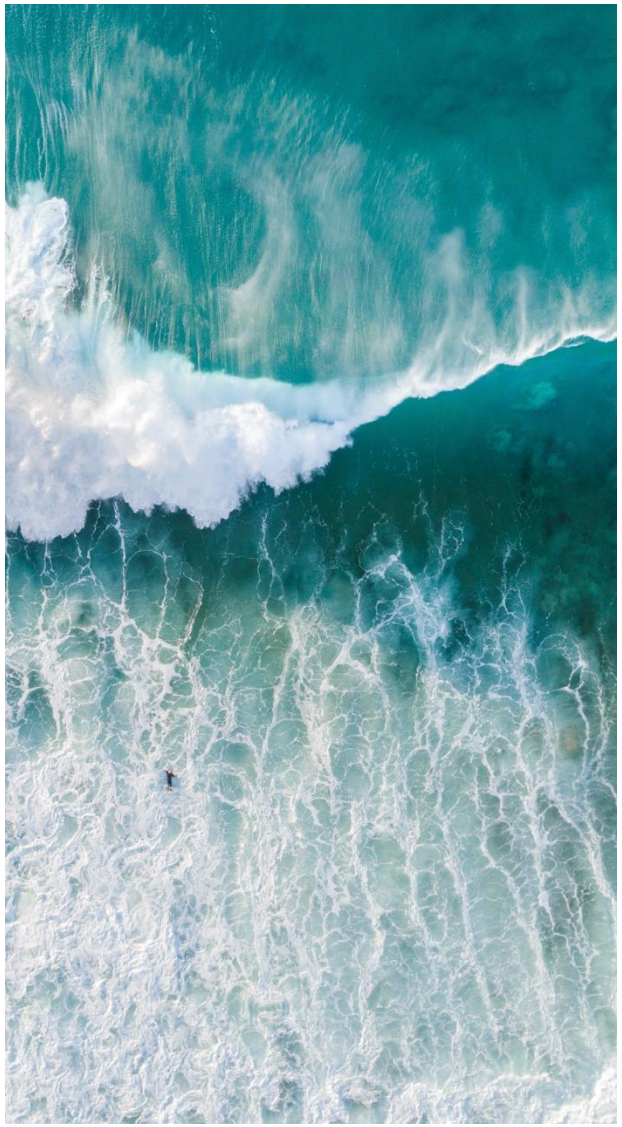
Evaluating Storm Selection for Machine Learning-Based Tropical Cyclone Surge Surrogate Models

Jen Irish, Shelby Bensi, Yang Shao, Meredith Carr, Constantinos Frantzis

September 2025



Sponsored in part by the U.S. Coastal Research Program, U.S. Army Corps of Engineers.

