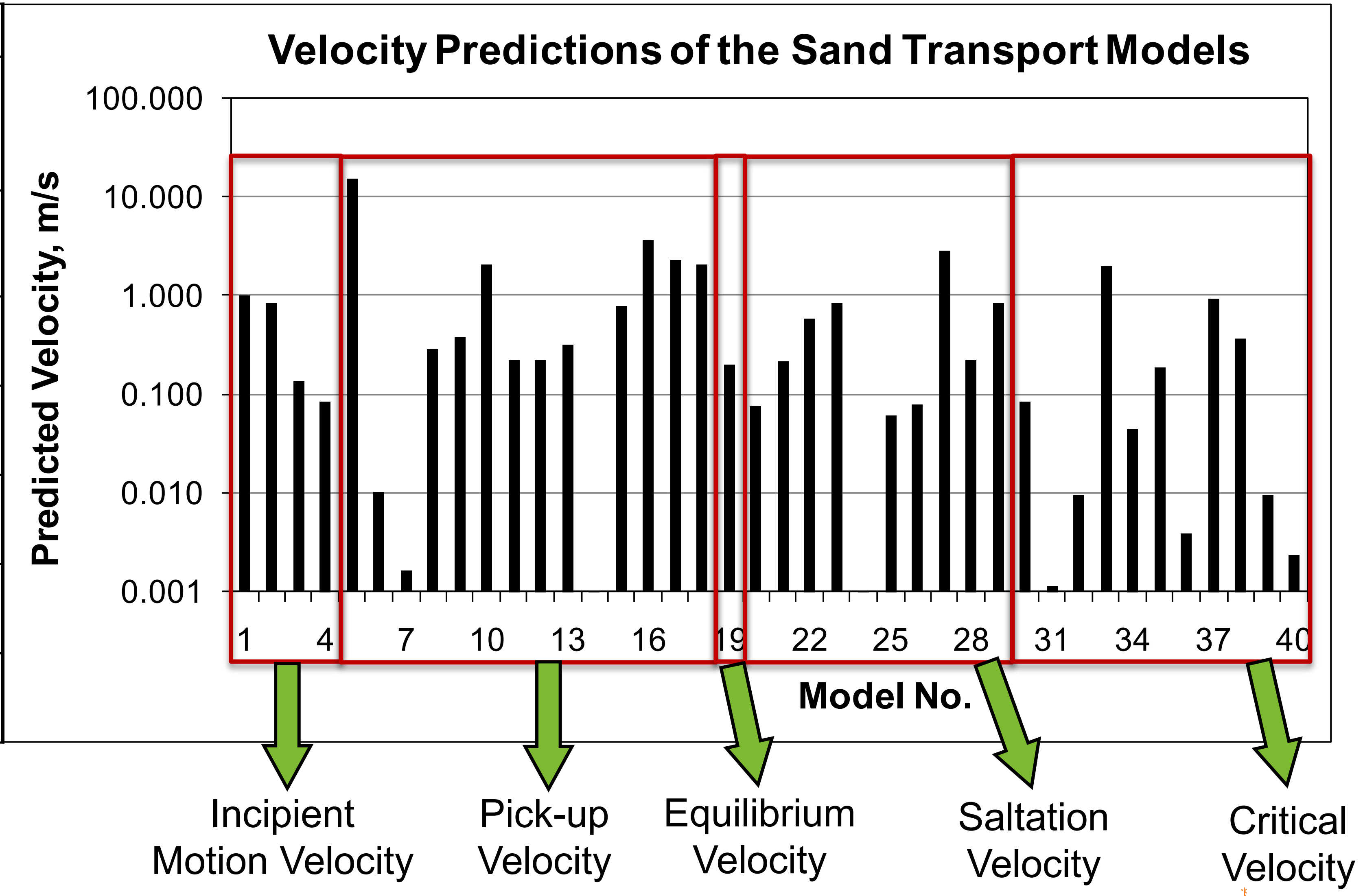
Machine learning suggested
experimental campaigns
and first-principle model refinement

Feature selection

Incorporating expert knowledge
to feature selection
for hybrid models

Solids transport challenges

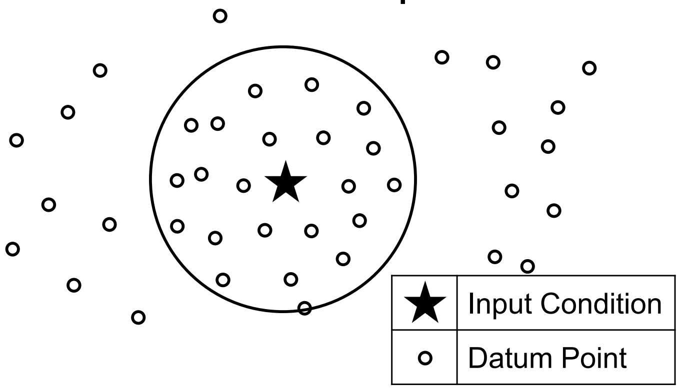
Typical field condition	
Particle Concentration (ppm)	5
Pipe Inclination Angle (degree)	0
Fluid Density (kg/m³)	850
Fluid Viscosity (mPa-s)	15
Particle Density (kg/m³)	2630
Particle Diameter (μm)	60
Hydraulic Diameter (m)	0.21



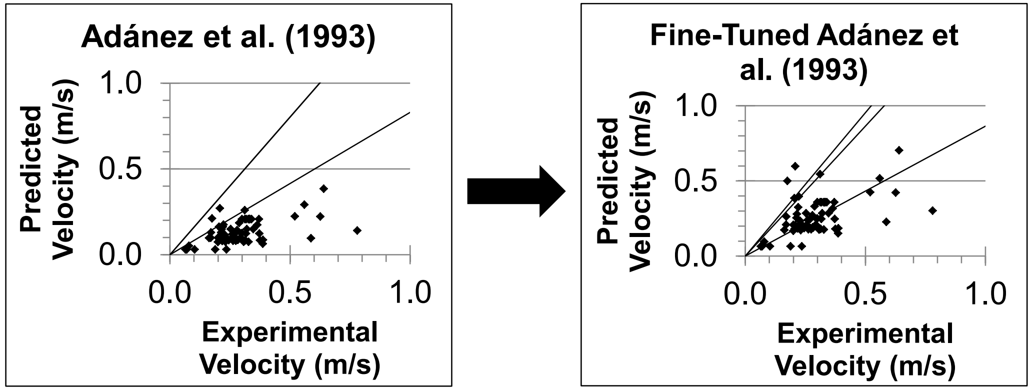
AUBURN UNIVERSITY

Hybrid models with data clustering, model tuning and evaluation

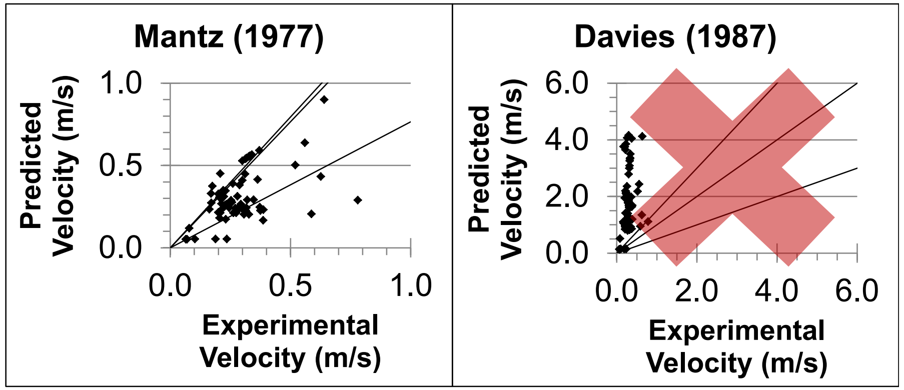
1. **Cluster data** to generate a dataset that contains similar data to input condition



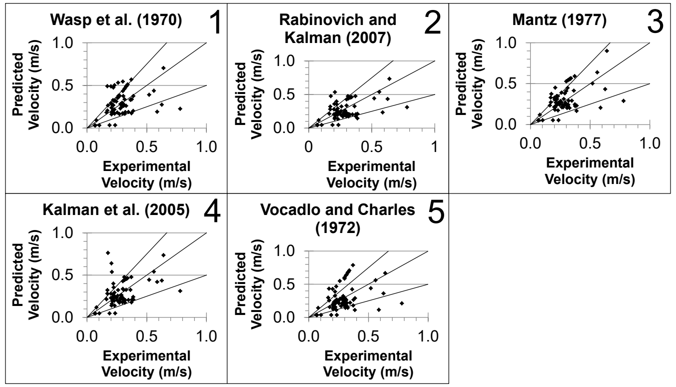
2. **Tune models** to reduce bias in predictions



3. **Model screening** to discard inaccurate models



4. **Rank** models and quantify prediction uncertainty



Soepyan et al. (2013). Journal of Petroleum Science and Engineering, 110, 210-224.
Soepyan et al. (2014). AIChE Journal, 60 (1), 76-122.
Soepyan et al. (2016). Computers & Chemical Engineering, 93, 143-159.
Soepyan et al. (2017). Journal of Petroleum Science and Engineering, 151, 128-142.

1. Cluster data – Incorporating expert knowledge into Euclidean distance

1. Normalize independent variables to remove unintended contribution due to scale
2. Calculate the weighted Euclidean distance between each data point i and input condition

$$d_i = \left(\boxed{|h_1|} \times \left| \overline{\log C_i} - \overline{\log C_0} \right|^2 + \boxed{|h_2|} \times \left| \overline{\cos\theta_i} - \overline{\cos\theta_0} \right|^2 + |h_3| \times \left| \boxed{\overline{\log\left(\rho_{S,i}/\rho_{f,i}\right)}} - \overline{\log\left(\rho_{S,0}/\rho_{f,0}\right)} \right|^2 + |h_4| \times \left| \overline{\log\left(D_i/d_{p,i}\right)} - \overline{\log\left(D_0/d_{p,0}\right)} \right|^2 + |h_5| \times \boxed{\left| \overline{\log Ar_i} - \overline{\log Ar_0} \right|^2} \right)^{1/2}$$

h_l = correlation between independent and dependent variables

3. Sort data based on distance, calculate differences in distance between consecutive points
4. Locate “jumps” using outlier detection, populate dataset with data using “jumps”

Soepyan et al. (2017). Journal of Petroleum Science and Engineering, 151, 128-142.

3. Model screening

Discard model if:

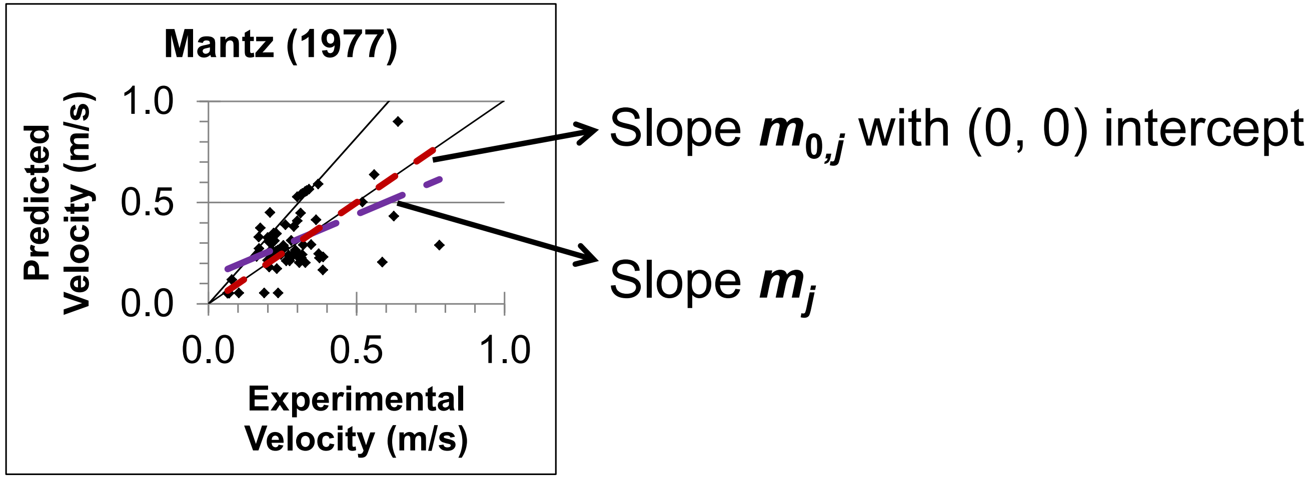
$$R^2_{adj,j} < 0$$

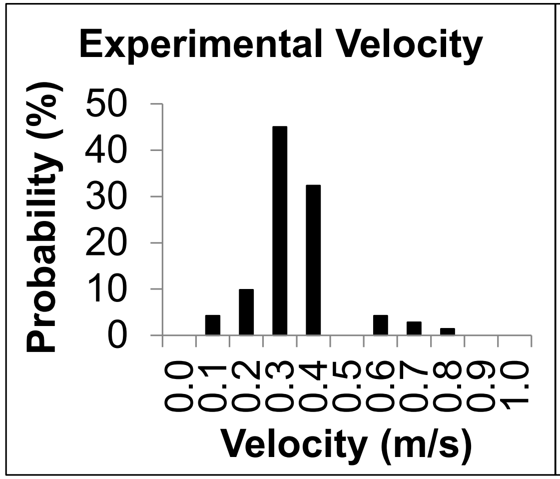
$$m_{0,j} < 0.5$$

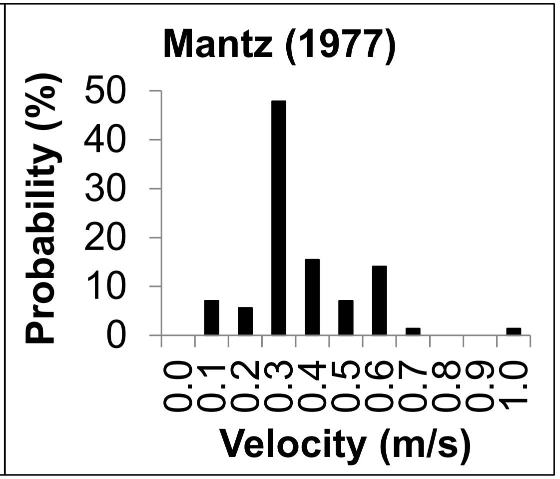
$$m_{0,j} > 1.5$$

$$m_j < 0$$

Different velocity distributions





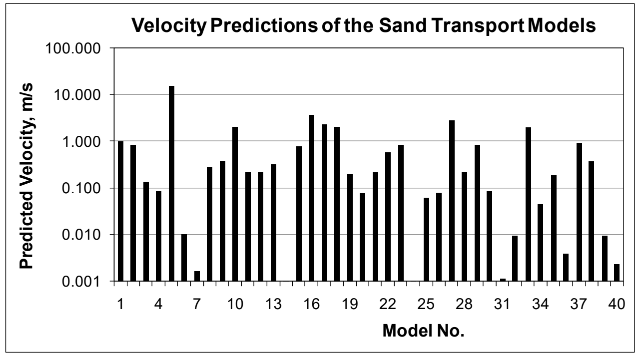


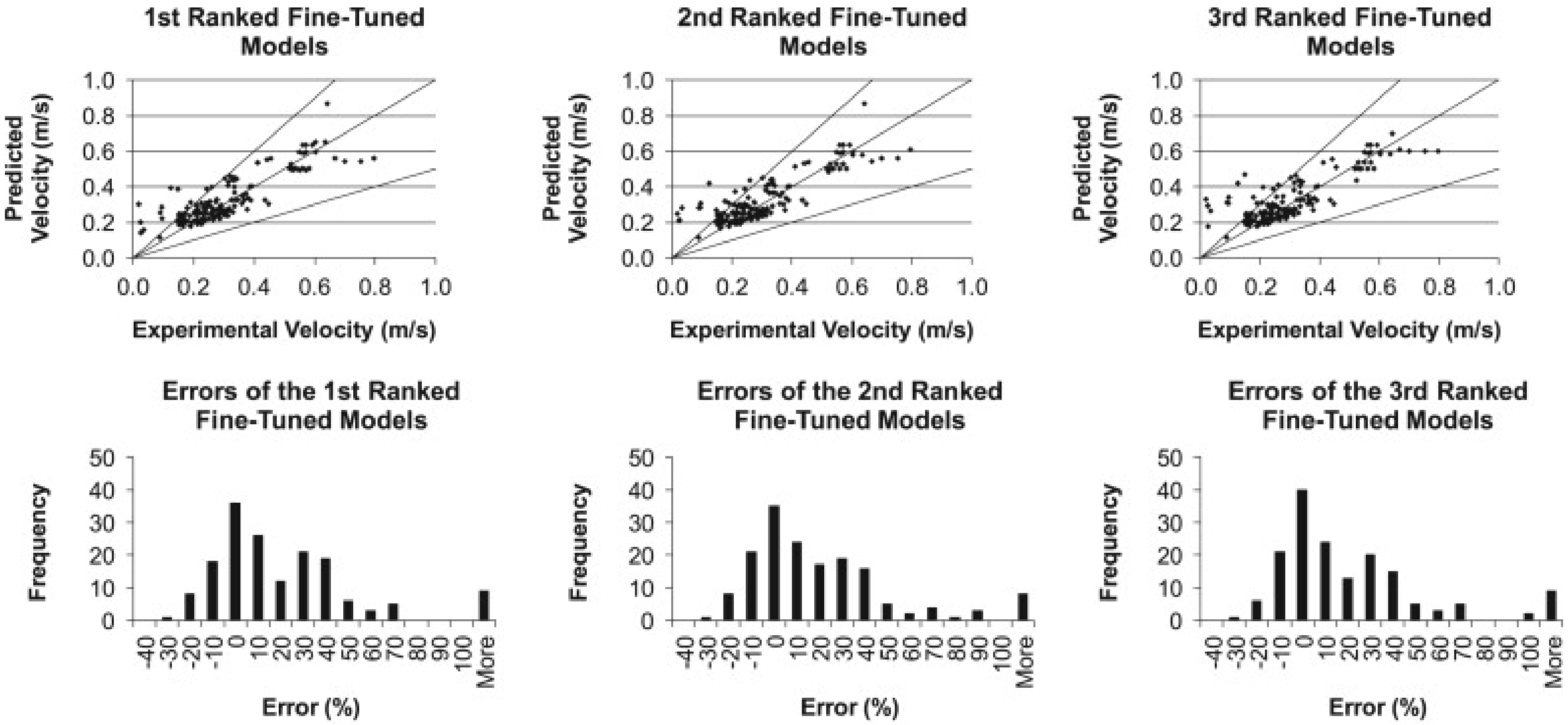
Soepyan et al. (2013). Journal of Petroleum Science and Engineering, 110, 210-224.
Soepyan et al. (2017). Journal of Petroleum Science and Engineering, 151, 128-142.

4. Model ranking

Rank using score S_j

$$S_j = w_1\overline{E_{MS,j}} + w_2\overline{E_{MAP,j}} + w_3\overline{E_{MA,j}} + w_4\overline{R^2_{adj,dev,j}}$$
$$+w_5\overline{max_{\%,j}} + w_6\overline{P_{\%,j}} + w_7\overline{v_{0,dev,j}} + w_8\overline{k_{j,non}}$$





Soepyan et al. (2013). Journal of Petroleum Science and Engineering, 110, 210-224.
Soepyan et al. (2017). Journal of Petroleum Science and Engineering, 151, 128-142.