Motivation

Key Findings

Methods

Results

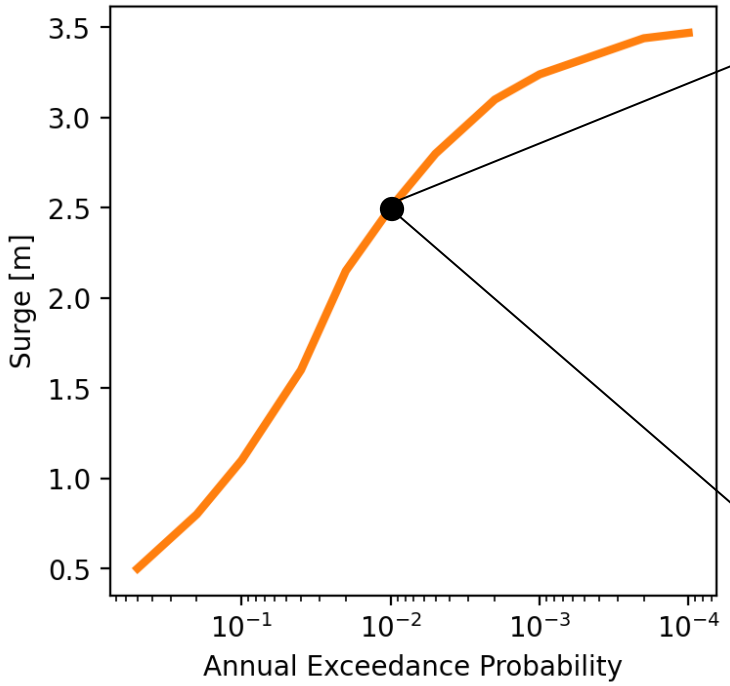
Conclusions

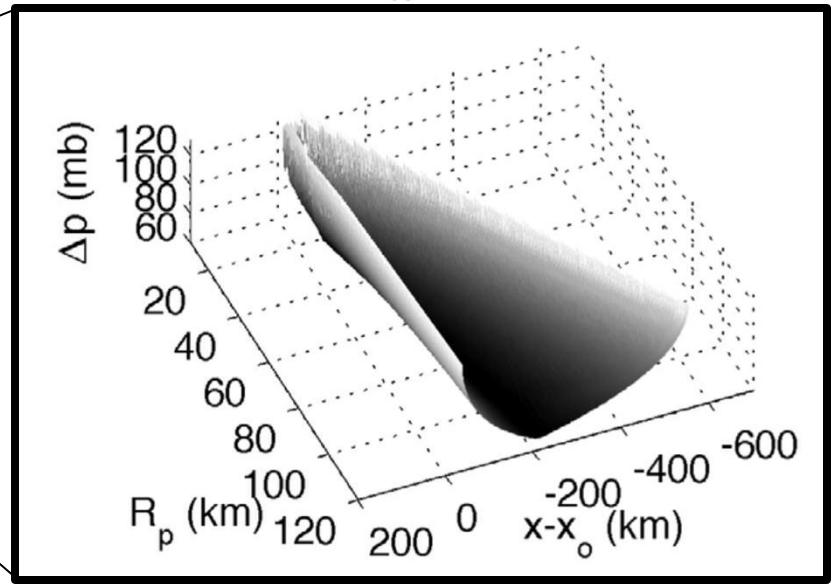
- Essential for engineering, planning and design
- High-fidelity, surge models computationally expensive
- Emerging popularity of surge surrogate models

Note:

High-fidelity surge model set = ***Machine learning training sets***

Tropical Cyclone Surge Hazard Assessment





Surge Surrogate Models

Millions of storm surge estimates

Top right figure from Irish et al. 2011, JGR.

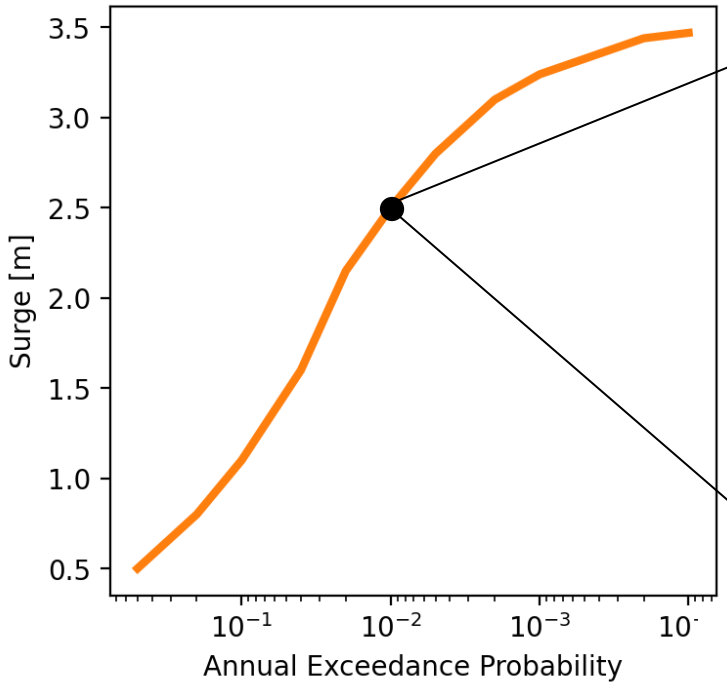
Tropical Cyclone Surge Hazard Assessment

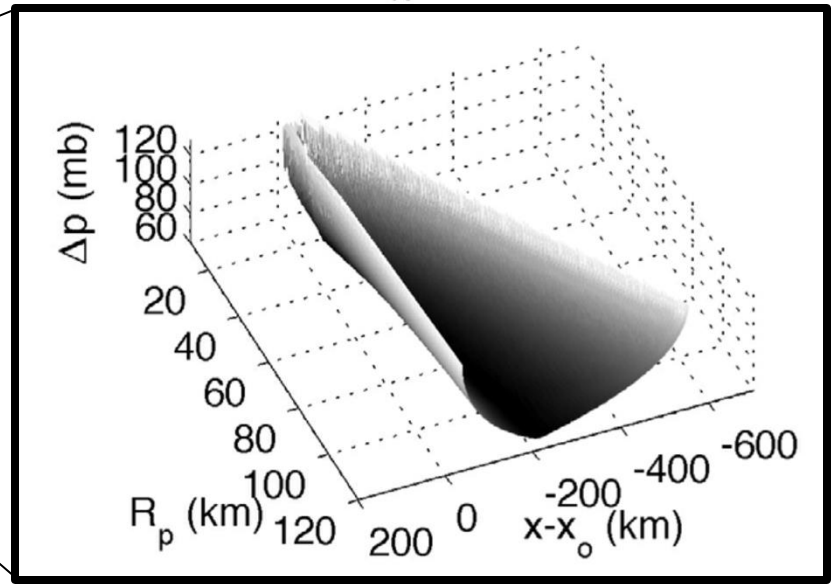
Motivation
Key Findings
Methods
Results
Conclusions

- Essential for engineering, planning and design
- High-fidelity, surge models computationally expensive
- Emerging popularity of surge surrogate models

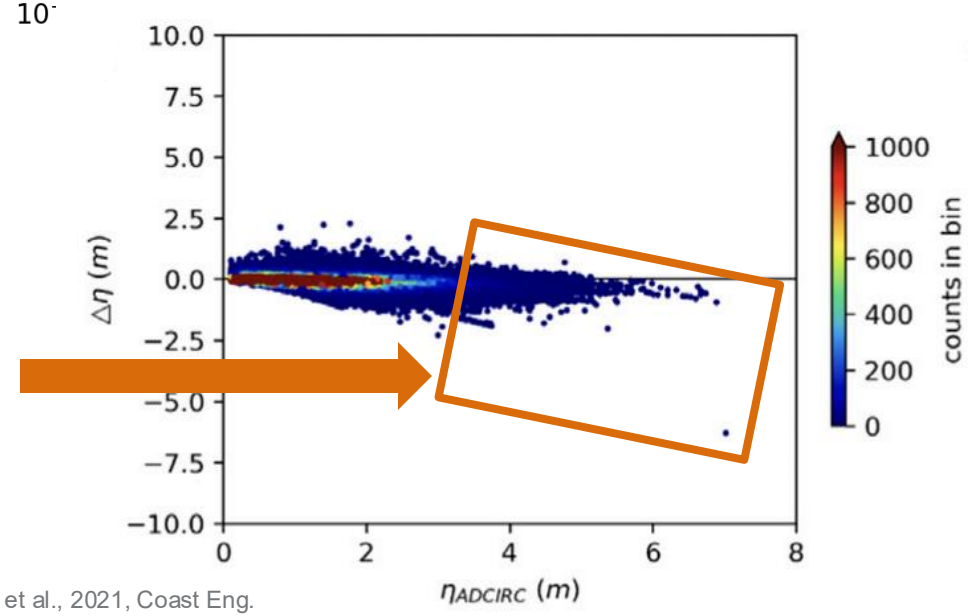
Note:

High-fidelity surge model set = ***Machine learning training sets***

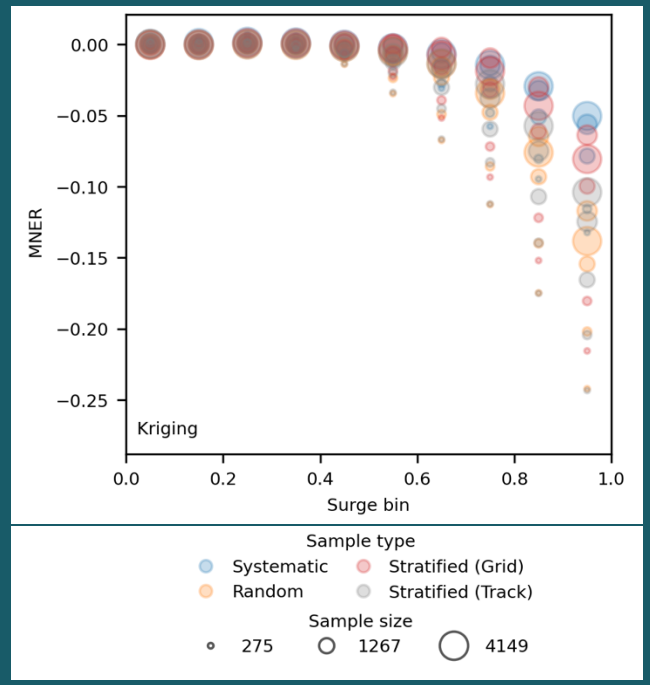




≤ 1% annual exceedance probability



Top right figure from Irish et al. 2011, JGR. Lower figure from Lee* et al., 2021, Coast Eng.