4. Prediction uncertainty quantification

Uncertainty due to model errors

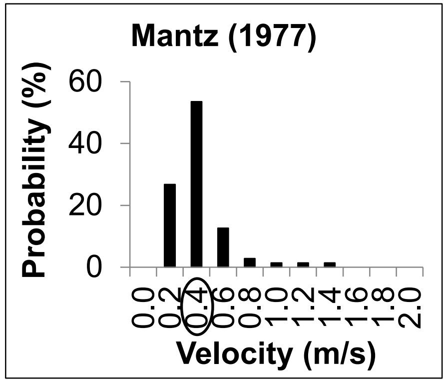
Calculate percent error for each data point

Correct velocity prediction for input

Generate velocity prediction histogram

Dataset		
v_{exp} (m/s)	v_{calc} (m/s)	Percent Error (%)
0.23	0.26	13.7
0.21	0.21	2.1
0.29	0.26	-9.8
0.21	0.18	-11.5
…	…	…

Corrected Velocity at Input Condition (m/s)
0.26
0.29
0.33
0.33
…



Uncertainty due to input condition, experimental data, and model errors

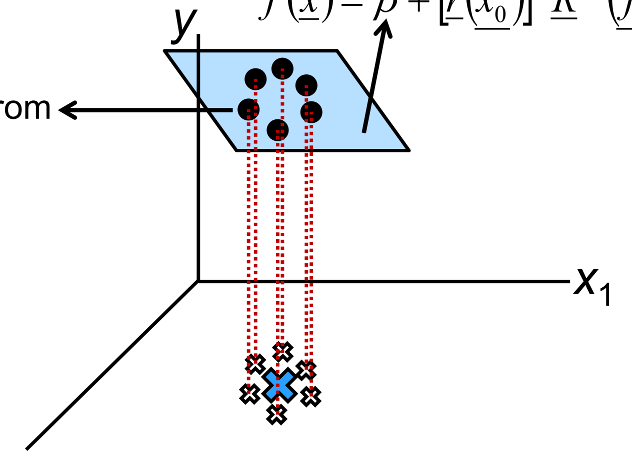
Propagate input condition uncertainty using Monte Carlo simulation

Perturb model predictions using random samples from experimental data uncertainty

Generate prediction uncertainty using Kriging models of percentiles

$$\hat{f}(\underline{x}) = \hat{\beta} + [\underline{r}(\underline{x}_0)]^T \underline{R}^{-1}(\underline{f}_{true} - \hat{\beta}\underline{u})$$

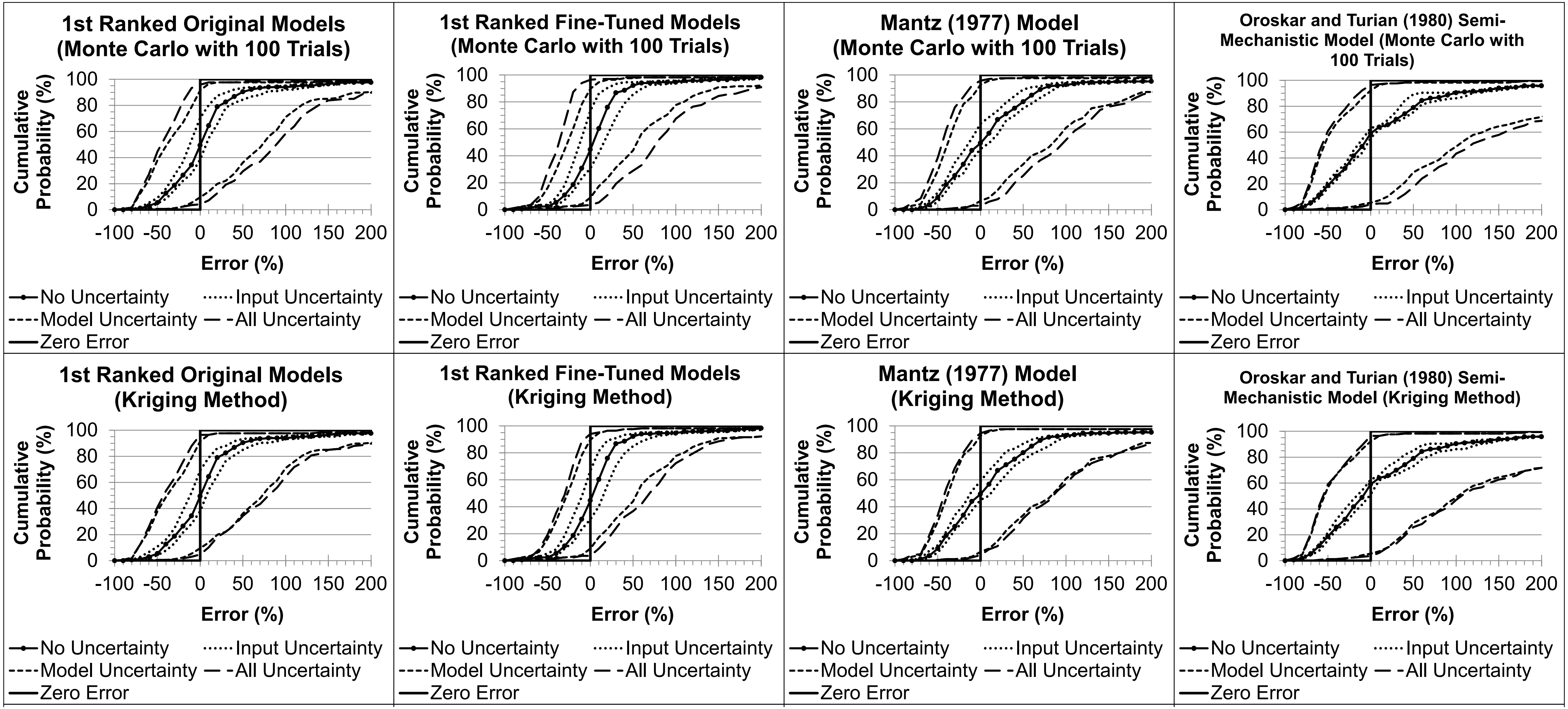
Xth percentile of prediction of model from 1 Monte Carlo trial



Symbols:	
	Input Condition
	Perturbed Input Condition

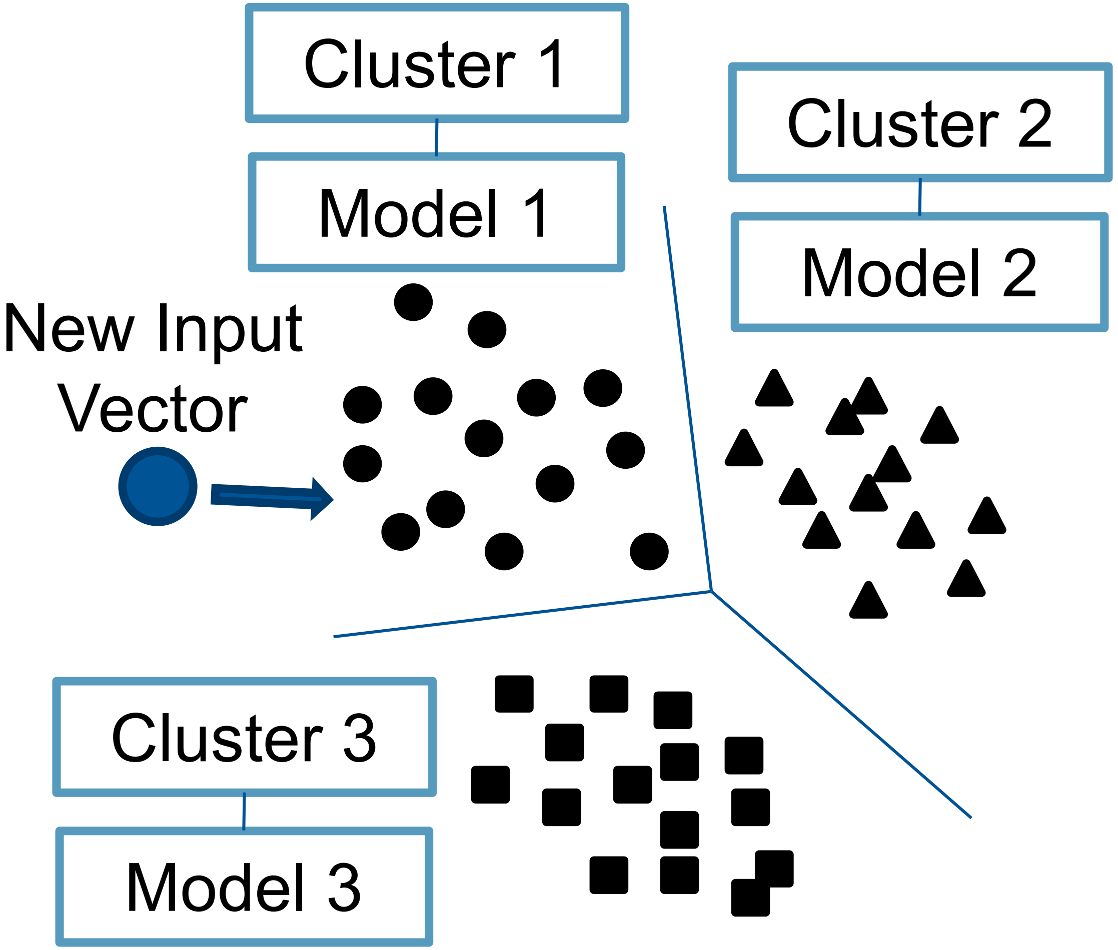
Soepyan et al. (2016). Computers & Chemical Engineering, 93, 143-159.
Soepyan et al. (2017). Journal of Petroleum Science and Engineering, 151, 128-142.

Model prediction uncertainty for case study



Soepyan et al. (2016). Computers & Chemical Engineering, 93, 143-159.

Hybrid models with data clustering and model evaluation



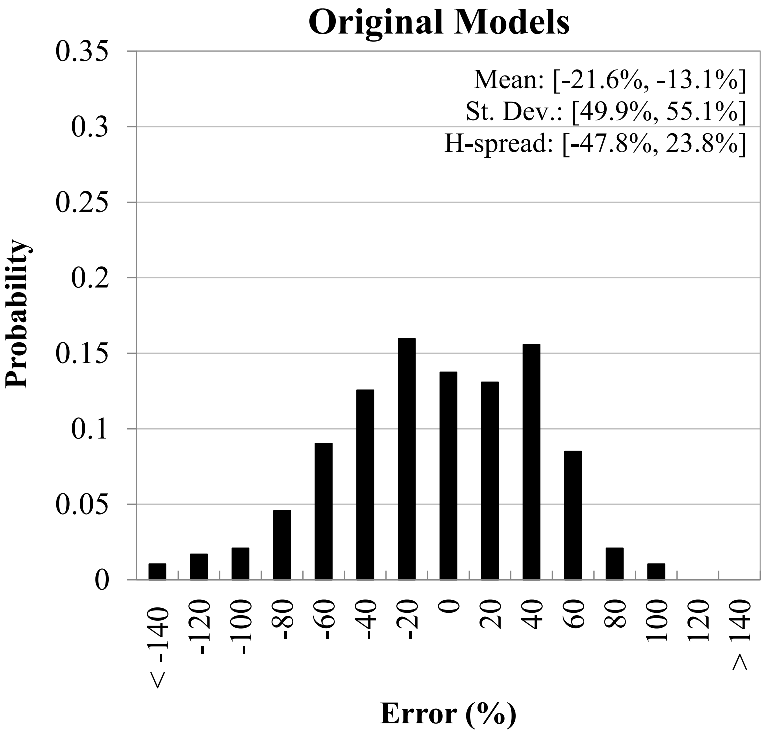
Cluster experimental data
using k-means clustering
with modified Euclidean distance
to integrate knowhow
on threshold velocity and its type

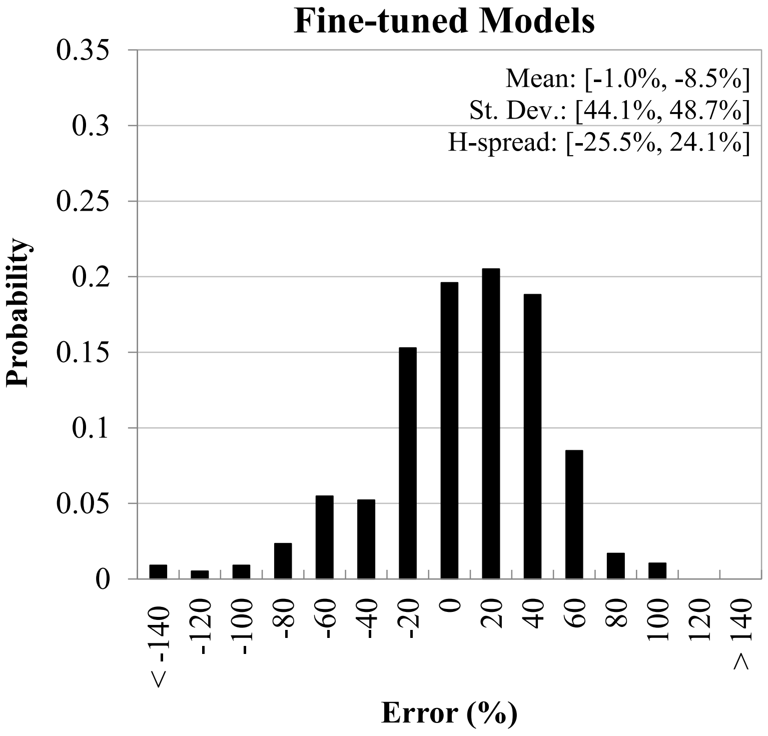
$$Dist = \left[\sum_{l=1}^{M} \left|R_{l,t}\right| \left|x_l^{(j)} - cent_{l,j}\right|^p\right]^{1/p}$$

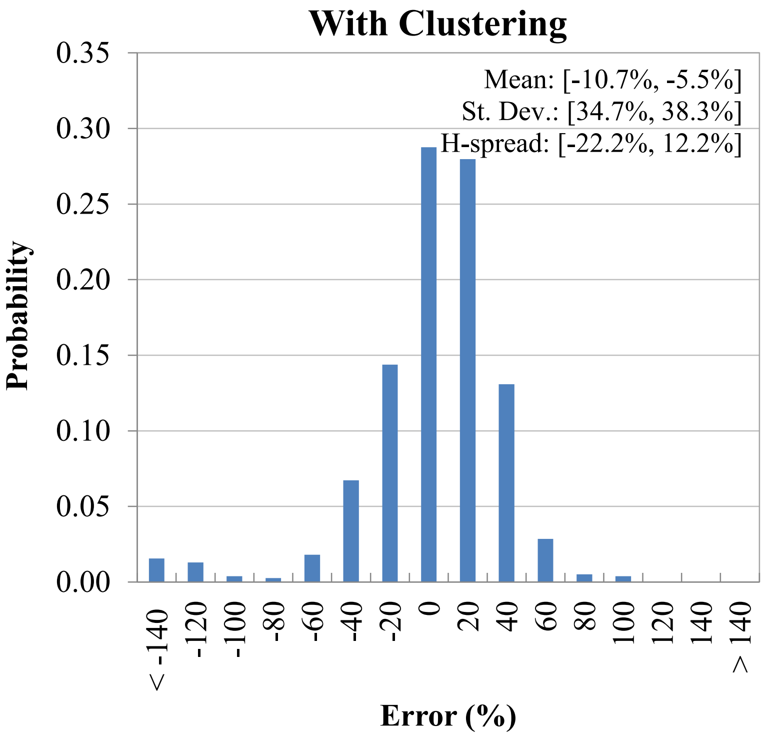
Determine the best model
for each cluster
using model screening and ranking

Shin et al. (2015). Computers & Chemical Engineering, 81, 355-363.

Clustering reduces prediction uncertainty







Shin et al. (2015). Computers & Chemical Engineering, 81, 355-363.

Our hybrid modeling efforts in flow assurance

Data clustering

Group data into similar sets
Identify best models for each set

Uncertainty quantification

Error analysis
Propagate input and data uncertainty

Discrepancy modeling

Machine learning
for capturing mismatch between
observation and first-principle model

Model refinement

Machine learning suggested
experimental campaigns
and first-principle model refinement

Feature selection

Incorporating expert knowledge
to feature selection
for hybrid models

Accurate erosion predictions is important and challenging

Erosion is a complex process
- Geometry of flow lines
- Construction material
- Particle properties
- Flow conditions

U.S. has the largest network of energy pipelines

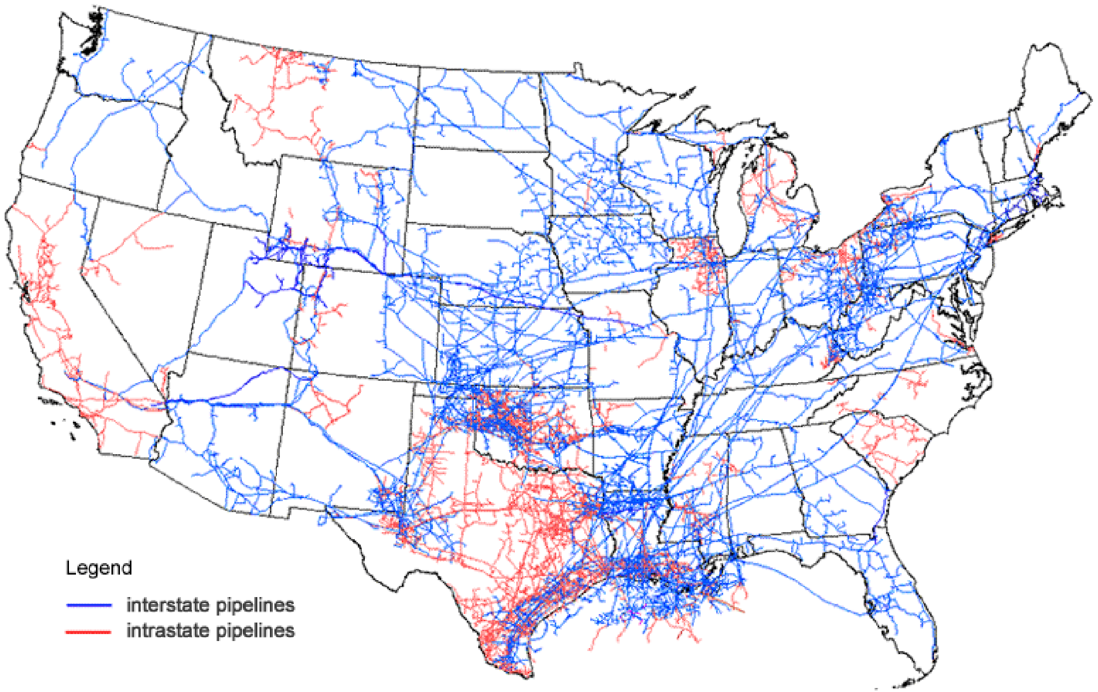
Approximately 72,000 miles of crude oil lines

Sand production is an unplanned event
- Filtration equipment; Chemical inhibitors

Possible outcome
- Environment impact; Lost production; Repair costs

Modeling of erosion is imperfect
- Little information is available on model or data uncertainty
- Measurement takes a long time

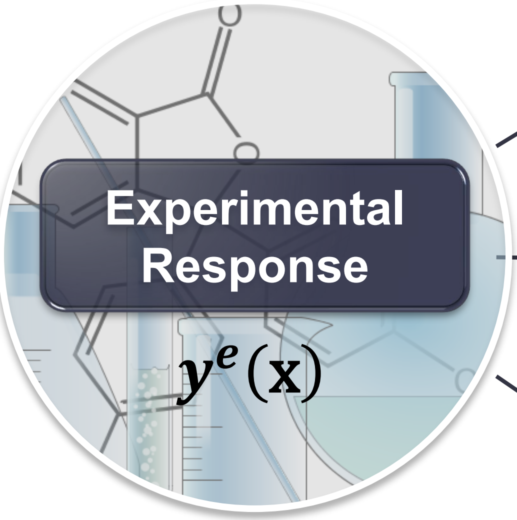
Map of U.S. interstate and intrastate natural gas pipelines



Source: U.S. Energy Information Administration, *About U.S. Natural Gas Pipelines*

A hybrid modeling approach with discrepancy modeling*

$$\boxed{y^e(\mathbf{x}) = y^m(\mathbf{x}) + \delta(\mathbf{x}) + \varepsilon}$$



$\mathbf{x}$: inputs

Model Prediction $y^m(\mathbf{x})$

- **Model form uncertainty**
- **Parameter uncertainty**
- **Interpolation uncertainty**

Model Discrepancy $\delta(\mathbf{x})$

The deviation between the model prediction and the true experimental response

Data Uncertainty ε

Measurement errors due to both instrument and human errors

*Kennedy, M.C. and O'Hagan, A., 2001, Bayesian calibration of computer models, *Journal of the Royal Statistical Society*.

Parallel structure hybrid model

$$\boldsymbol{y^e}(\mathbf{x}) = \boldsymbol{y^m}(\mathbf{x}) + \boldsymbol{\delta}'(\mathbf{x}) + \boldsymbol{\varepsilon} \xrightarrow{\text{assume } \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \boldsymbol{var}(\boldsymbol{\varepsilon}))} \boldsymbol{y^e}(\mathbf{x}) = \boldsymbol{y^m}(\mathbf{x}) + \boldsymbol{\delta}(\mathbf{x})$$

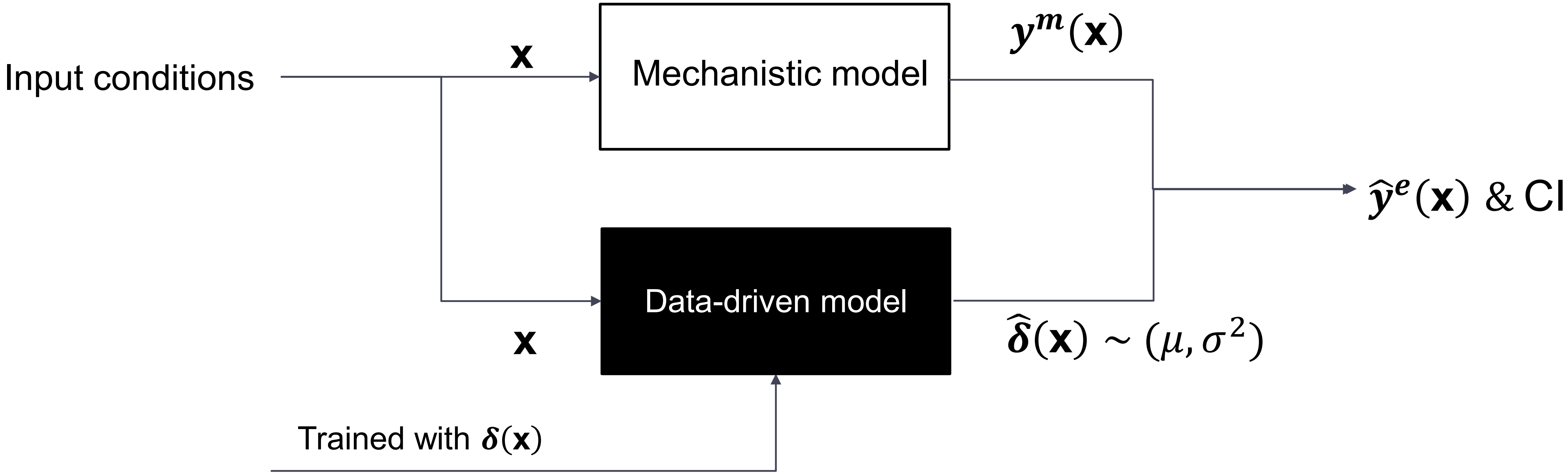
$\boldsymbol{y^e}(\mathbf{x})$- experimental measurement