Preliminary Findings

- **Systematic** and **stratified (grid)** sets outperform **random** and **stratified (track)** sets

- **Negative** bias at extremes, regardless of set size / type or model type
  - **Extremes underestimated**
  - Bias reduced with larger set size
  - Bias reduced with **systematic** and **stratified (grid)** sets

- Aggregate error statistics overstate performance

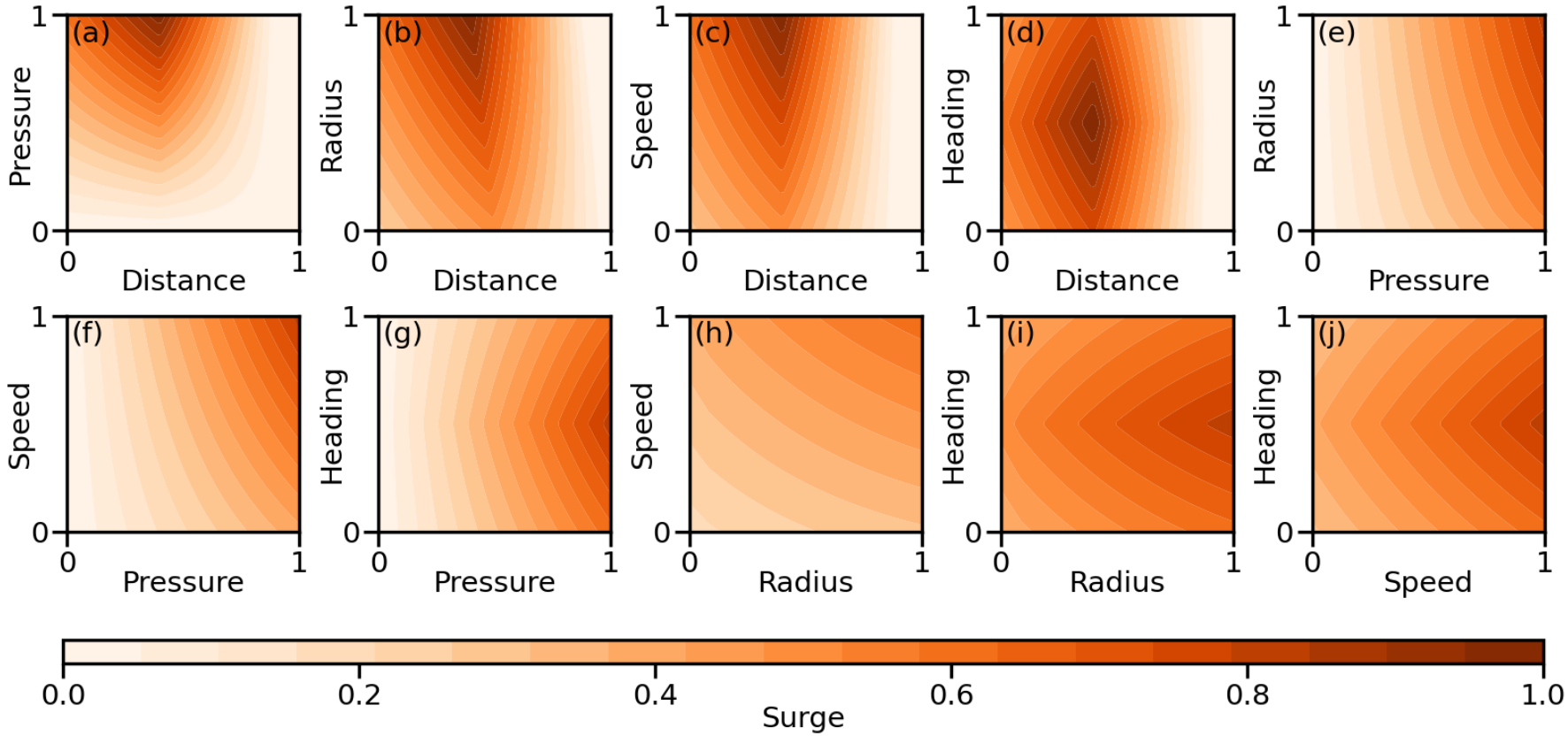
- At extremes, Kriging using larger **systematic** or **stratified (grid)** sets outperform ANN

- For moderate surges, ANN outperforms Kriging

Surge Response Surface Basis

- Single geographic location
- Dimensionless "unit" surge
- Follows Irish et al. (2008, *J Phys Ocean* & 2010, *Ocean Eng*)
- Five-dimensional "unit" track parameters:
  - Distance (landfall location)
  - Pressure (pressure deficit)
  - Radius
  - Heading
  - Speed (forward speed)