$\boldsymbol{y^m}(\mathbf{x})$-prediction from mechanistic models

$\boldsymbol{\delta}'(\mathbf{x})$-model bias

$\boldsymbol{\varepsilon}$-experimental uncertainty

$\boldsymbol{\delta}(\mathbf{x})$-model discrepancy



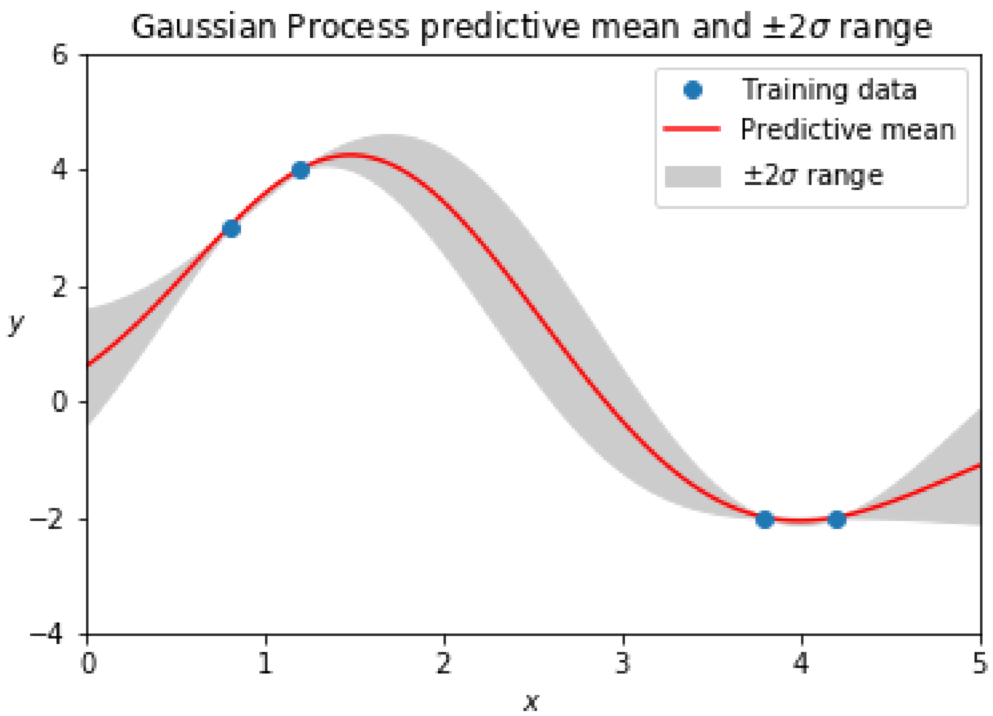
Data-driven model - Gaussian Process Modeling[1]

$$\boldsymbol{\delta} \sim \mathcal{GP}(\boldsymbol{m}, \boldsymbol{k})$$

Non-parametric approach to map inputs and outputs

Relies on the covariance function ($\boldsymbol{k}$) to define similarity at different observation locations

Gives the prediction along with its variance (confidence interval, CI)



$$\boldsymbol{m}(\boldsymbol{x}) = \boldsymbol{a}, \text{ and } \boldsymbol{k}(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{\sigma_y^2} \, \boldsymbol{exp}\left(-\frac{(\boldsymbol{x} - \boldsymbol{x}')^2}{2\boldsymbol{l}^2}\right)$$

$$\text{Hyper-parameters } \theta = \{a, \sigma_y, l\},$$

1 Rasmussen, C. E., Williams, C. K., 2006, Gaussian Process for Machine Learning, *The MIT Press*.

Model discrepancy differs significantly for erosion predictions

Input variables

- h_B: pipeline hardness (vicker)
- μ_f: flow viscosity (cp)
- ρ_f: flow density (kg/m^3)
- v_f: flow velocity (m/s)
- ρ_p: density of sand particles (kg/m^3)
- $\mathrm{d_p}$: particle size (μm)
- D: pipe diameter (inch)
- Geometry of the pipeline
- Orientation of the pipeline
- Flow regime

Categorical attributes

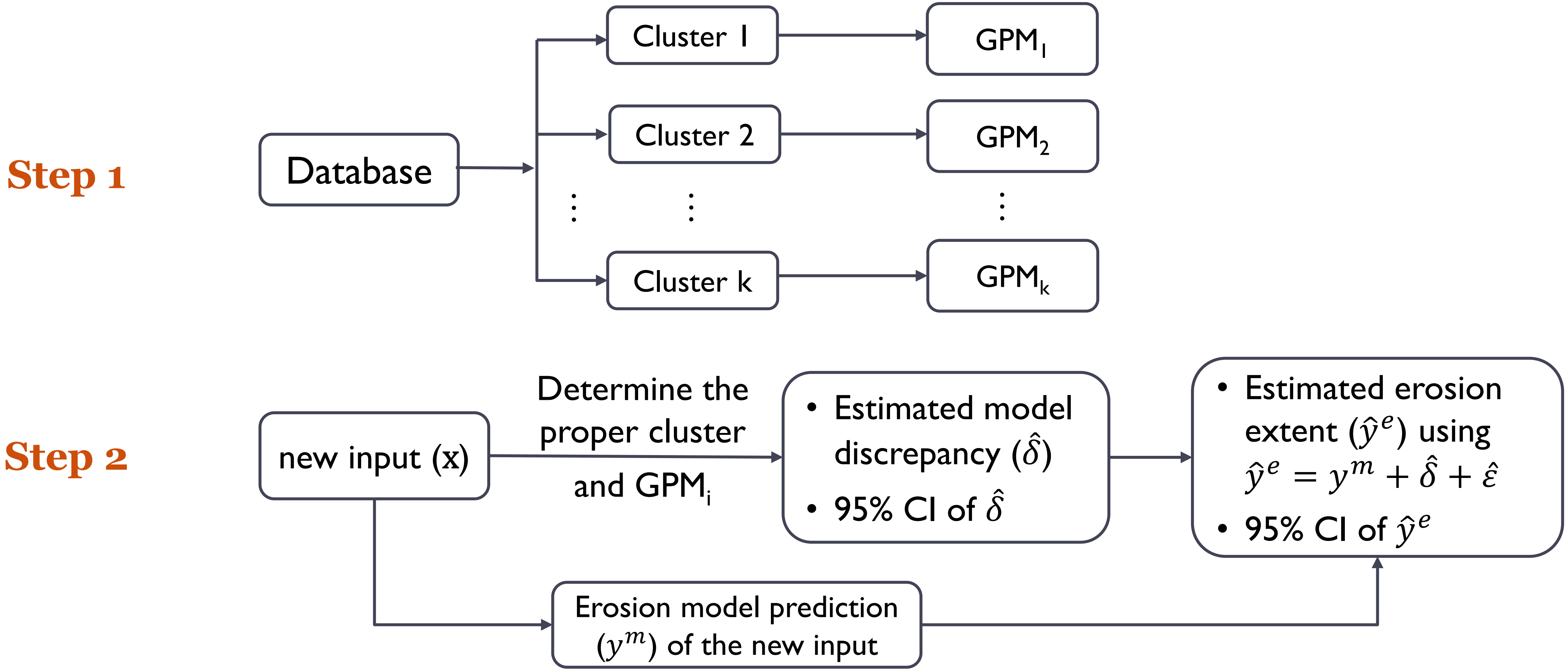
The inputs cover a wide ranges and operating condition differs greatly from each other.

There are six orders of magnitude differences in the model discrepancies for the erosion database.

A single large-scale model is not enough…

Operating conditions	Max δ	Min δ
Pipe diameter (inch)	2	3
Particle size (micron)	350	20
Flow viscosity (cp)	1	1
Liquid velocity (m/s)	0	0.45
Gas velocity (m/s)	122	15.2
Flow regime	Gas	Mist
y^m (mils/lb)	1.3	7.0×10^{-5}
y^e (mils/lb)	3.8	7.13×10^{-5}
δ (mils/lb)	**2.5**	**1.3×10^{-6}**

Hybrid model to predict erosion and its confidence interval combines clustering, GPM discrepancy, and 1D-SPPS

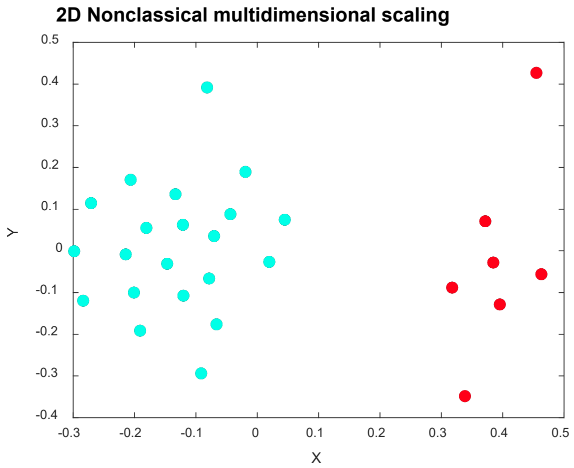


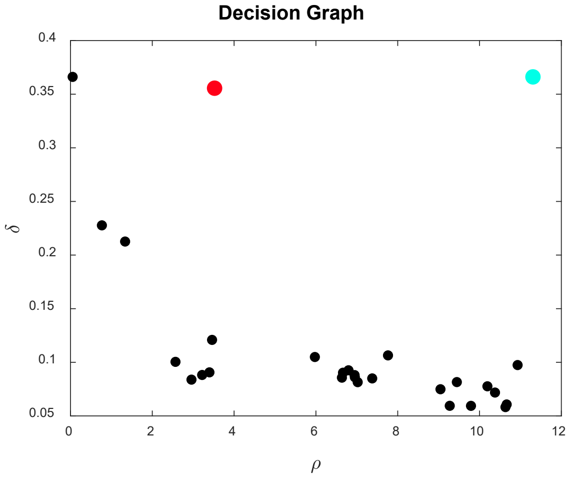
Dai et al. (2021). Computers & Chemical Engineering, 156, 107577.
Dai et al. (2019). Computers and Chemical Engineering, 127, 175-185.
Dai et al. (2019). Chemical Engineering Research and Design, 147, 187-199.
Dai et al. (2018). Wear, 408, 108-119.

Clustering input space of erosion measurement database

Object-cluster similarity metric (**OCIL**[1]) is used because erosion input data contains both categorical and numerical attributes.

We developed a special density-based[2] initialization approach[3].

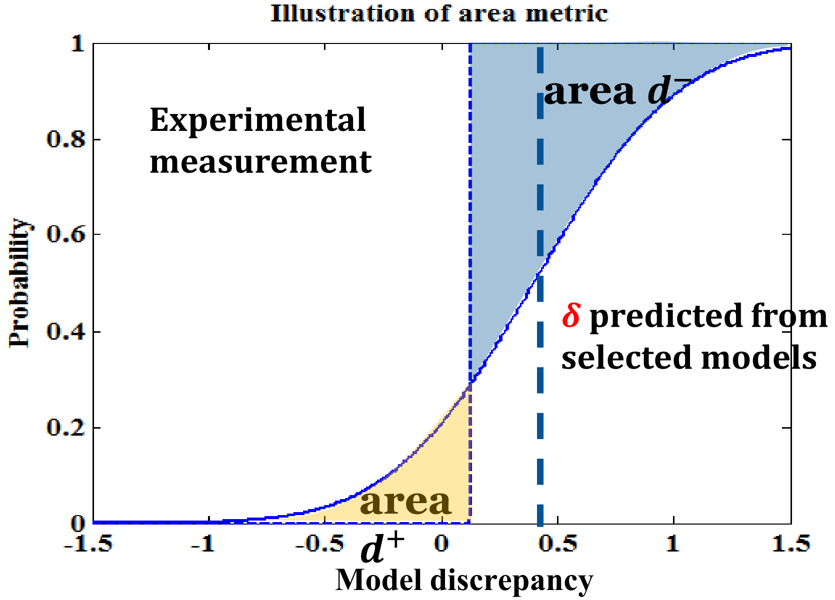




$\boldsymbol{\delta}$: distance

$\boldsymbol{\rho}$: density

Area metric[4]



$$\mathbf{AM} = \boldsymbol{d}^{-} + \boldsymbol{d}^{+}$$

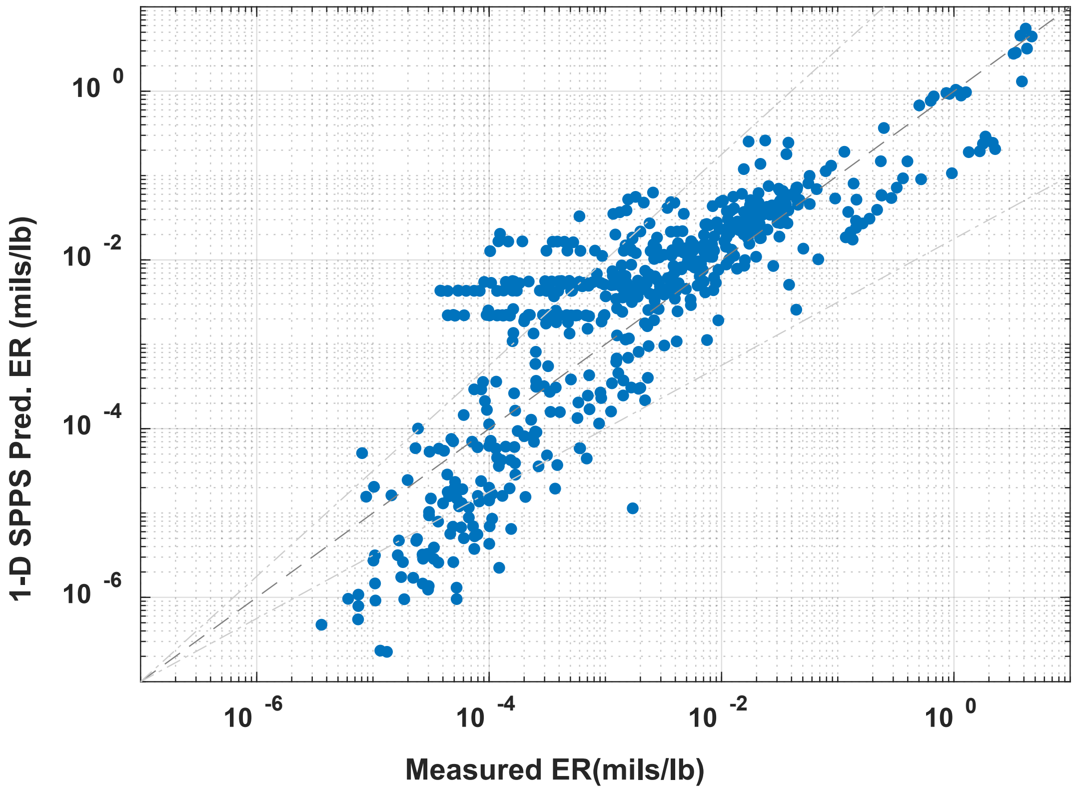
RMSE

$$RMSE(\delta) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left\|\delta_i - \hat{\delta}_i\right\|^2} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\|y^e - \hat{y}^e\|^2}$$

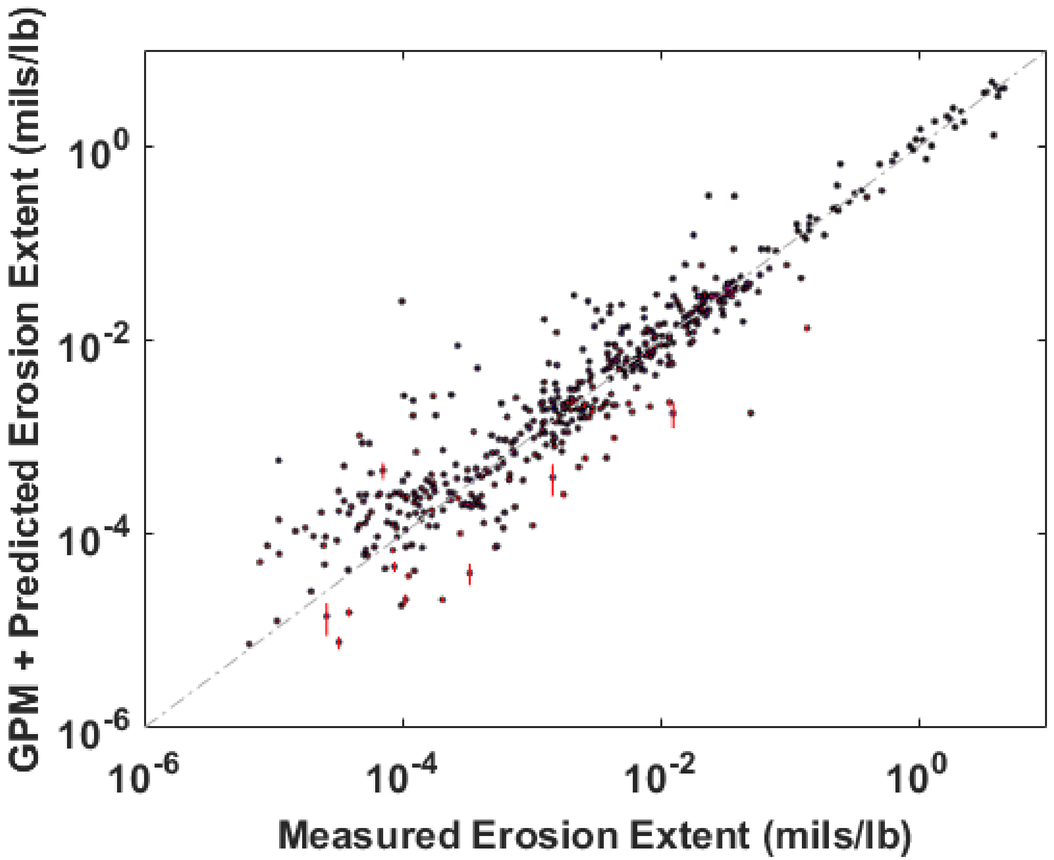
1 Cheung, Y. M. and Jia, H. (2013). Pattern Recognition, 46(8), 2228-2238.
2 Rodriguez, A. and Laio A.. (2014). Science, 344(6191), 1492-1496.
3 Dai et al. (2018). Wear, 408, 108-119.
4 Ferson S. and Oberkampf W.L. (2009). Int Journal of Reliability and Safety, 3, 3-22.

Hybrid model improves erosion predictions with high confidence

Predictions of SPPS 1D version 5.1

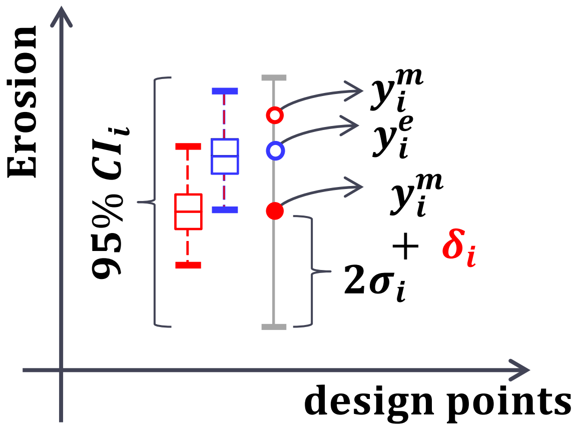


Predictions of the hybrid model



Dai et al. (2018). Wear, 408, 108-119.

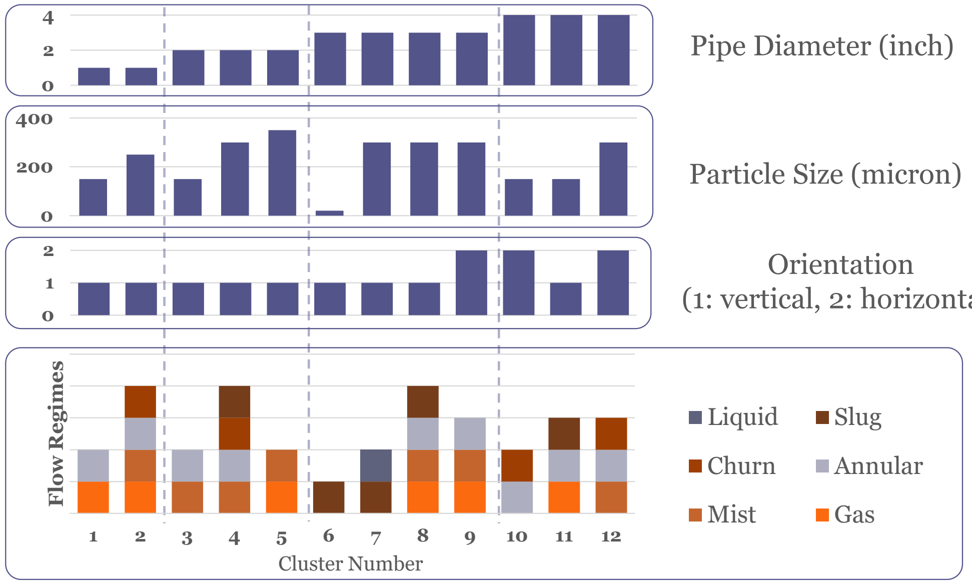
Prediction results reveal accurate predictions for each cluster

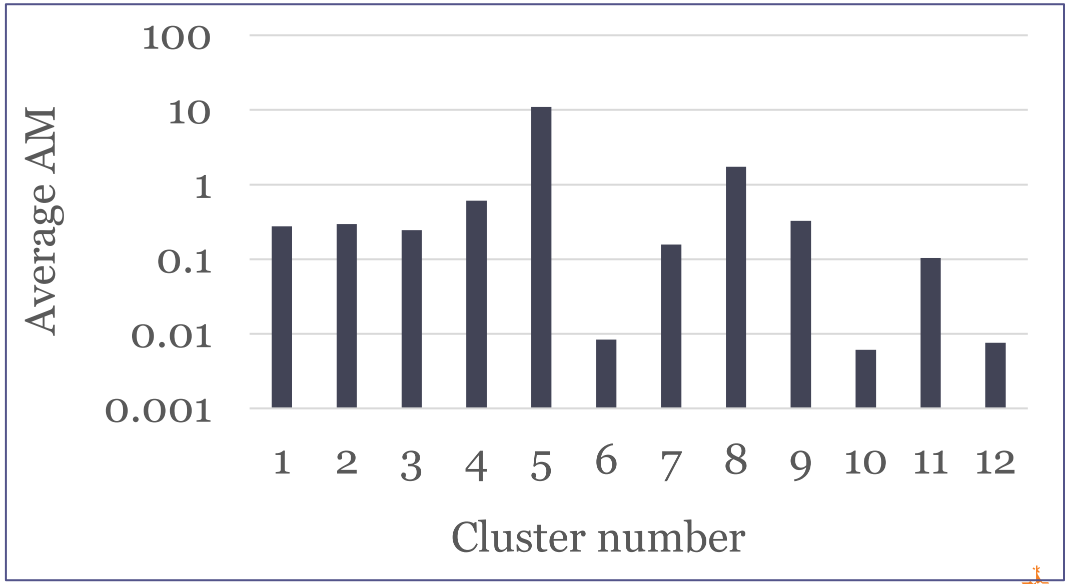


Even though most data were measured using ER probe, where data uncertainty can be 4 times of the erosion extent, the numbers of outliers in each cluster are less than 5%.