"True" Oceanic Surge from Analytical Model

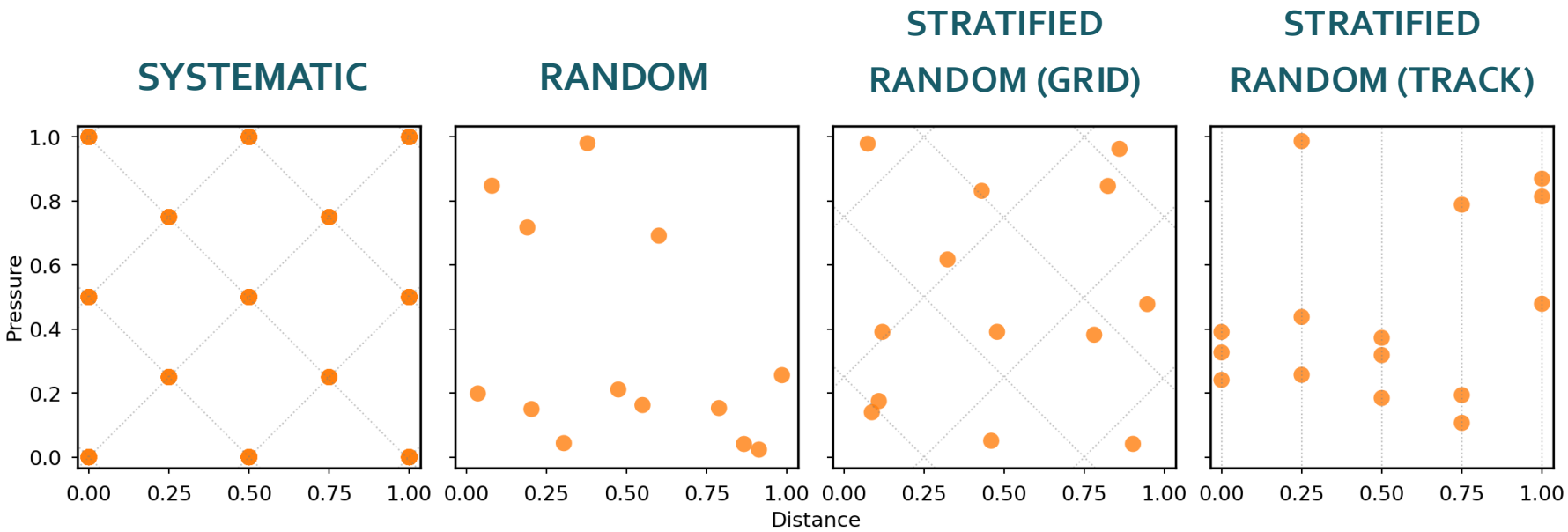


- Triangular on Distance (skewed), along-coast peak offset by Radius
- Triangular on Heading (symmetric)
- Linear on all other parameters

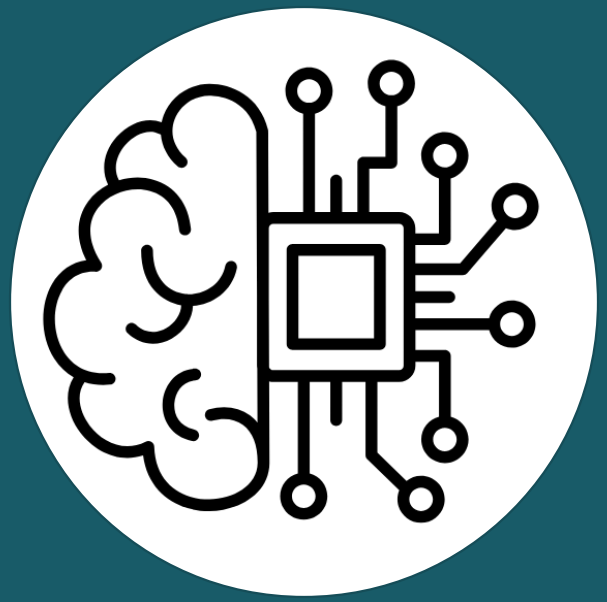
"Unit" Storm Track Parameters

- Distance (landfall location)
- Pressure (pressure deficit)
- Radius
- Heading
- Speed (forward speed)

Training Sets: Storm Sample Types & Sample Sizes



- Sample size constrained by systematic, triangular grid:
  - 275 unique storms
  - 486 unique storms
  - 1267 unique storms
  - 2048 unique storms
  - 4149 unique storms



Created by Nanik haq
from Noun Project

Surrogate Model Types

MULTILINEAR INTERPOLATION

- Five-dimensional Delaunay triangulation
- Python built-in function
- No extrapolation

KRIGING (GAUSSIAN PROCESS REGRESSION)

- Hyperparameter tuning using 15 randomized trials (nugget/noise term, radial-basis function length scale)
- 3-fold cross-validation using 80 / 20 train / validation split

ARTIFICIAL NEURAL NETWORK (ANN)

- Hyperparameter tuning using 15 randomized trials (hidden layers, units per layer, learning rate)
- 3-fold cross-validation using 80 / 20 train / validation split