In most cases, the experimental data lies within the 95% CI. The length of the CI is of the same order of magnitude to the experimental data.

Sum(AM) = 14.71 (36% lower than clustering based on flow regimes (Sum(AM) = 23.03)





Our hybrid modeling efforts in flow assurance

Data clustering

Group data into similar sets
Identify best models for each set

Uncertainty quantification

Error analysis
Propagate input and data uncertainty

Discrepancy modeling

Machine learning
for capturing mismatch between
observation and first-principle model

Model refinement

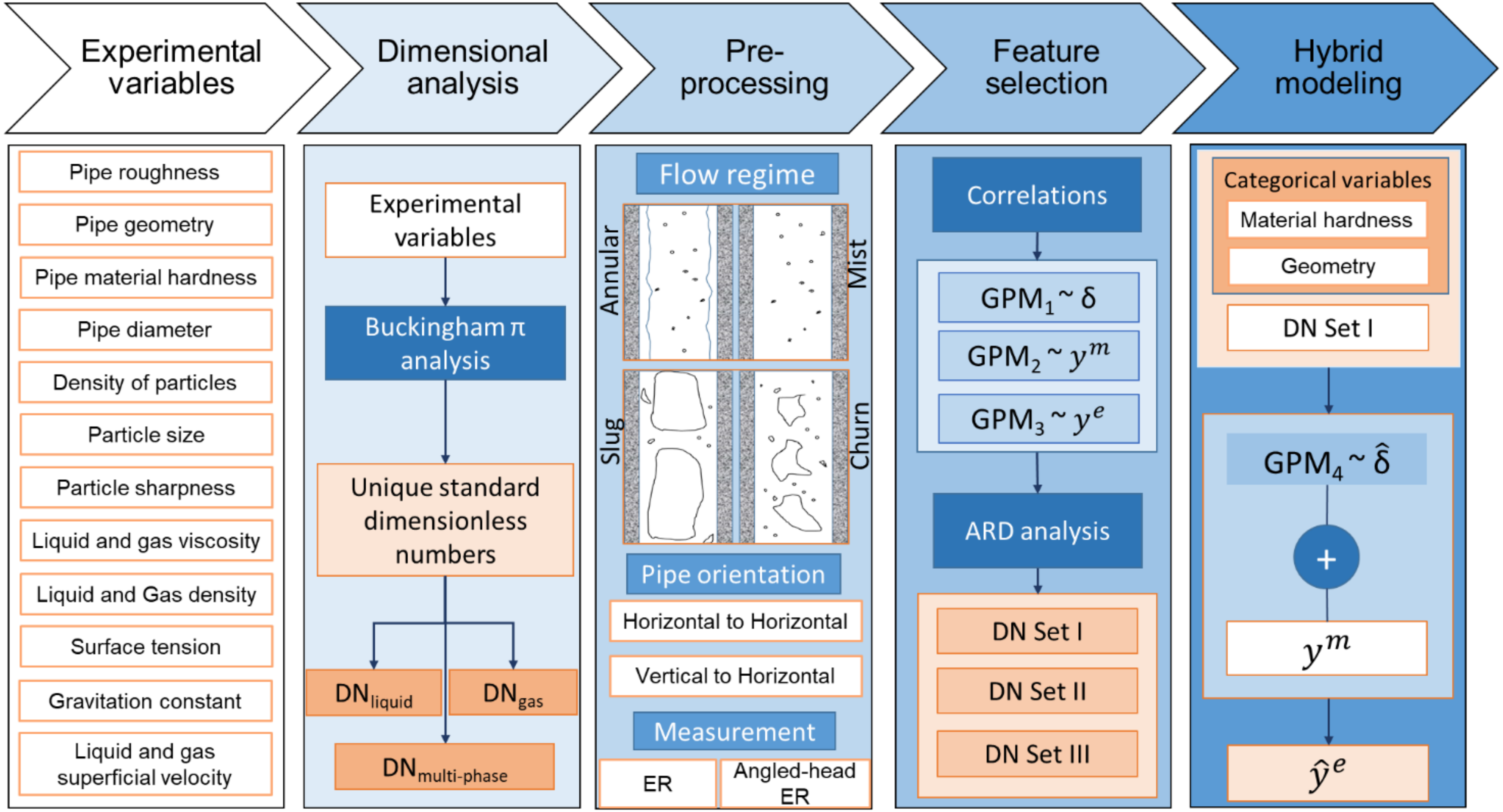
Machine learning suggested
experimental campaigns
and first-principle model refinement

Feature selection

Incorporating expert knowledge
to feature selection
for hybrid models

AUBURN UNIVERSITY

Can data-driven models inform semi-mechanistic models?



Dai et al. (2021). Computers & Chemical Engineering, 156, 107577.

Incorporating dimensional analysis to hybrid models

$$\delta = f\left(d_p, \rho_p, \mu_g, \rho_g, \mu_w, \rho_w, D, g, v_g, v_w, \Upsilon\right) \text{ - 11 variables}$$

$$\delta = f\left(\frac{dp}{D}, \frac{\rho_p D^2 g}{\Upsilon}, \frac{\rho_g D^2 g}{\Upsilon}, \frac{\rho_w D^2 g}{\Upsilon}, \frac{\mu_g\sqrt{Dg}}{\Upsilon}, \frac{\mu_w\sqrt{Dg}}{\Upsilon}, \frac{v_g}{\sqrt{Dg}}, \frac{v_w}{\sqrt{Dg}}\right) - \text{8 variables}$$

- **T: time**
- **M: mass**
- **L: length**

Original data: 585N×11D

Buckingham Pi analysis: 585N×648D

Unique dimensionless groups: 648D → 366D

d_p, v_g, γ	$\frac{dp}{D}$	$\frac{\rho_p d_p v_g^2}{\Upsilon}$	$\frac{\rho_w d_p v_g^2}{\Upsilon}$	$\frac{\rho_g d_p v_g^2}{\Upsilon}$	$\frac{v_w}{v_g}$	$\frac{\mu_w v_g}{\Upsilon}$	$\frac{\mu_g v_g}{\Upsilon}$	$\frac{d_p g}{v_g^2}$
d_p, v_w, ρ_w	$\frac{dp}{D}$	$\frac{\rho_g}{\rho_w}$	$\frac{\rho_p}{\rho_w}$	$\frac{v_w}{v_g}$	$\frac{\mu_w}{d_p\rho_w v_w}$	$\frac{\mu_g}{d_p\rho_w v_w}$	$\frac{\rho_w d_p v_w^2}{\Upsilon}$	$\frac{d_p g}{v_w^2}$

Remove transformed dimensionless groups: 366D →63D

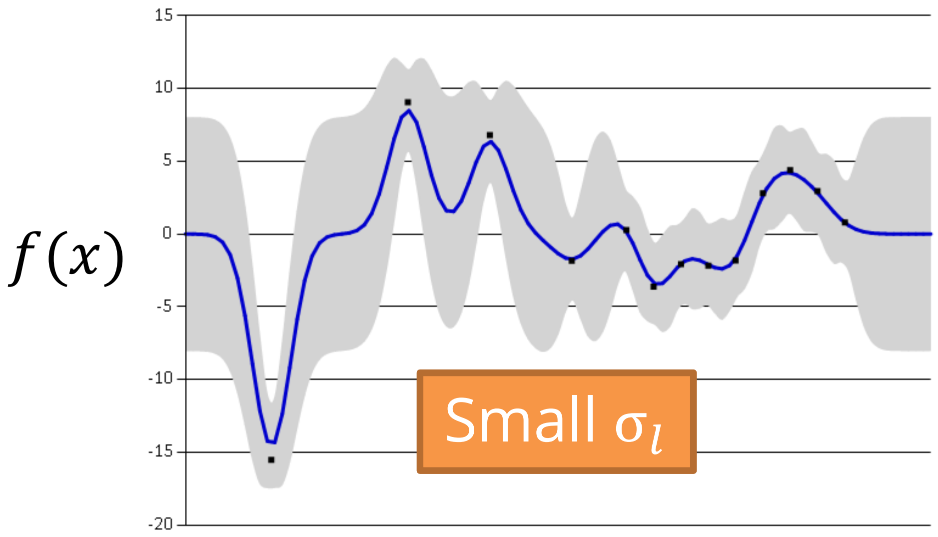
N^th root of dimensionless groups: $v_g\sqrt{\frac{d_p\rho_w}{\sigma}}$ $vs.$ $\frac{{v_g}^2 d_p\rho_w}{\sigma}$ (Weber number)

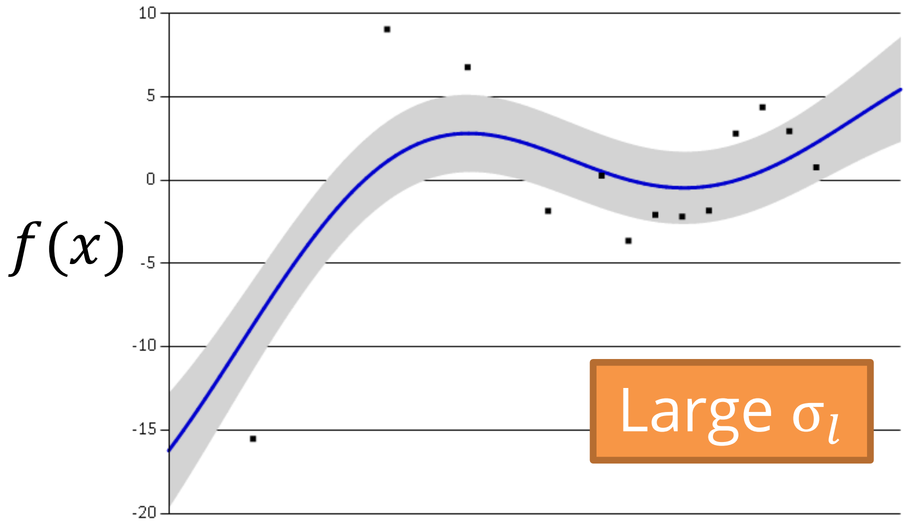
Dai et al. (2021). Computers & Chemical Engineering, 156, 107577.

Feature relevance using automatic relevance determination (ARD)

Use the squared exponential covariance function to select the most relevant inputs (ARD)

$$k\left(x_i, x_j \middle| \theta\right) = \sigma_f^2 \exp\left(-\frac{1}{2}\frac{\left(x_i - x_j\right)^2}{\sigma_l^2}\right)$$

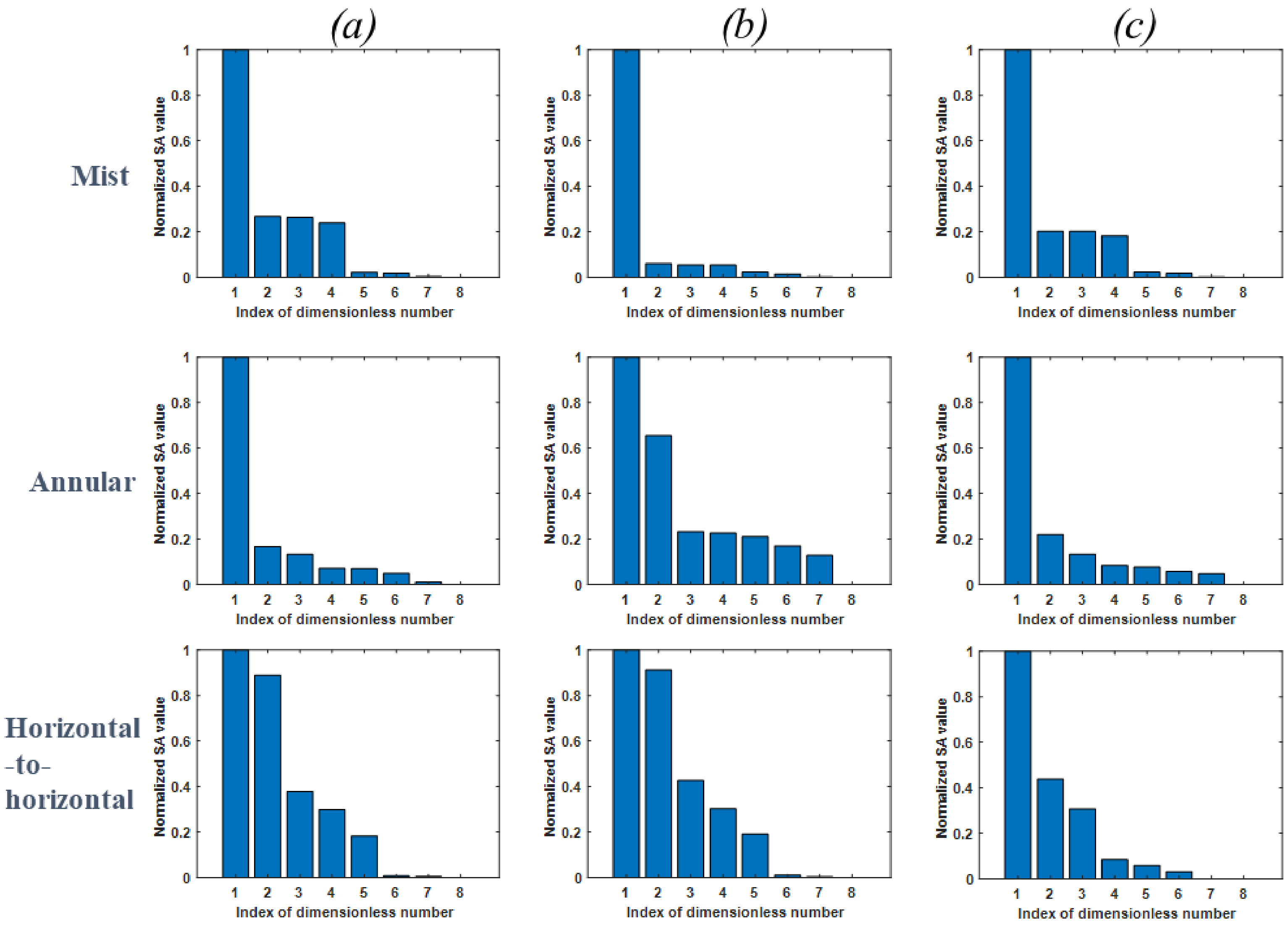




A smaller σ_l indicates a more relevant dimensionless input

Dai et al. (2021). Computers & Chemical Engineering, 156, 107577.

Example improvement suggestions for semi-mechanistic model



Sensitivity for mist and annular flow regimes and horizontal-to-horizontal pipeline orientation.

$(a)\,\delta \sim GPM_1$ (discrepancy)

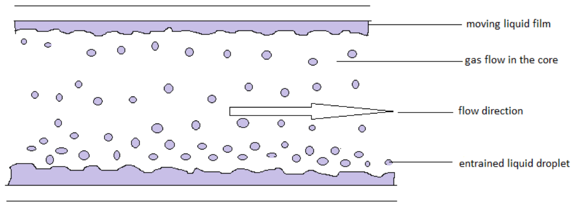
$(b)\,y^m \sim GPM_2$ (SPPS 1D model)

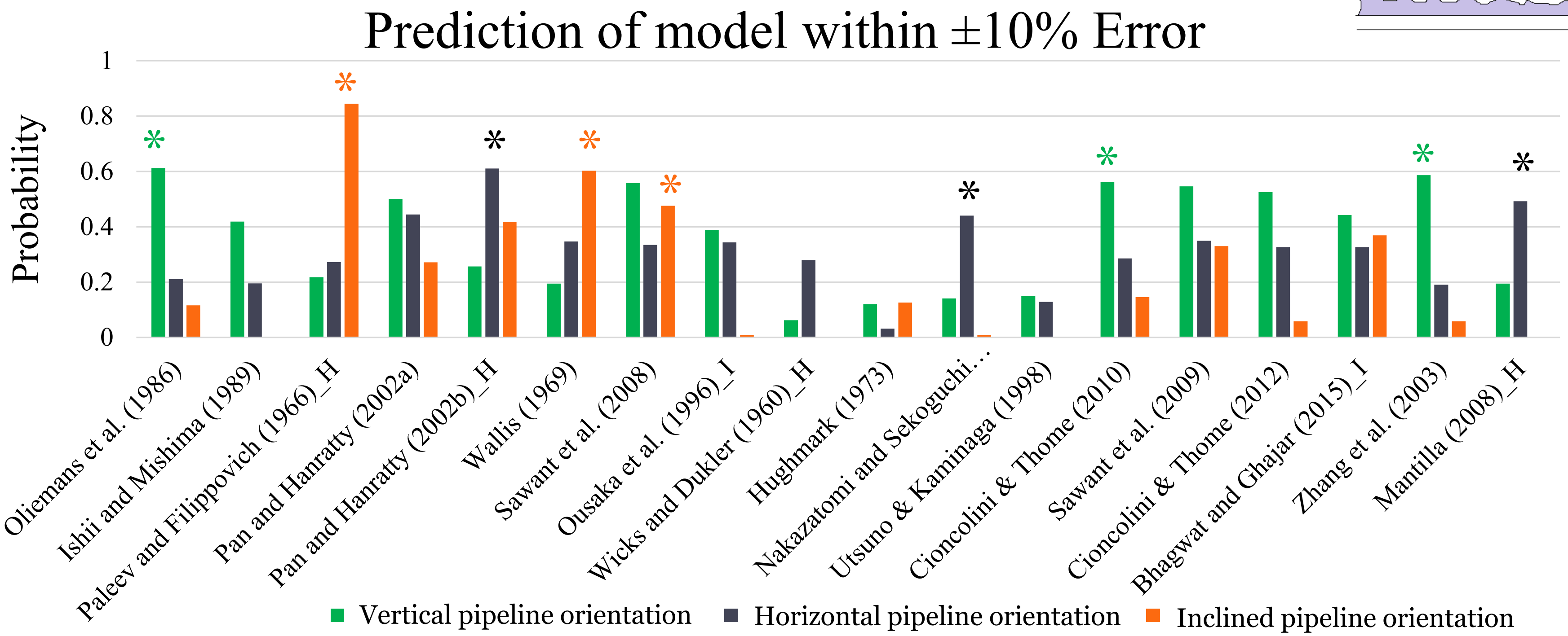
$(c)\,y^e \sim GPM_3$ (erosion data)

Dai et al. (2021). Computers & Chemical Engineering, 156, 107577.

Liquid entrainment fraction predictions of models differ significantly

Current models predict the entrainment fraction between 4% (very little liquid entrained) to 90% (almost all liquid entrained) at typical LNG conditions.





Horizontal pipeline orientation
- Pan and Hanratty (2002b)
- Mantilla (2008)
- Nakazatomi & Sekoguchi (1996)

Vertical pipeline orientation
- Oliemans et al. (1986)
- Zhang et al. (2003)
- Cioncolini & Thome (2010)

Inclined pipeline orientation
- Paleev & Filippovich (1966)
- Wallis (1969)
- Sawant et al. (2008)

Deng et al. (2022). Computers & Chemical Engineering, 162, 107796.