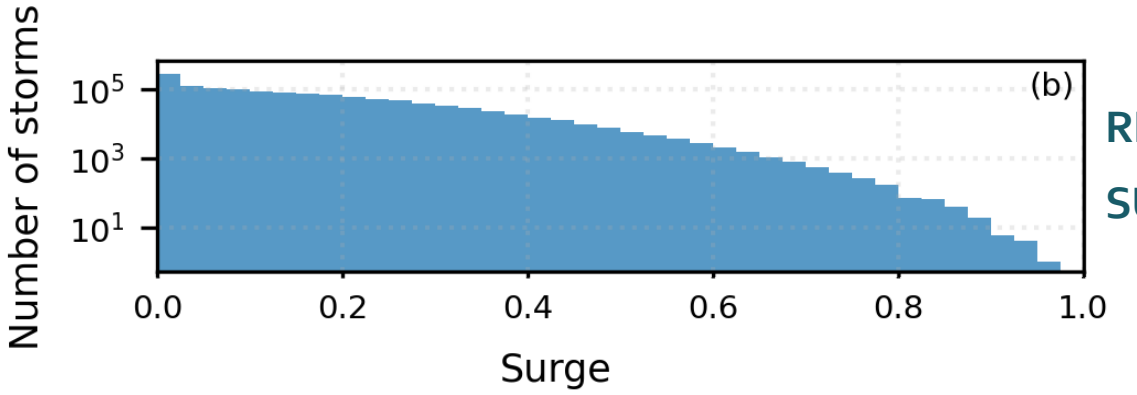
BASIS

- 1.3-million storm systematic triangular set
- Error = Surrogate - Analytical

Error Characterization

STATISTICS

- Root-mean-square error (RMSE)
- Mean error (MNER)
- Standard deviation of errors (STDE)
- Quartiles: Q1, Q2 (median), Q3
  - Interquartile range: IQR = Q3-Q1
- Considered:
  - **Aggregate** for entire 1.2-million storm set
  - **Binned by surge magnitude** (0.1-unit intervals)

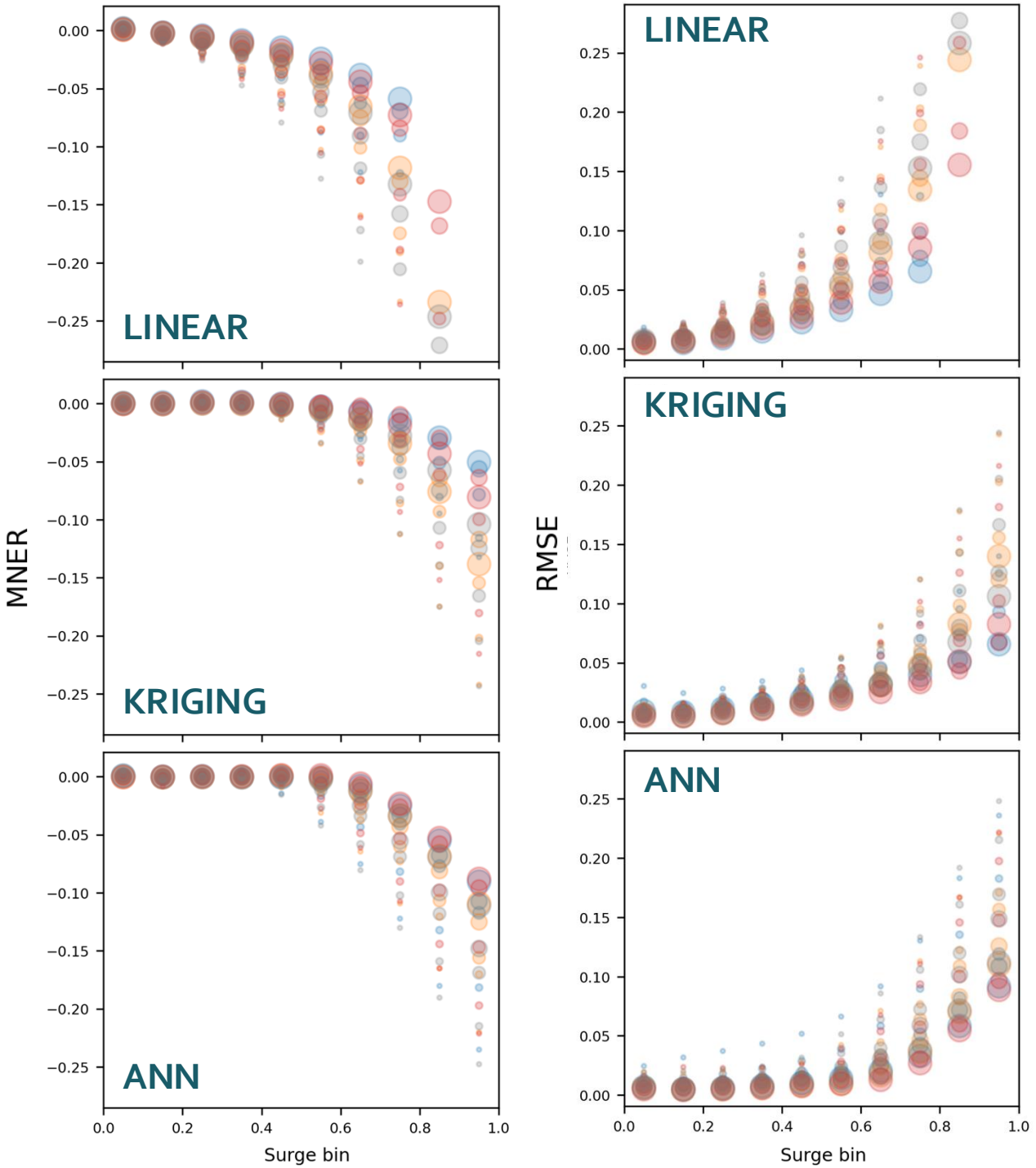


RESPONSE SURFACE 2

Aggregate Error Statistics for all Models and all Samples

- $-0.016 \leq MNER \leq 0.001$
- $0.005 \leq RMSE \leq 0.038$

Ensemble-Averaged Mean & Root-Mean Square Error



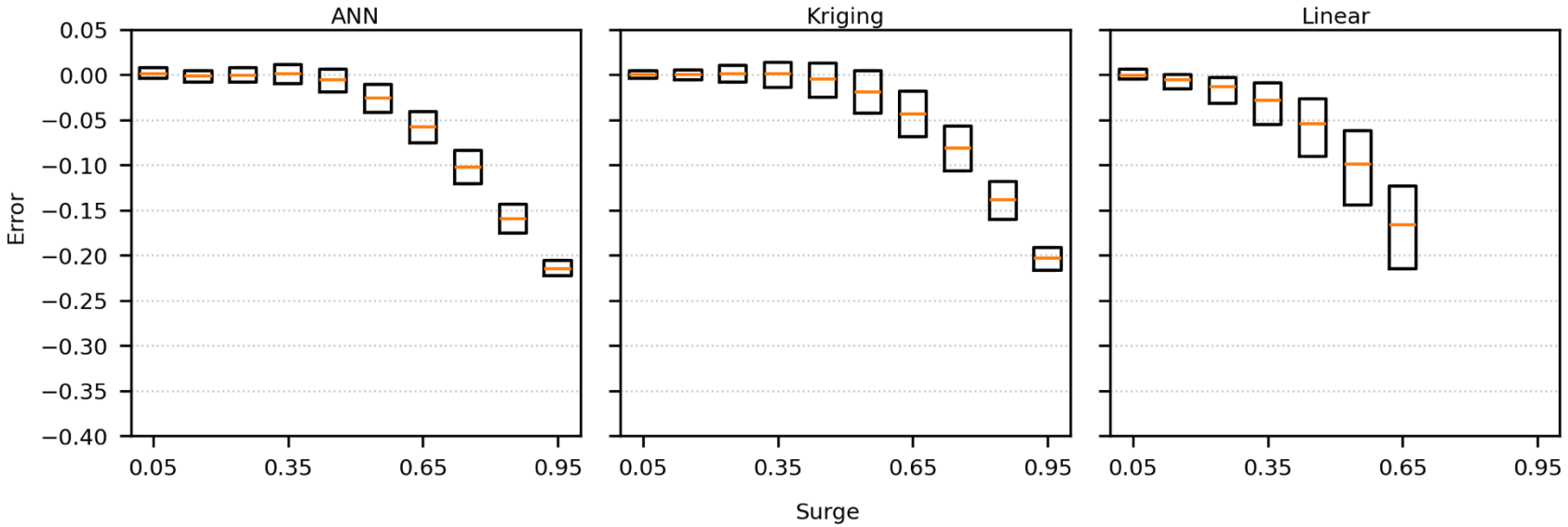
VIRGINIA TECH

Aggregate Error Statistics for all Models and all Samples