Hybrid model for entrainment fraction estimation

$$\boldsymbol{y^e}(\mathbf{x}) = \boldsymbol{y^m}(\mathbf{x}) + \boldsymbol{\delta}(\mathbf{x})$$

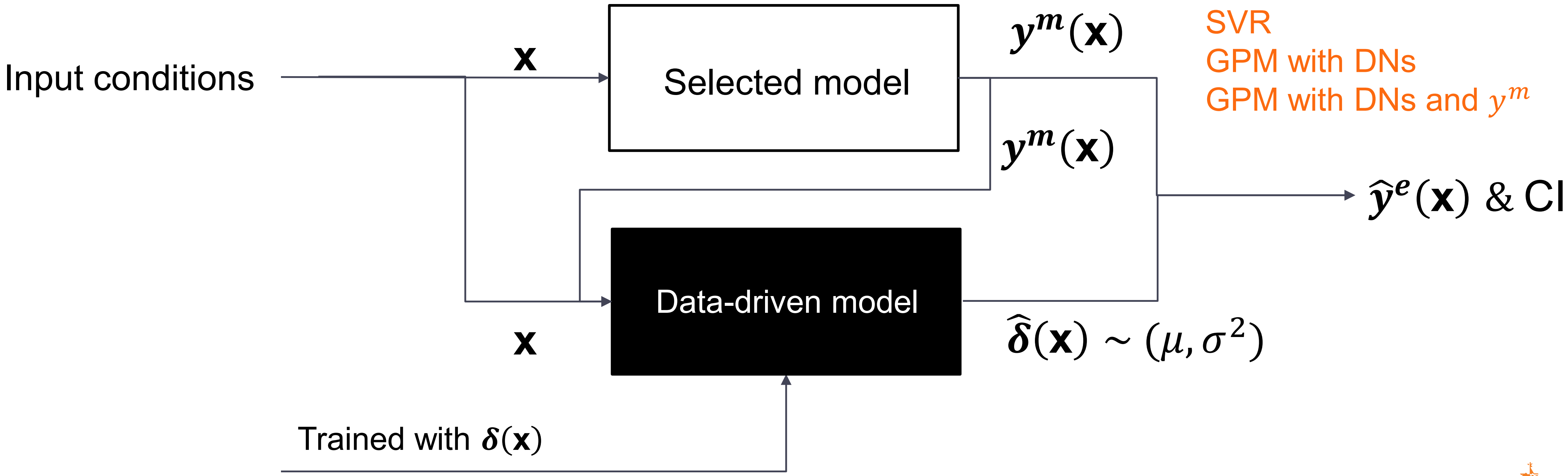
$\boldsymbol{y^e}(\mathbf{x})$- entrainment fraction measurement

$\boldsymbol{y^m}(\mathbf{x})$-entrainment fraction estimation from selected models

$\boldsymbol{\delta}(\mathbf{x})$-model discrepancy

Data driven modeling approaches

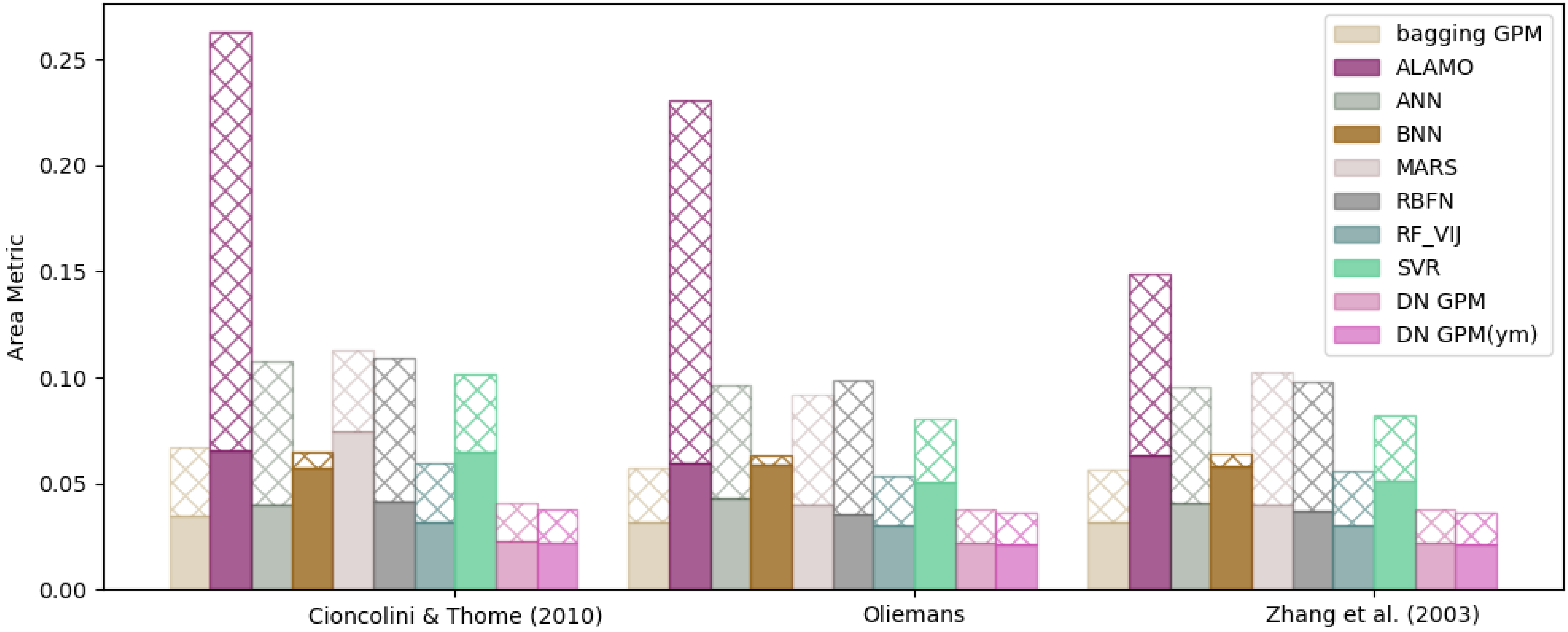
- bagging GPM
- ALAMO
- ANN
- BNN
- MARS
- RBFN
- RF
- SVR
- GPM with DNs
- GPM with DNs and y^m



AUBURN UNIVERSITY

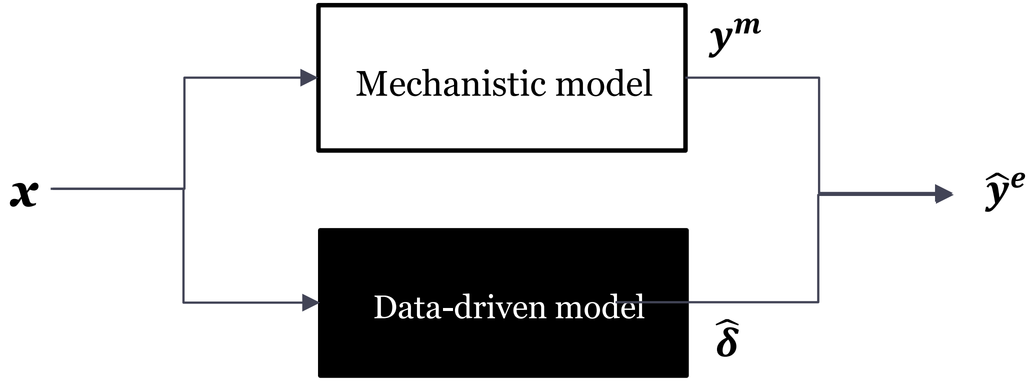
Using dimensionless numbers as inputs improves hybrid model entrainment fraction predictions

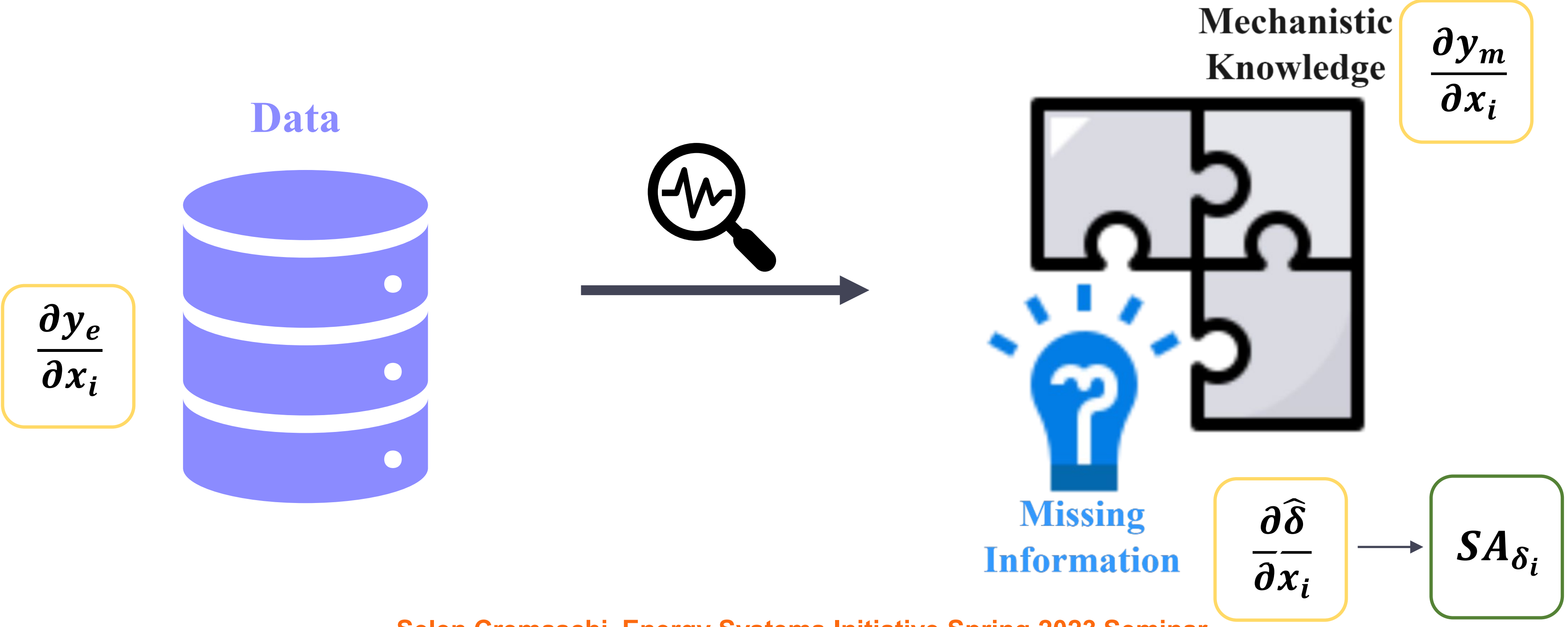
Vertical flow orientation



Missing information quantified using hybrid model

$$y^e(\mathrm{x}) = y^m(\mathrm{x}) + \delta(\mathrm{x})$$

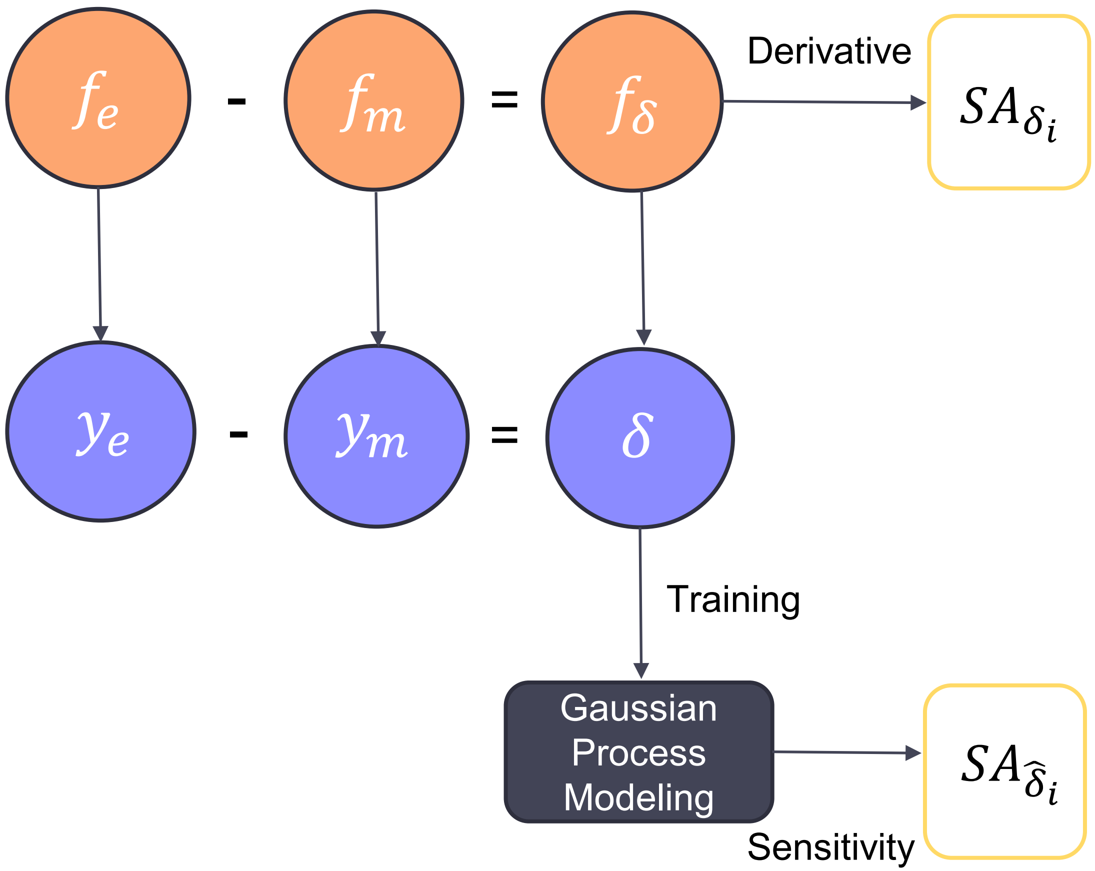




AUBURN UNIVERSITY

Validation experiments

$$f_e(X_1, \ldots, X_n) = f_m(X_1, \ldots, X_n) + f_\delta(X_1, \ldots, X_n)$$



Validation functions f_e and f_m are simulated using

- Sobol G-function[1]
- Polynomial function[2]
- Ishigami function[3]

Details of experiments[1]	
Number of variables	[3, 10]
Range of parameter change	0% - 100 %
Number of experiments	100

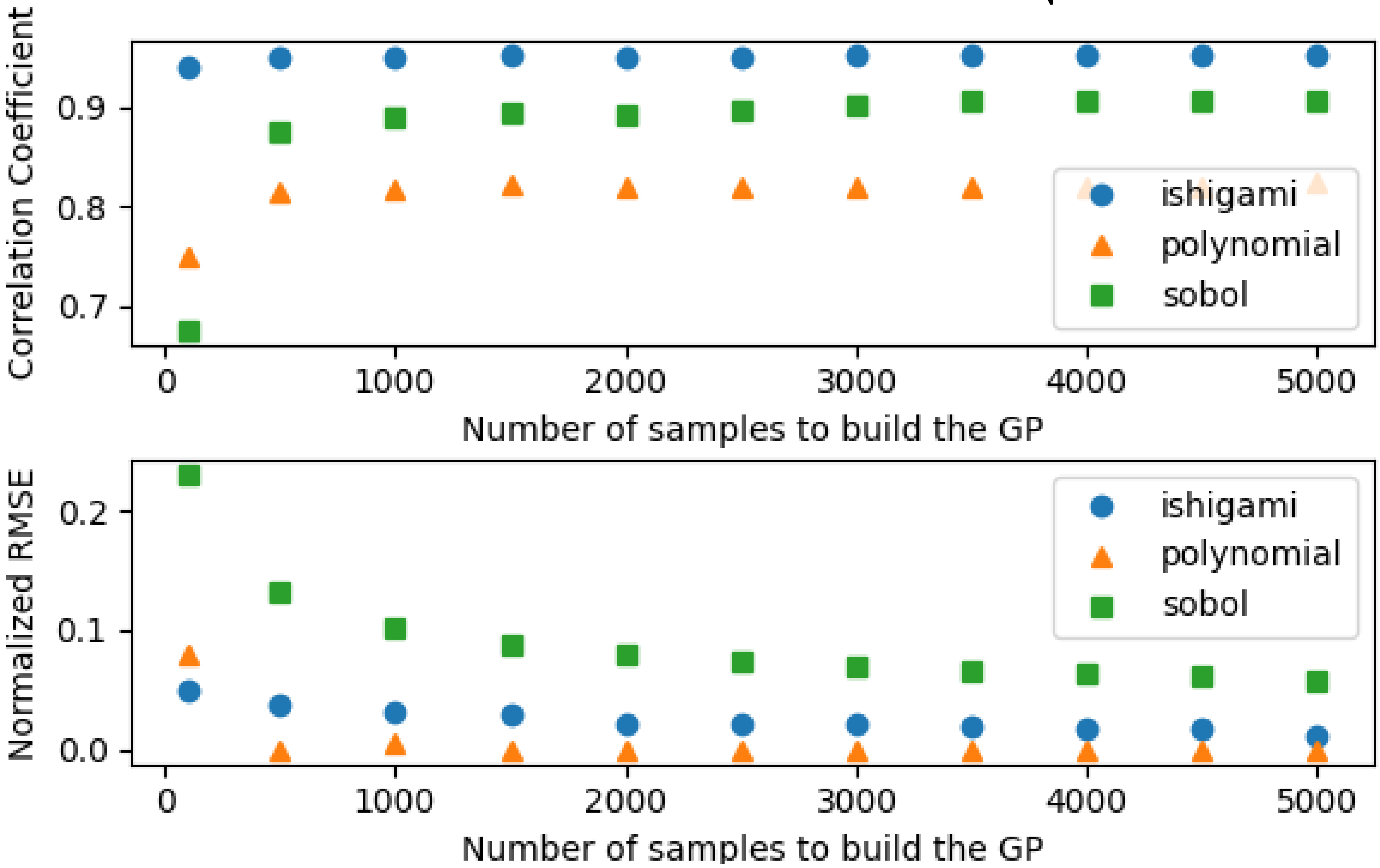
SA_{δ_i} is compared to the known analytical sensitivity $|V_{Ti}(f_e) - V_{Ti}\ (f_m)|$.

1. Saltelli, A., Annoni, P., Azzini, I., Campolongo, F., Ratto, M., & Tarantola, S. (2010). Variance based sensitivity analysis of model output. Design and estimator for the total sensitivity index. *Computer Physics Communications*
2. Mara, T.A., Tarantola, S., 2012. Variance-based sensitivity indices for models with dependent inputs. Reliab. Eng. Syst. Saf. 107, 115–121. https://doi.org/10.1016/J.RESS.2011.08.008
3. Kala, Z., 2018. Benchmark of goal oriented sensitivity analysis methods using Ishigami function. Int. J. Math. Comput. Methods 3.

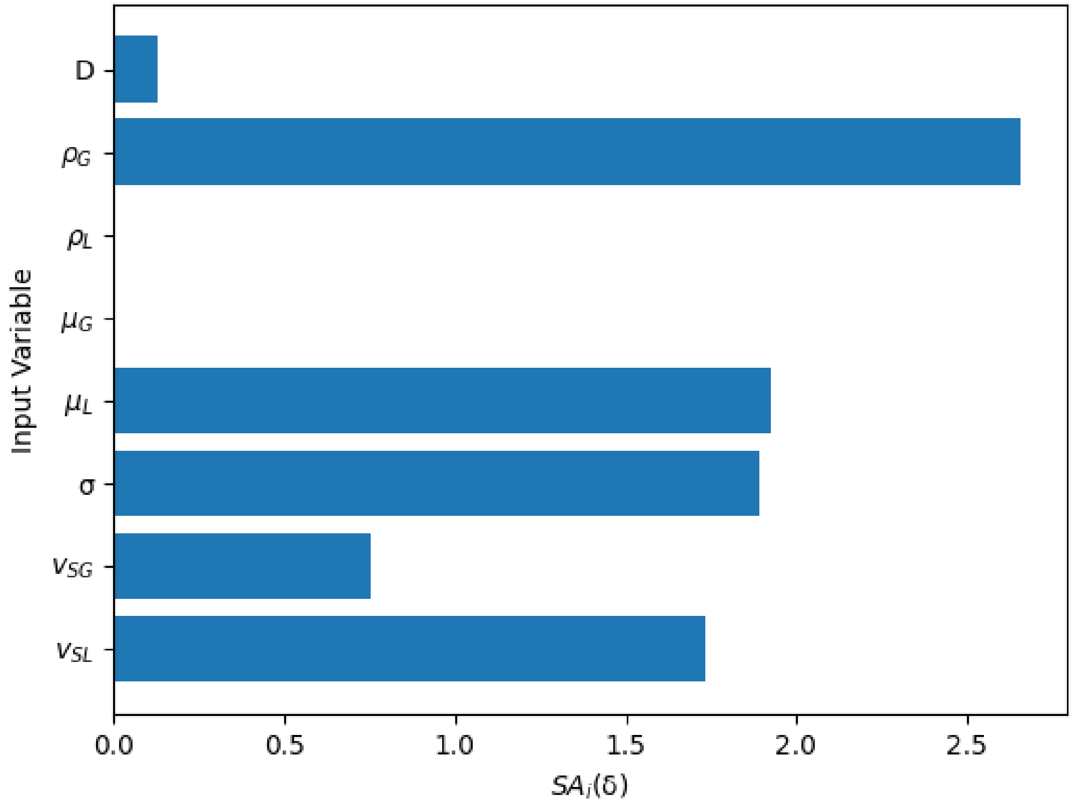
Validation experiment results

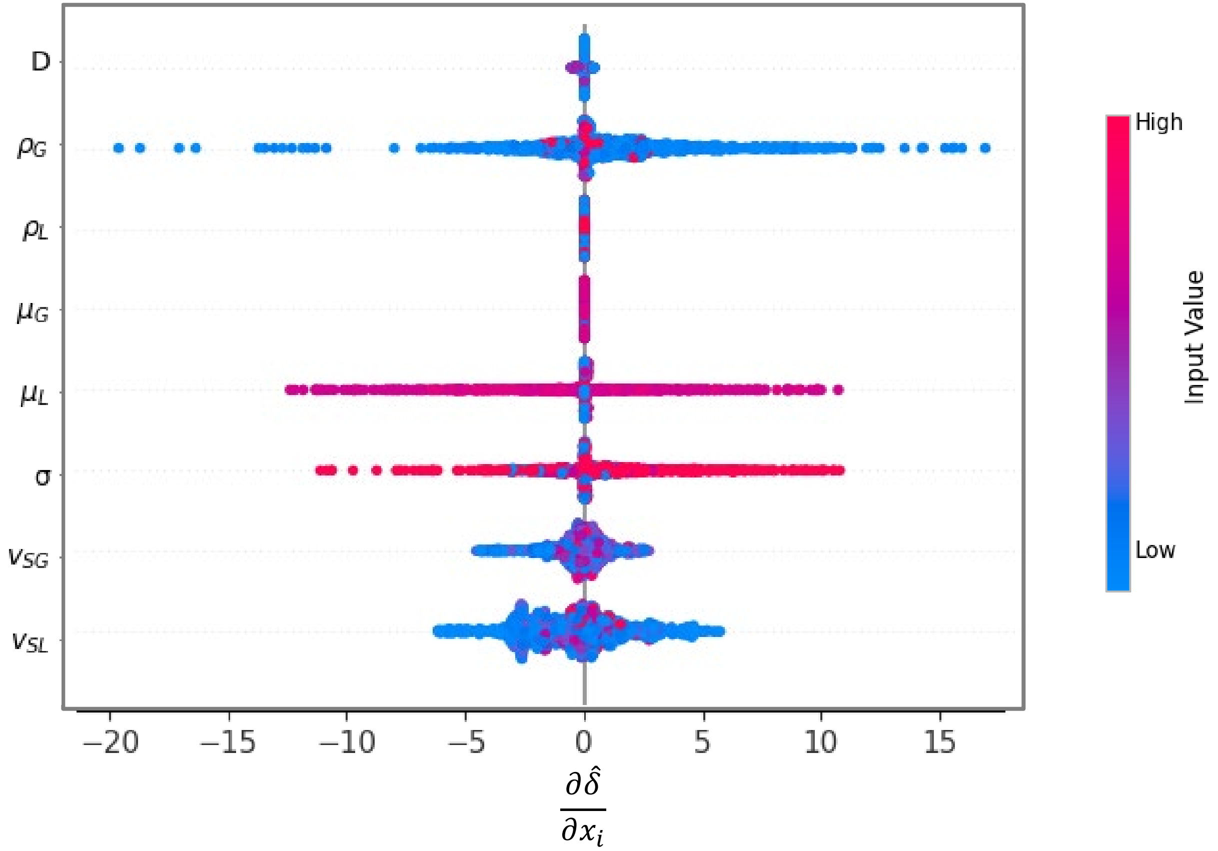
Correlation coefficient = $\mathrm{corr}(SA_{\hat{\delta}}, |V_{Ti}(f_e) - V_{Ti}(f_m)|)$

Normalized RMSE = $\sqrt{\frac{1}{p}(normalized\ SA_{\delta_i} - normalized\ SA_{\hat{\delta}_i})^2}$



Results – Zhang et al. (2003) model (vertical flow)



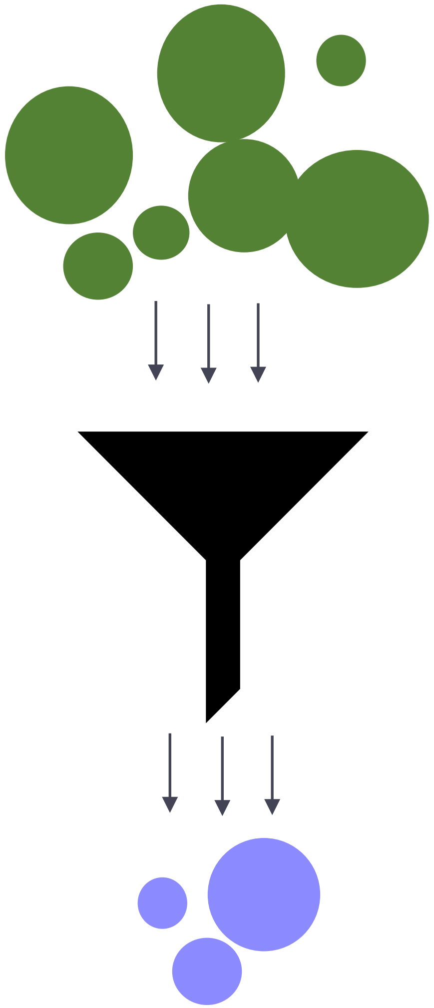


AUBURN UNIVERSITY

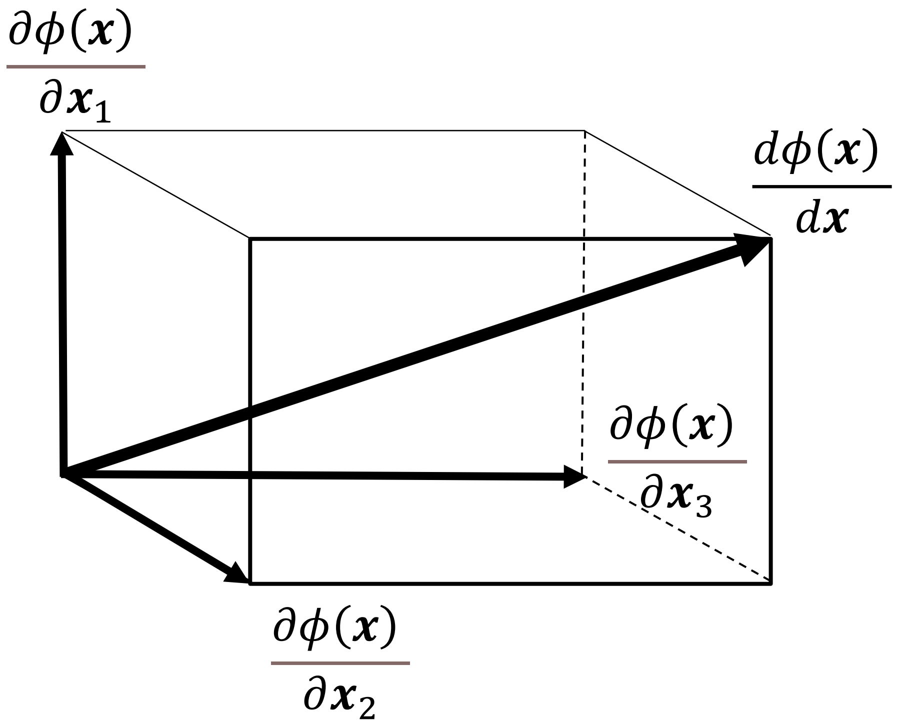
Can we quantify cumulative feature importance for better hybrid model performance?

The presence of irrelevant input variables result in unnecessary computational time, model overfitting and poor model performance.

Current Gaussian Process embedded feature selection methods are based on ranking results without a standard to discriminate relevant and irrelevant features.



Derivative decomposition ratio (DDR)



For n dimensional data, for each data point

$$\left\|\frac{d\phi}{dx}\right\| = \sqrt{\sum_{i=1}^{n}\left(\frac{\partial \phi}{\partial x_i}\right)^2}$$

The feature importance of the h^{th} input feature is defined as the derivative decomposition ratio (DDR)

$$DDR_h = \frac{\left|\frac{\partial \phi(\boldsymbol{x})}{\partial x_h^i}\right|^2}{\left|\frac{d\phi(\boldsymbol{x})}{d\boldsymbol{x}}\right|^2}$$

Two ways of calculating overall feature importance

	V_1	V_2	V_3	…	V_h	…	V_p
S_1							
S_2							
…							
S_i							
…							
S_N							

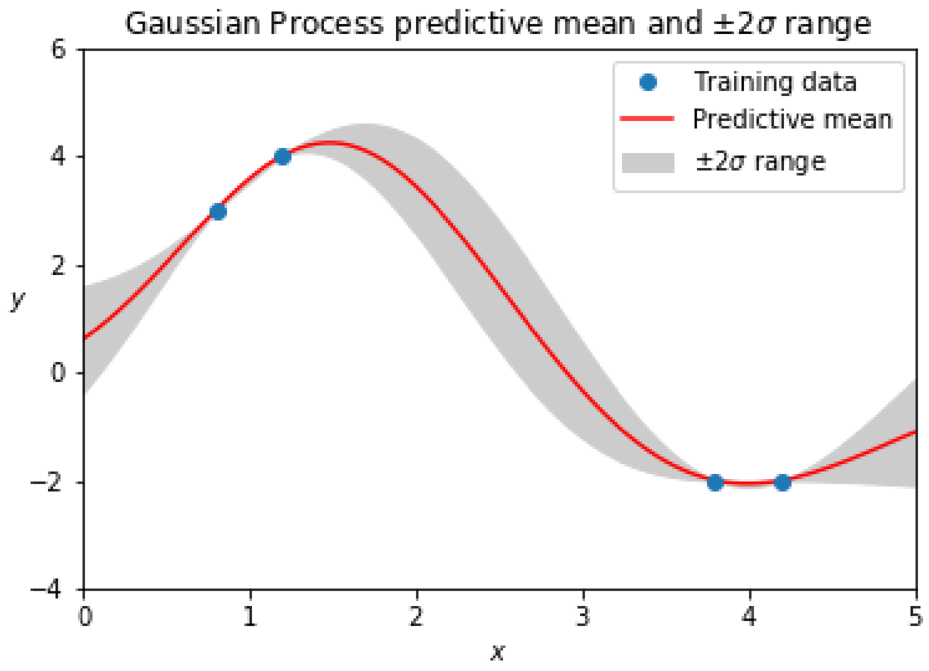
	V_1	V_2	V_3	…	V_h	…	V_p
S_1							
S_2							
…							
S_i							
…							
S_N							

$$\text{Averaged } DDR_h = \frac{1}{N}\sum_{i=1}^{N}\frac{\left|\frac{\partial\phi(\boldsymbol{x})}{\partial x_h^i}\right|^2}{\left|\frac{d\phi(\boldsymbol{x})}{d\boldsymbol{x}}\right|^2}$$

$$s_h = \frac{1}{N}\sum_{n=1}^{N}\left(\frac{\partial\phi(\boldsymbol{x}_n)}{\partial x_n^h}\right)^2$$

$$NS_h = \frac{s_h}{\sum_{h=1}^{H} s_h}$$

DDR is applied to Gaussian Process[1]



$$\boldsymbol{\delta} \sim \mathcal{N}\left(\overline{f_*}, cov(f_*)\right)$$

Substitute the mean function into the DDR

$$\frac{\partial \overline{f_*}(\boldsymbol{x}_n)}{\partial x_n^h} = \frac{\alpha_p\left(x_p^h - x_q^h\right)}{l_h^2} k(\boldsymbol{x}_p, \boldsymbol{x}_q)$$

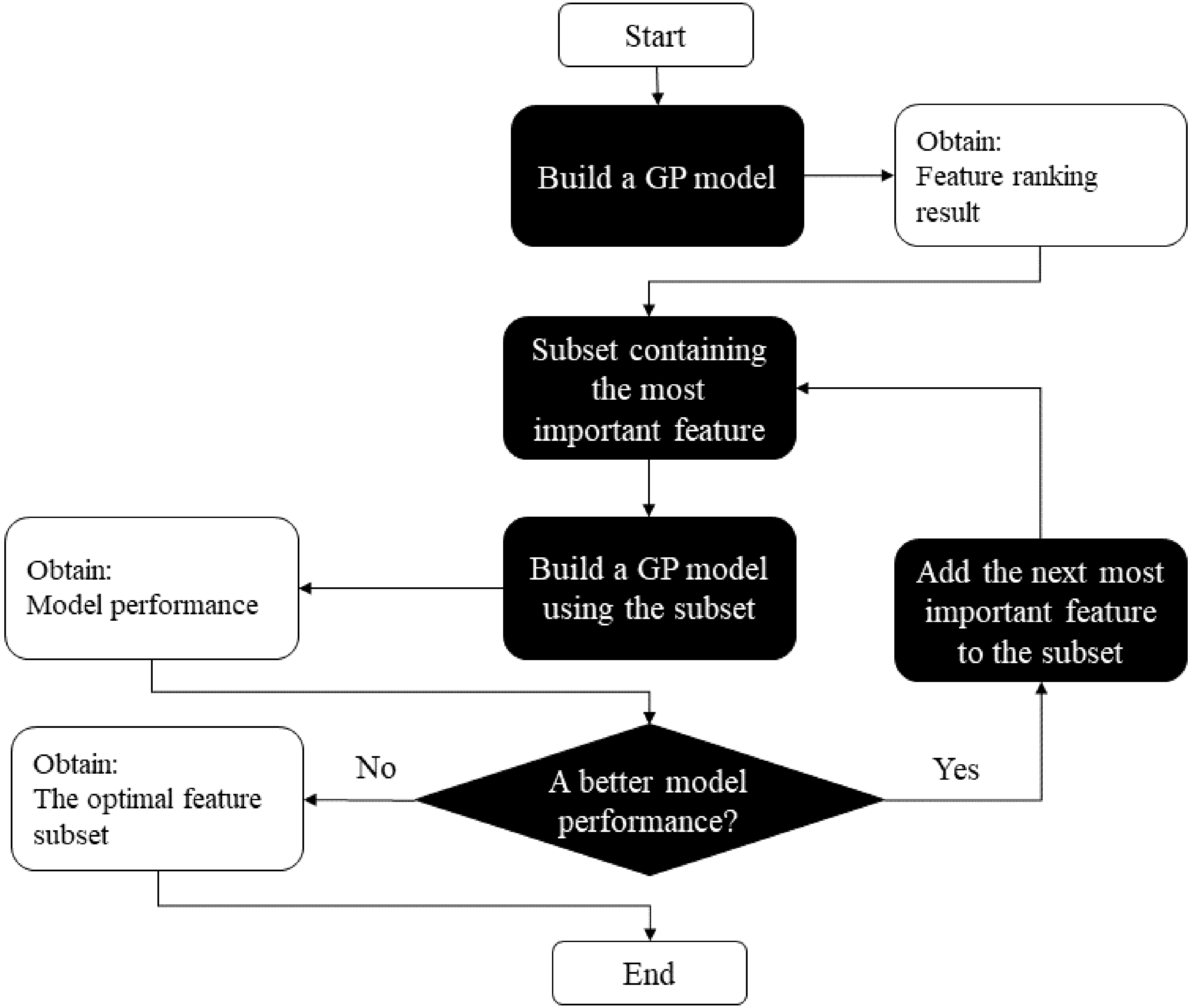
Averaged

$$DDR_h = \frac{1}{N}\sum_{i=1}^{N} \frac{\left(\sum_{p=1}^{N} \frac{\alpha_p\left(x_p^h - x_q^h\right)}{l_h^2} k(\boldsymbol{x}_p, \boldsymbol{x}_q)\right)^2}{\sum_{h=1}^{H}\left(\sum_{p=1}^{N} \frac{\alpha_p\left(x_p^h - x_q^h\right)}{l_h^2} k(\boldsymbol{x}_p, \boldsymbol{x}_q)\right)^2}$$

$$NS_h = \frac{\left(\sum_{p=1}^{N} \frac{\alpha_p\left(x_p^h - x_q^h\right)}{l_h^2} k(\boldsymbol{x}_p, \boldsymbol{x}_q)\right)^2}{\sum_{h=1}^{H}\left(\sum_{p=1}^{N} \frac{\alpha_p\left(x_p^h - x_q^h\right)}{l_h^2} k(\boldsymbol{x}_p, \boldsymbol{x}_q)\right)^2}$$

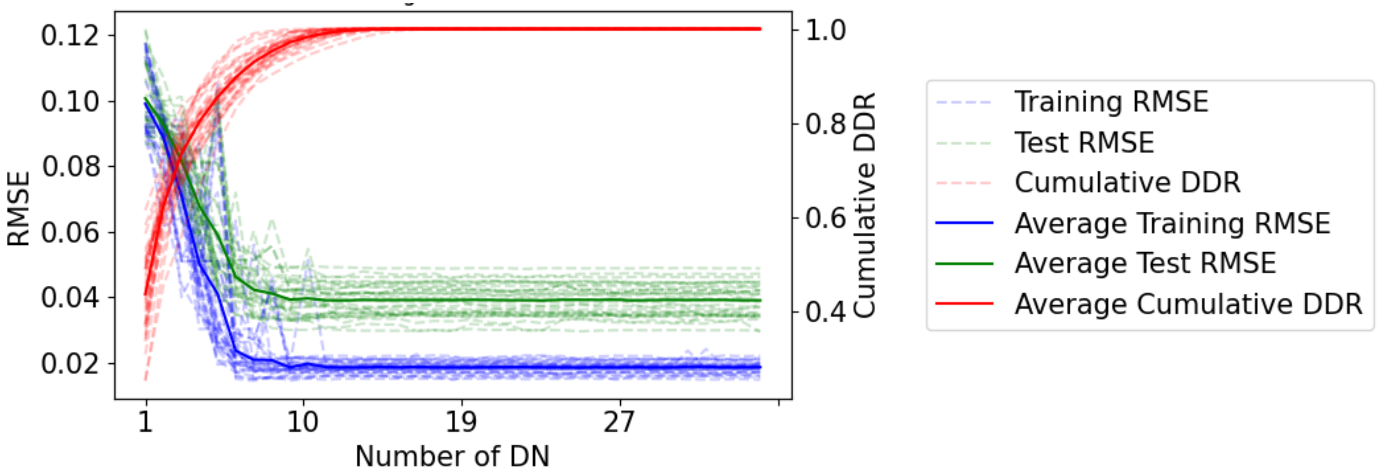
1. Rasmussen, C. E., Williams, C. K., 2006, Gaussian Process for Machine Learning, *The MIT Press*.

Feature selection validation experiment

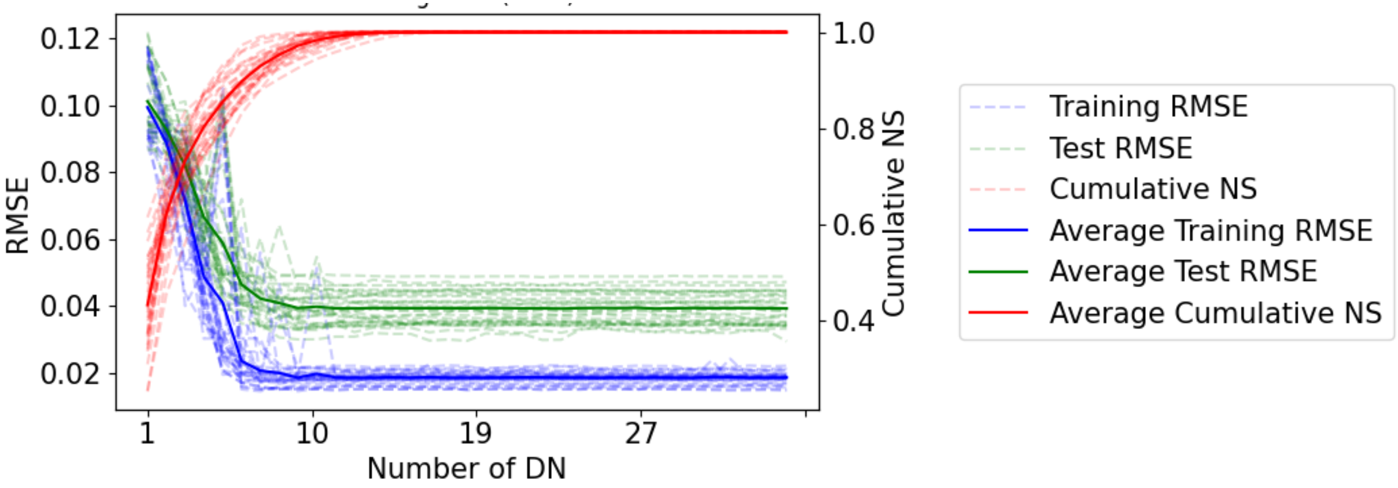


Validation experiments results – vertical flow

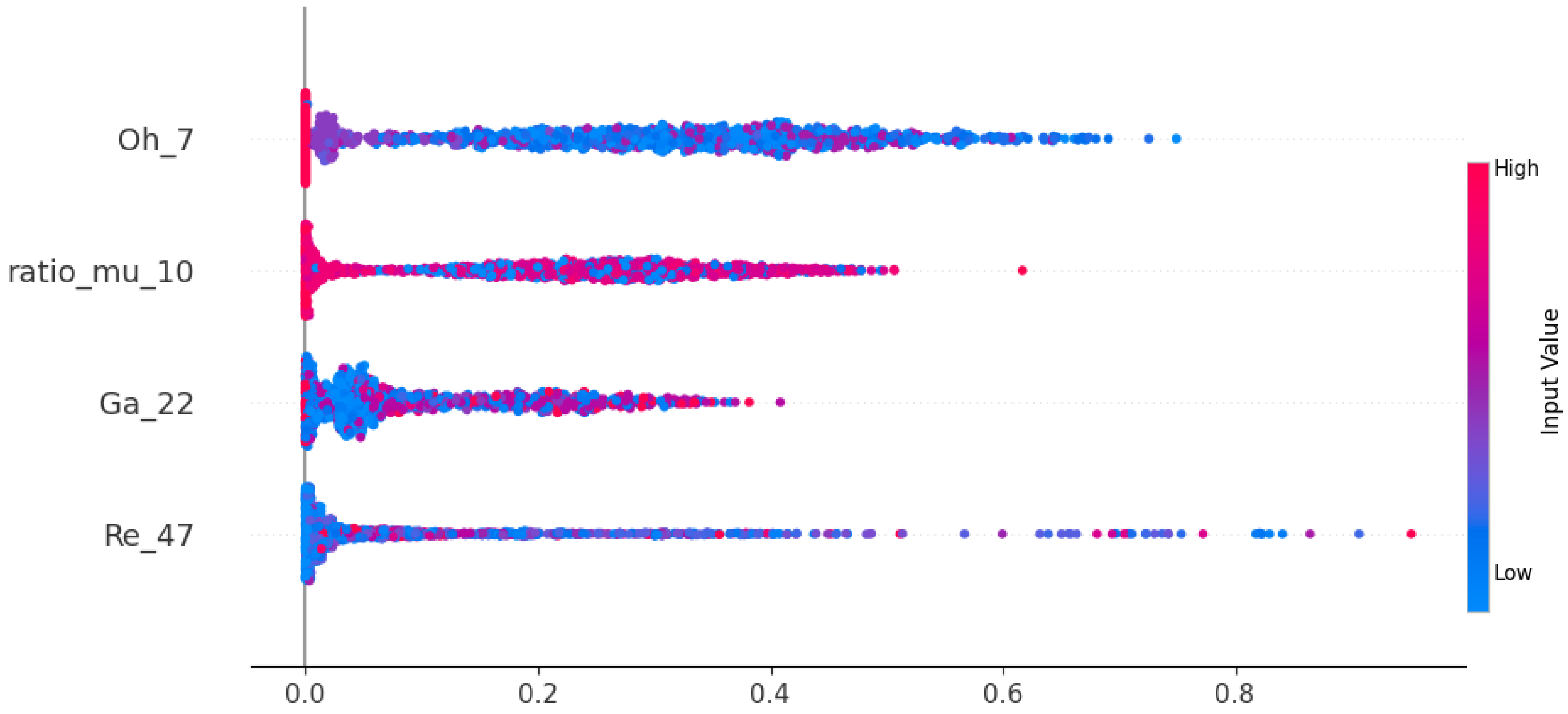
Cumulative Derivative Decomposition Ratio (DDR)



Cumulative Normalized Sensitivity (NS)



Local feature importance heatmap – vertical flow



$$\left|\frac{\partial \overline{f}_*(\boldsymbol{x}_n)}{\partial x_n^h}\right|$$

DNs with physical explanation	
Oh_7	$\frac{\mu_L}{\sqrt{\sigma d \rho_L}}$
Ratio_mu_10	$\frac{\mu_G}{\mu_L}$

DNs without physical explanation	
Ga_22	$\frac{D^3 {\rho_L}^2 g}{{\mu_G}^2}$
Re_47	$\frac{d \rho_G V_{SL}}{\mu_G}$

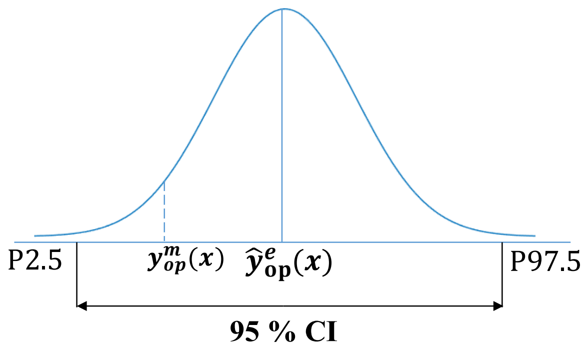
AUBURN UNIVERSITY

Conclusions

- Hybrid modeling techniques that integrate domain knowledge with machine learning techniques has great potential to increase the confidence in predictions of commonly used models in flow assurance problems
  - Critical velocity predictions for sand transport
  - Erosion extend predictions for pipelines
  - Liquid entrainment predictions in conduits
- Identifying the right data set for the application is essential.
- Gaussian Process modeling is powerful for regression applications with limited data sets.
- Incorporating domain knowledge is necessary for increased confidence in hybrid model predictions.
- Selecting the right input set for flow assurance applications requires expert knowledge.
- Feature selection capabilities of machine learning models can aid in identifying regions for further experiments and semi-mechanistic model refinement.

Auburn University Quantification of Uncertainty in Erosion by Sand in pipe Ties (AUQUEST)

$$\boldsymbol{y^e}(\mathrm{x}) = \boldsymbol{y^m}(\mathrm{x}) + \boldsymbol{\delta}(\mathrm{x}) + \boldsymbol{\varepsilon}$$



$\boldsymbol{x_{op}}, \boldsymbol{y^m_{op}(x)}$ → **AUQUEST** (A, B) → $\hat{\boldsymbol{\delta}}_{op}(\mathrm{x})$@P2.5, P50, P97.5, $\hat{\boldsymbol{\varepsilon}}_{op}$,

$$\hat{\boldsymbol{y}}^e_{op}(\boldsymbol{x}) = \boldsymbol{y}^m_{op}(\boldsymbol{x}) + \hat{\boldsymbol{\delta}}_{op}(\mathrm{x}) + \hat{\boldsymbol{\varepsilon}}_{op}$$

$\boldsymbol{x_{exp}}, \boldsymbol{y^m_{exp}(x)}, \boldsymbol{y^e_{exp}(x)}$ → A

$\boldsymbol{x_{exp}}, \boldsymbol{y^e_{exp}(x)}$ → B

A: estimate $\hat{\delta}_t(x)$ and Sign($\hat{\delta}(x)$): $\hat{\delta}(x)$ = $\hat{\delta}_t(x)$ Sign($\hat{\delta}(x)$)

- **Pre-process data**
  - Normalization x
  - Log transformation $\delta_t(x) = \log(\delta(x))$
  - Sign of $\delta(x)$
- **Cluster data**
  - Flow regime
  - Measurement approaches (MA)
  - OCIL and density-based initialization
- Dimensional analysis? yes →
- **Generate dimensionless numbers (DNs)**
  - Buckingham Pi analysis
  - ARD for selecting the most influential DNs
- no →
- **Train GPMs for estimating $\hat{\delta}_t(x)$ and Sign($\hat{\delta}(x)$) for each cluster**
  - Training options
  - -Kernel functions: squared exponential, neural network or spectral mixture function
  - -Optimization methods: conjugate gradient or Bayesian optimization algorithm

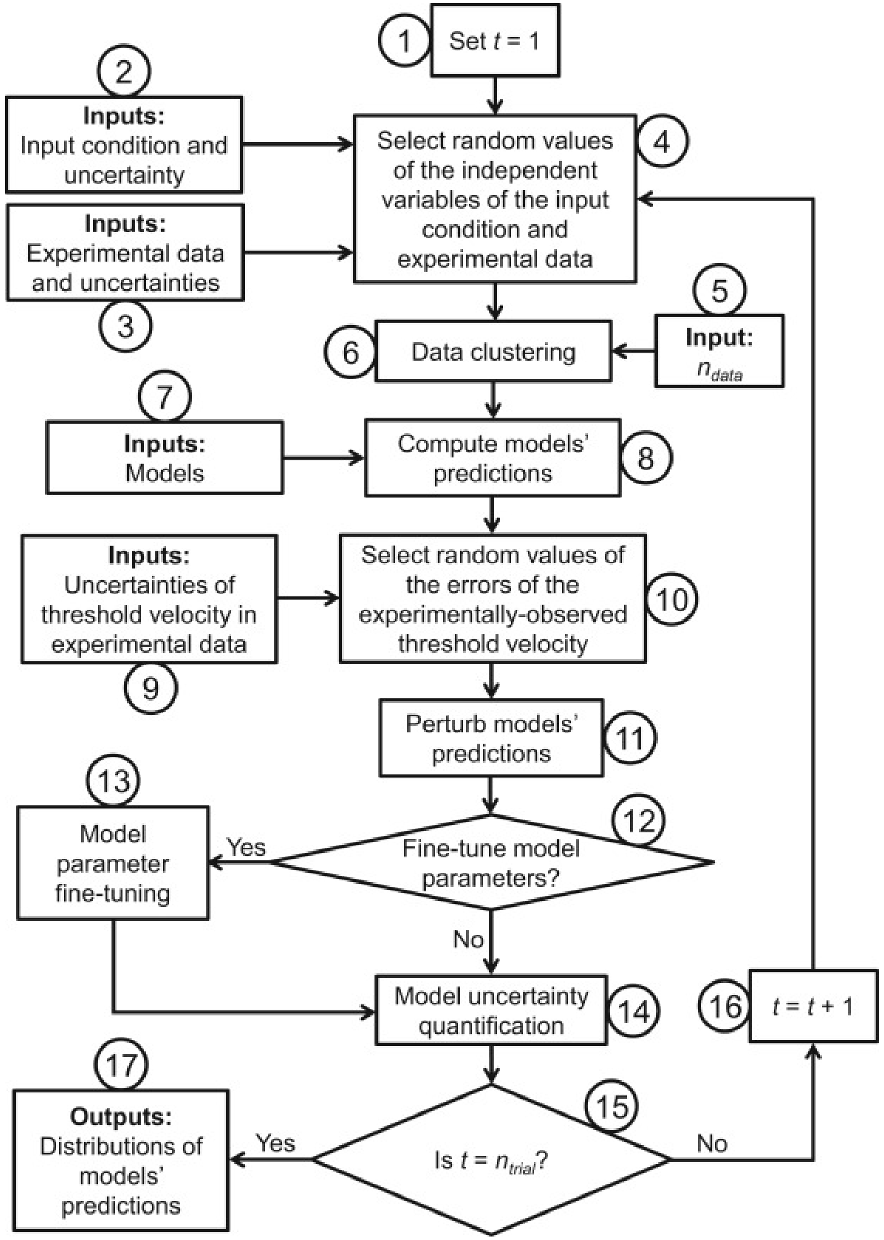
→ **GPMs**

B: estimate $\hat{\varepsilon}$

1. Separate the data based on MA
2. Find S_{MA} for each MA
3. Identify best fit and calculate MR
4. Combine MR of all the sets for each MA
5. Construct histograms of MRs
6. Fit histograms with kernel distributions

→ **Distribution of $\hat{\varepsilon}_{MA}$**

$\boldsymbol{x}$: operating conditions
$\boldsymbol{y^e}(x)$: experimental measurement ($\hat{\boldsymbol{y}}^e(\boldsymbol{x})$: estimation)
$\boldsymbol{y^m}(x)$: 1-D SPPS prediction at x
$\boldsymbol{\delta}(x)$: model discrepancy ($\hat{\boldsymbol{\delta}}(\mathrm{x})$: estimation)
$\boldsymbol{\varepsilon}$: data uncertainty ($\hat{\boldsymbol{\varepsilon}}$: estimation)
ARD: automatic relevance determination
DN: dimensionless numbers
GPM: Gaussian process modeling
MA: measurement approaches
MR: modified residual
OCIL: Iterative clustering learning based on object-cluster similarity metric
S_{MA}: a series of data points from the same measurement approach with only one variable changing



AUBURN UNIVERSITY

The most relevant dimensionless numbers identified for mist flow

	$\delta \sim$ GPM₁		$y^m \sim$ GPM₂		$y^e \sim$ GPM₃	
	Number	**Definition**	**Number**	**Definition**	**Number**	**Definition**
1	Oh	$\frac{\mu_g}{\sqrt{\sigma d_p \rho_l}}$	Oh	$\frac{\mu_g}{\sqrt{\sigma d_p \rho_l}}$	Oh	$\frac{\mu_g}{\sqrt{\sigma d_p \rho_l}}$
2	Fr	$\frac{v_g}{\sqrt{Dg}}$	We	$\frac{D\rho_l V_g^2}{\sigma}$	Fr	$\frac{v_g}{\sqrt{Dg}}$
3	We	$\frac{D\rho_l V_g^2}{\sigma}$	Fr	$\frac{v_g}{\sqrt{Dg}}$	We	$\frac{D\rho_l V_g^2}{\sigma}$
4	Fr and Re	$\frac{\mu_l g}{\rho_l v_g^3}$	Fr and Re	$\frac{\mu_l g}{\rho_l v_g^3}$	Fr and Re	$\frac{\mu_l g}{\rho_l v_g^3}$
5	Oh	$\frac{\mu_g}{\sqrt{\sigma D \rho_g}}$	Oh	$\frac{\mu_g}{\sqrt{\sigma D \rho_g}}$	Oh	$\frac{\mu_g}{\sqrt{\sigma D \rho_g}}$
6	Oh	$\frac{\mu_g}{\sqrt{\sigma d_p \rho_g}}$	Oh	$\frac{\mu_g}{\sqrt{\sigma d_p \rho_g}}$	Oh	$\frac{\mu_g}{\sqrt{\sigma d_p \rho_g}}$
7	Bo	$\frac{gD^2\rho_g}{\sigma}$	Bo	$\frac{gD^2\rho_g}{\sigma}$	Bo	$\frac{gD^2\rho_g}{\sigma}$
8	Oh	$\frac{\mu_l}{\sqrt{\sigma d_p \rho_g}}$	Oh	$\frac{\mu_l}{\sqrt{\sigma d_p \rho_g}}$	Oh	$\frac{\mu_l}{\sqrt{\sigma d_p \rho_g}}$

The most relevant dimensionless numbers identified for annular flow

	$\delta \sim$ GPM1		$y^m \sim$ GPM2		$y^e \sim$ GPM3	
	Number	Definition	Number	Definition	Number	Definition
1	Oh	$\frac{\mu_l}{\sqrt{\sigma d_p \rho_g}}$	Oh	$\frac{\mu_l}{\sqrt{\sigma d_p \rho_g}}$	Oh	$\frac{\mu_l}{\sqrt{\sigma d_p \rho_g}}$
2	We	$\frac{d_p \rho_g V_g^2}{\sigma}$	Oh	$\frac{\mu_g}{\sqrt{\sigma D \rho_l}}$	Oh	$\frac{\mu_g}{\sqrt{\sigma D \rho_l}}$
3	Fr	$\frac{v_g}{\sqrt{Dg}}$	Bo	$\frac{g D^2 \rho_g}{\sigma}$	**Fr and Re**	$\boldsymbol{\frac{\mu_l g}{\rho_g v_g^3}}$
4	We	$\frac{d_p \rho_l V_g^2}{\sigma}$	Fr	$\frac{v_g}{\sqrt{Dg}}$	Bo	$\frac{g D^2 \rho_g}{\sigma}$
5	Oh	$\frac{\mu_g}{\sqrt{\sigma D \rho_l}}$	We	$\frac{d_p \rho_g V_g^2}{\sigma}$	**Mo**	$\boldsymbol{\frac{g \mu_g^4}{\rho_g \sigma^3}}$
6	**Ca=$\frac{We}{Re}$**	$\boldsymbol{\frac{\mu_l V_l}{\sigma}}$	**Fr and Re**	$\boldsymbol{\frac{\mu_l g}{\rho_g v_g^3}}$	Fr	$\frac{v_g}{\sqrt{Dg}}$
7	Bo	$\frac{g D^2 \rho_g}{\sigma}$	**Mo**	$\boldsymbol{\frac{g \mu_g^4}{\rho_g \sigma^3}}$	We	$\frac{d_p \rho_g V_g^2}{\sigma}$
8	**We**	$\boldsymbol{\frac{D \rho_g V_g^2}{\sigma}}$	We	$\frac{d_p \rho_l V_g^2}{\sigma}$	We	$\frac{d_p \rho_l V_g^2}{\sigma}$

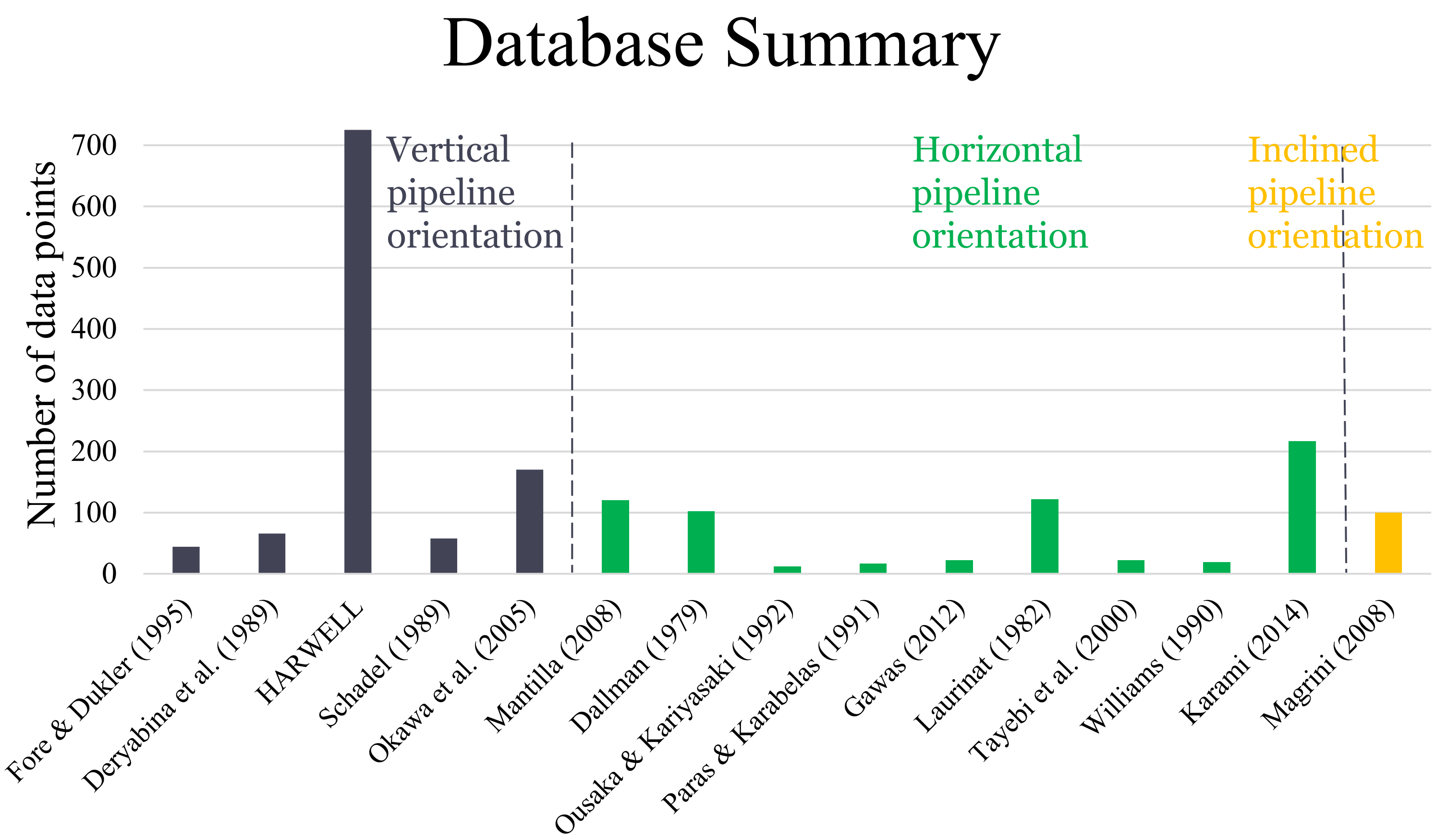
AUBURN UNIVERSITY

The most relevant dimensionless numbers identified for data collected in horizontal to horizontal orientation

	$\delta \sim$ GPM₁		$y^m \sim$ GPM₂		$y^e \sim$ GPM₃	
	Number	**Definition**	**Number**	**Definition**	**Number**	**Definition**
1	Oh	$\frac{\mu_l}{\sqrt{\sigma d_p \rho_l}}$	Ratio of diameters	$\frac{d_p}{D}$	Ratio of diameters	$\frac{d_p}{D}$
2	Ratio of diameters	$\frac{d_p}{D}$	Oh	$\frac{\mu_l}{\sqrt{\sigma d_p \rho_l}}$	Fr	$\frac{v_g}{\sqrt{Dg}}$
3	Fr	$\frac{v_g}{\sqrt{Dg}}$	Fr	$\frac{v_g}{\sqrt{Dg}}$	Oh	$\frac{\mu_l}{\sqrt{\sigma d_p \rho_l}}$
4	Oh	$\frac{\mu_l}{\sqrt{\sigma d_p \rho_g}}$	Oh	$\frac{\mu_l}{\sqrt{\sigma d_p \rho_g}}$	Oh	$\frac{\mu_l}{\sqrt{\sigma d_p \rho_g}}$
5	We	$\frac{d_p \rho_g V_g^2}{\sigma}$	We	$\frac{d_p \rho_g V_g^2}{\sigma}$	Oh	$\frac{\mu_l}{\sqrt{\sigma D \rho_g}}$
6	Oh	$\frac{\mu_l}{\sqrt{\sigma D \rho_g}}$	Oh	$\frac{\mu_l}{\sqrt{\sigma D \rho_g}}$	We	$\frac{d_p \rho_g V_g^2}{\sigma}$
7	Re	$\frac{\mu_l}{D \rho_g v_g}$	Re	$\frac{\mu_l}{D \rho_g v_g}$	Ratio of velocities	$\frac{v_g}{v_l}$
8	Ratio of velocities	$\frac{v_g}{v_l}$	Ratio of velocities	$\frac{v_g}{v_l}$	Re	$\frac{\mu_l}{D \rho_g v_g}$

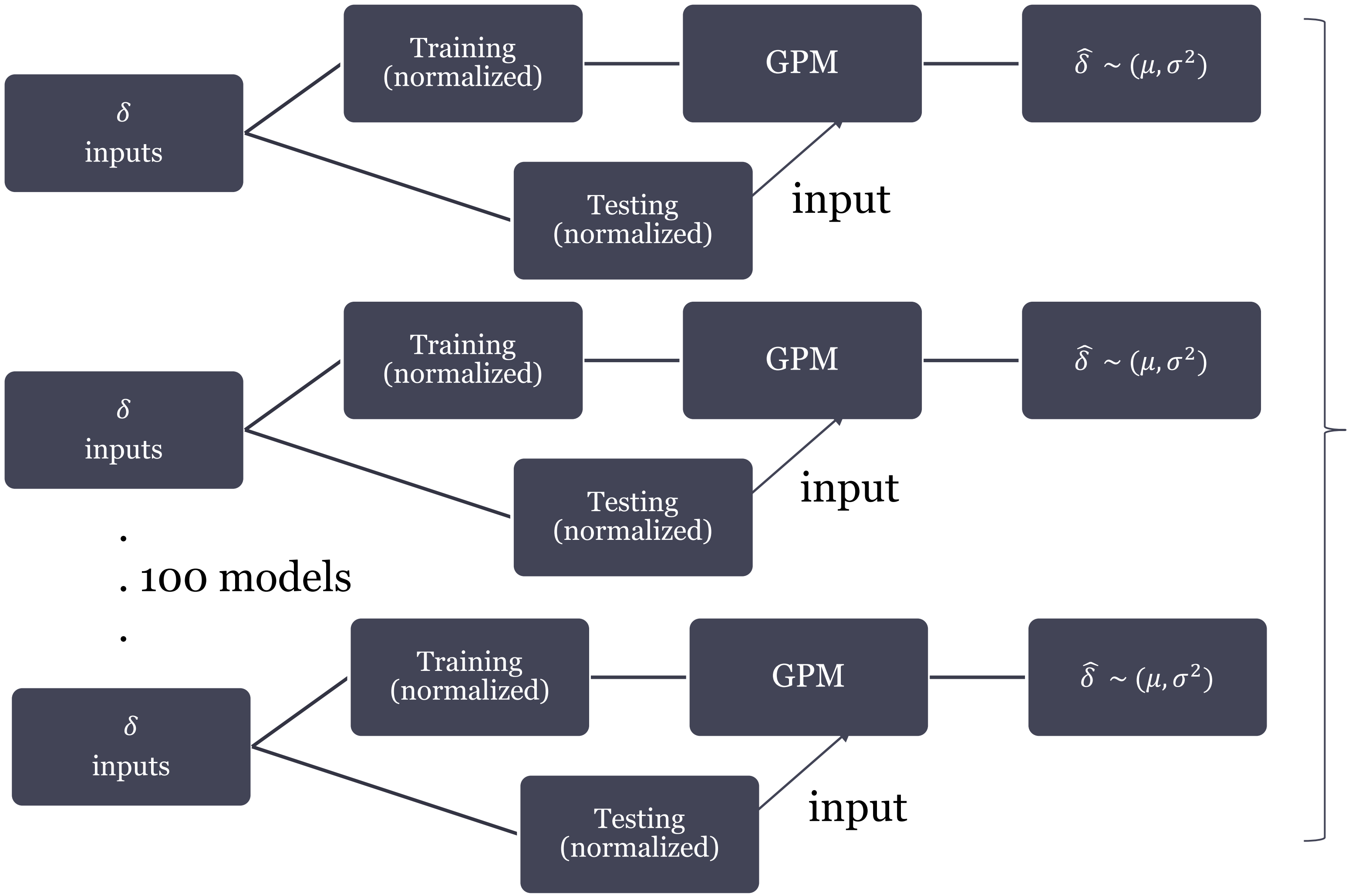
Database and models collected from literature

Model Name	Pipeline Orientation
Cionclini & Thome (2012)	Vertical
Cionclini (2010)	Vertical
Sawant et al. (2009)	Vertical
Sawant et al. (2008)	Vertical
Zhang et al. (2003)	Vertical
Pan & Hanratty (2002a)	Vertical
Utsuno & Kaminaga (1998)	Vertical
Nakazatomi & Sekguchi (1996)	Vertical
Ishii & Mishima (1989)	Vertical
Oliemans et al. (1986)	Vertical
Hughmark (1973)	Vertical
Wallis (1969)	Vertical
Pan & Hanratty (2002b)	Horizontal
Paleev & Filippovich (1966)	Horizontal
Wicks & Dukler (1960)	Horizontal
Mantilla (2008)	Horizontal
Ousaka et al. (1996)	Inclined
Bhagwat & Ghajar	Inclined



AUBURN UNIVERSITY

Gaussian process modeling (GPM) with bagging



Take average of μ, σ^2 to get $\hat{\delta}$ and 80% CI

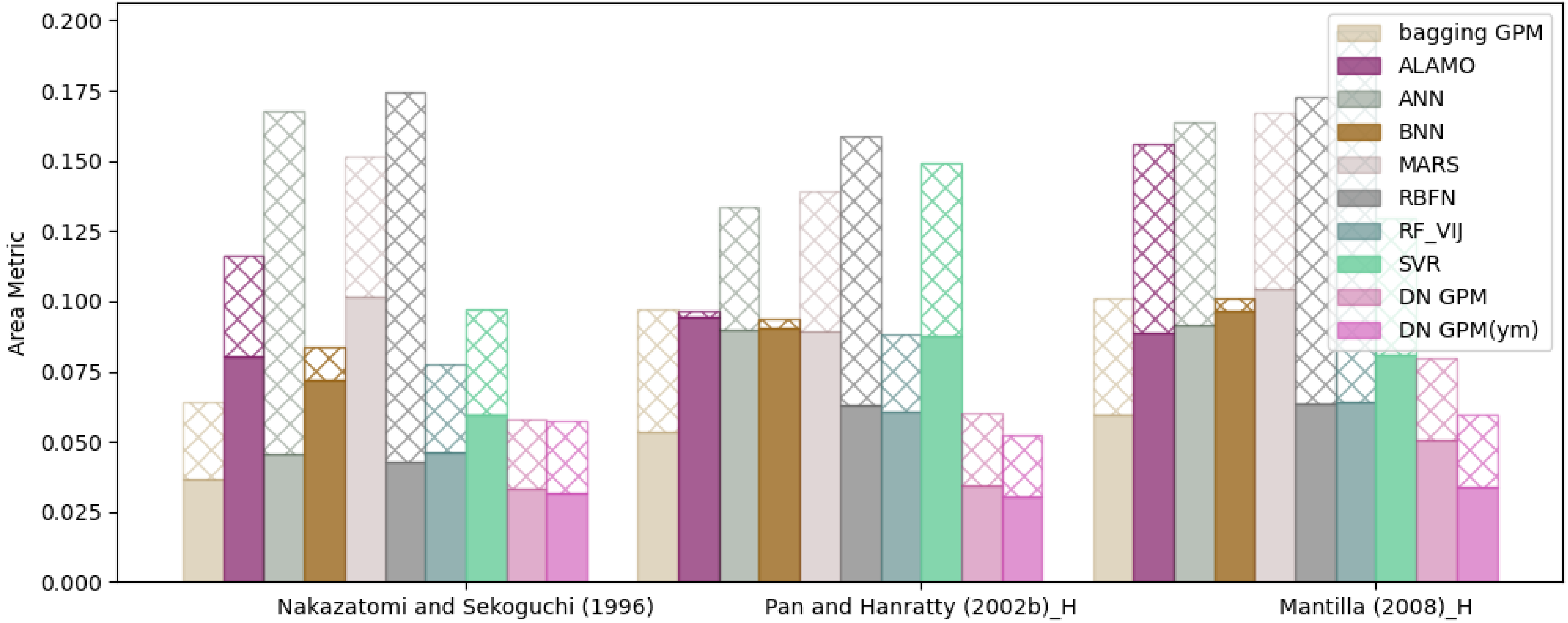
$$\bar{\mu} = \frac{1}{K}\sum_{k=1}^{K} \mu(k)$$

$$\overline{\sigma^2} = \frac{1}{K}\sum_{k=1}^{K} \sigma^2(k) + \frac{1}{K}\sum_{k=1}^{K} (\mu(k) - \bar{\mu})^2$$

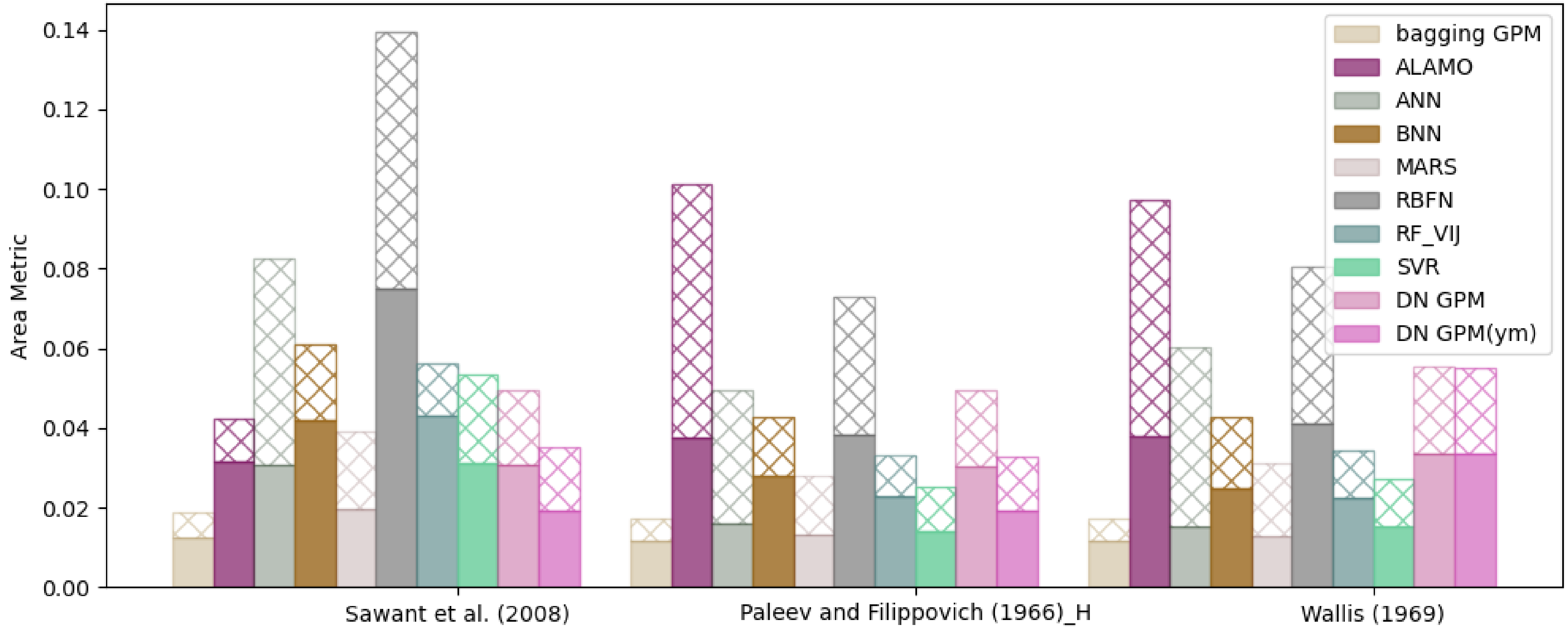
K is the number of models[1]

1 Chen, T.; Ren, J. Bagging for Gaussian process regression. *Neurocomputing* **2009**, *72*, 1605–1610.

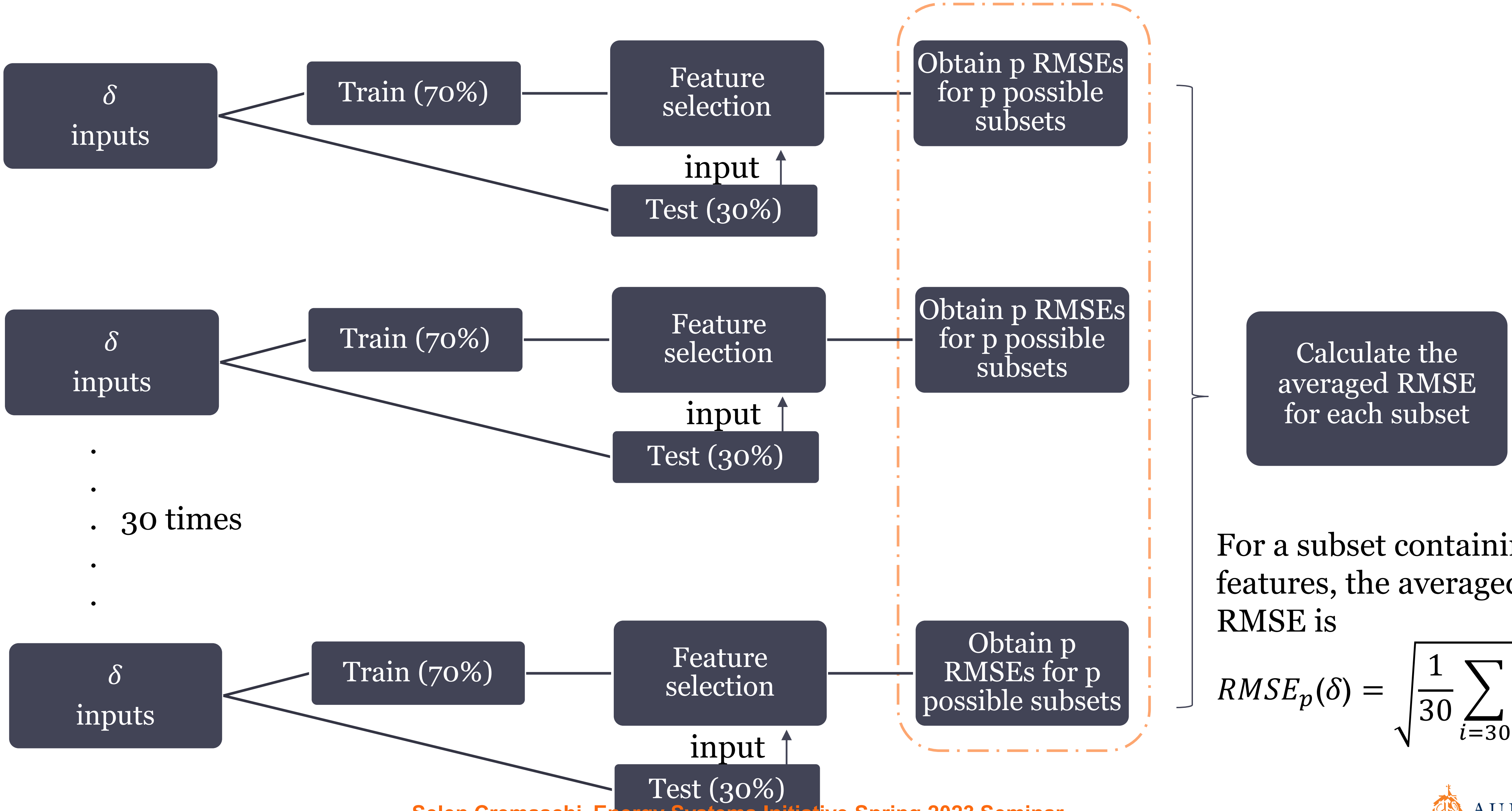
Horizontal flow



Inclined flow models



Validation experiment flowchart



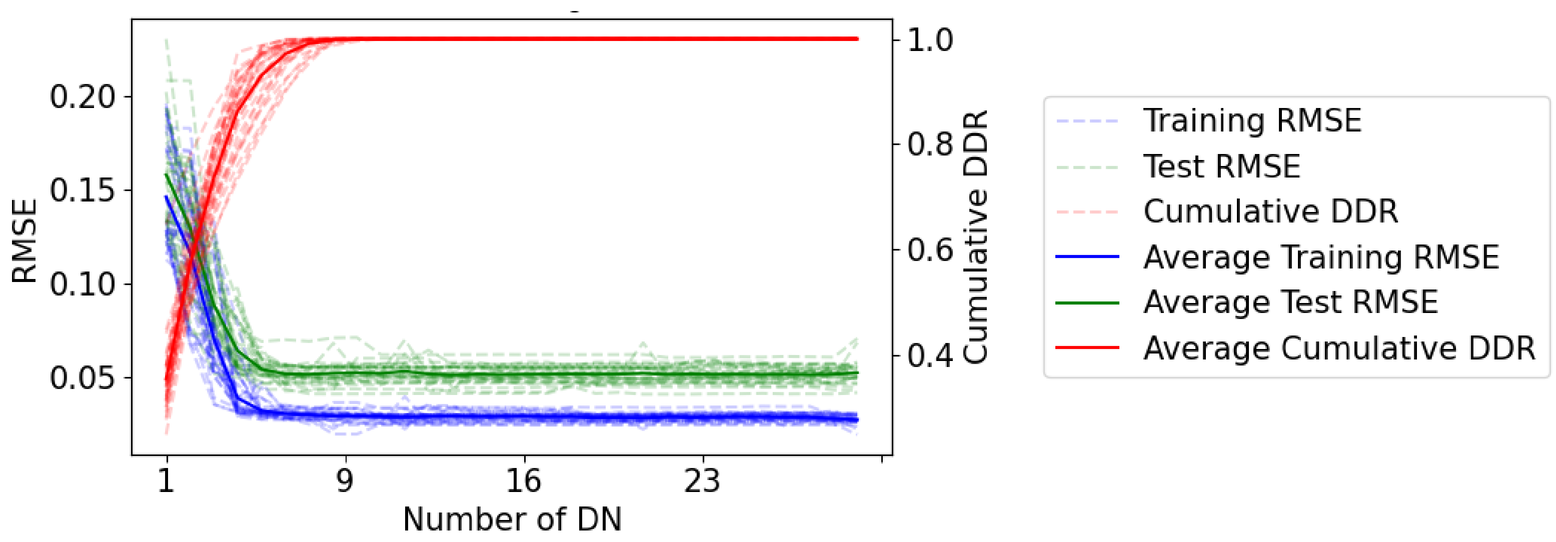
For a subset containing p features, the averaged RMSE is

$$RMSE_p(\delta) = \sqrt{\frac{1}{30}\sum_{i=30} RMSE_i}$$

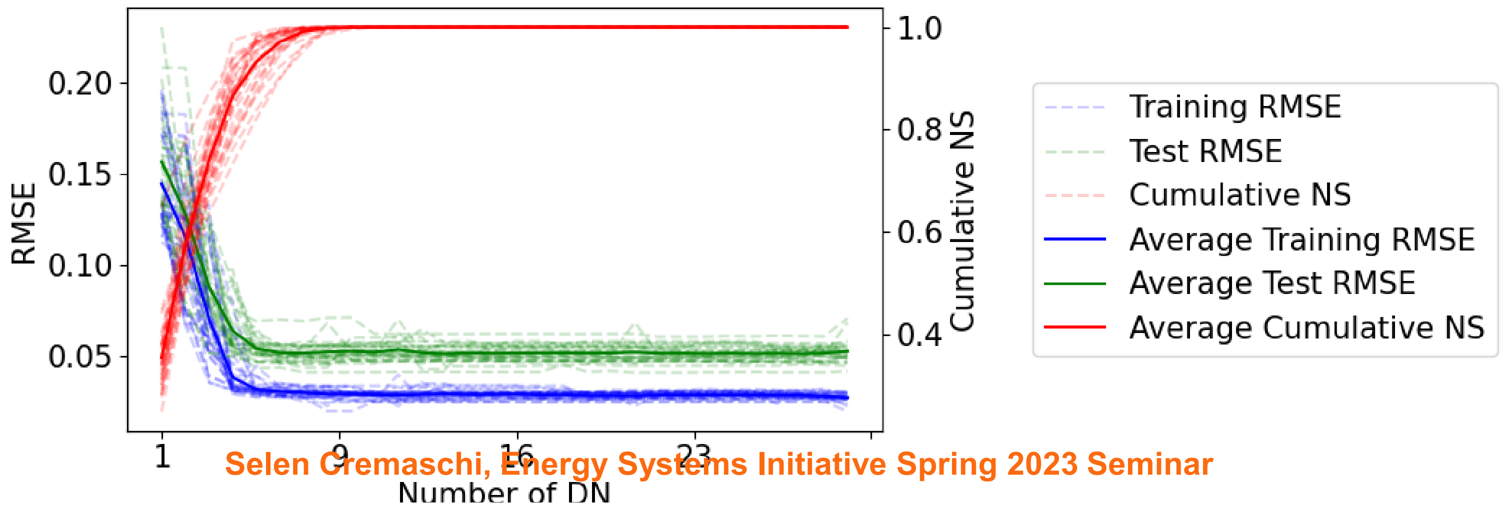
AUBURN UNIVERSITY

Validation experiments results – horizontal flow

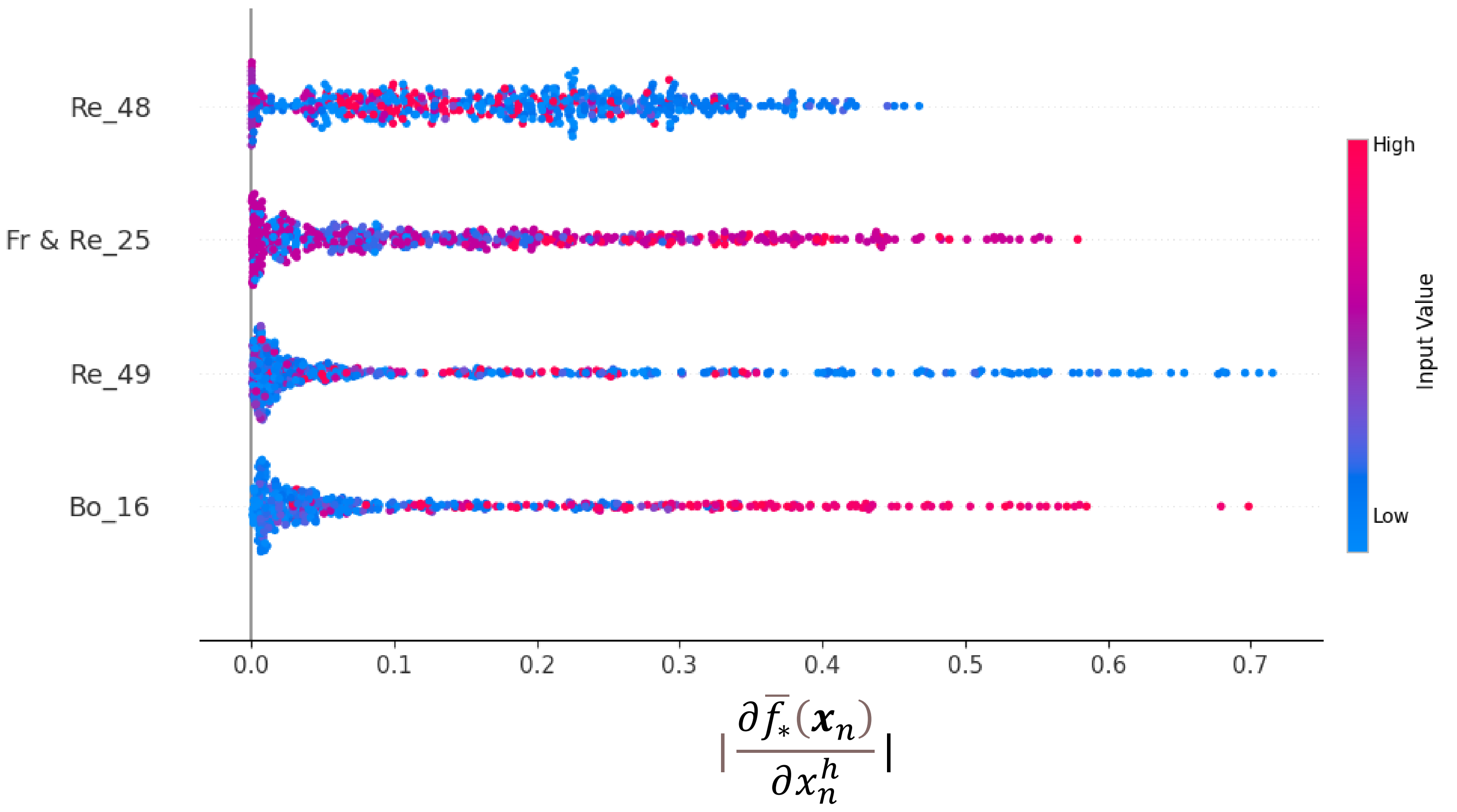
Cumulative Derivative Decomposition Ratio (DDR)



Cumulative Normalized Sensitivity (NS)



Local feature importance heatmap – horizontal flow



DNs with physical explanation	
Re_48	$\frac{d\rho_L V_{SL}}{\mu_L}$
Bo_16	$\frac{g d^2 \rho_L}{\sigma}$

DNs without physical explanation	
Fr & Re_25	$\frac{\mu_G g}{\rho_L {V_{SL}}^3}$
Re_49	$\frac{d\rho_L V_{SG}}{\mu_G}$

AUBURN UNIVERSITY