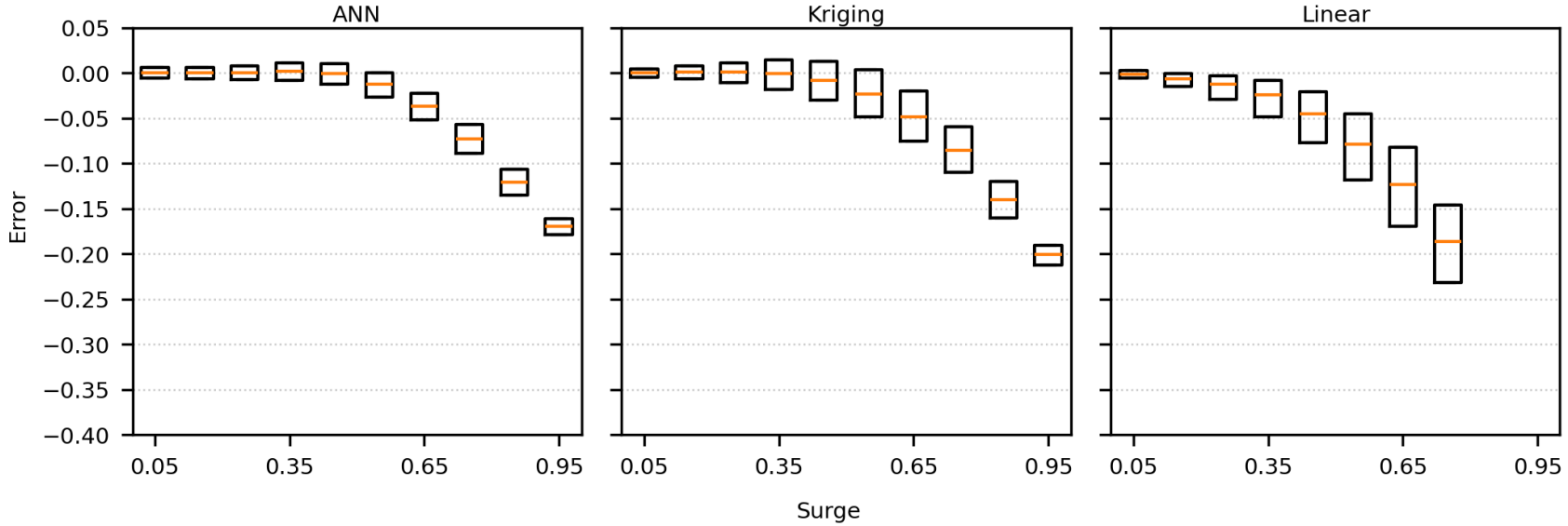
- $-0.016 \leq MNER \leq 0.001$
- $0.005 \leq RMSE \leq 0.038$

Box Plots – Ensemble-Averaged Median & Interquartile Range

STRATIFIED (TRACK) - 486



RANDOM - 486



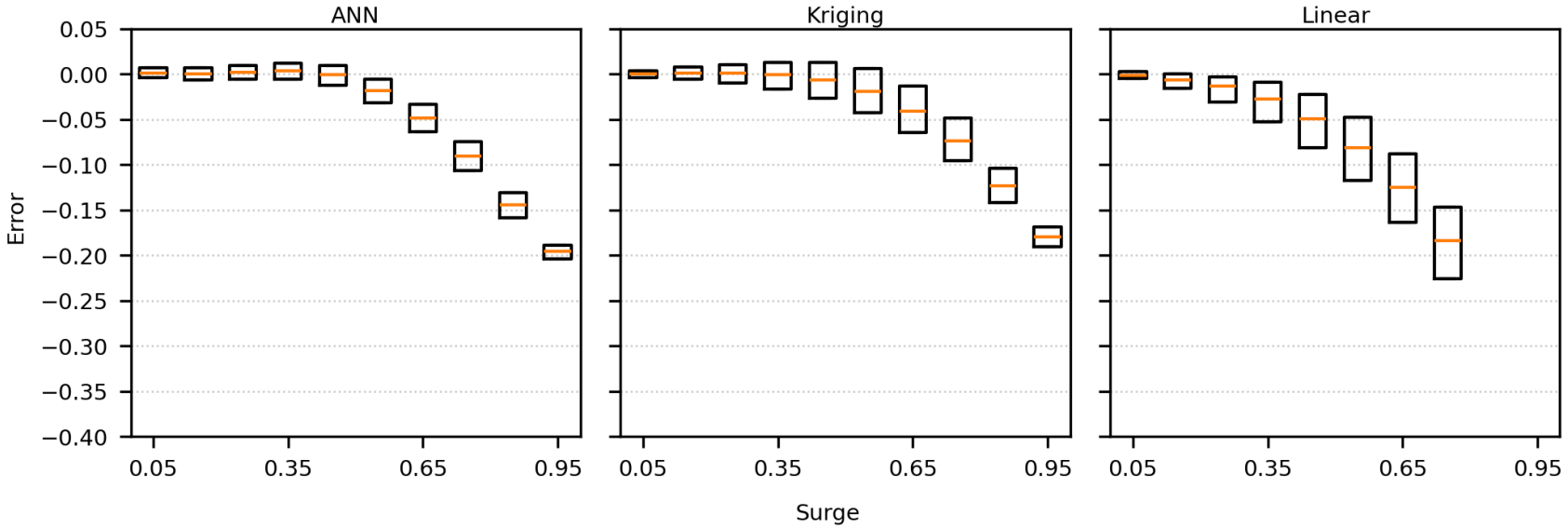
VIRGINIA TECH

Aggregate Error Statistics for all Models and all Samples

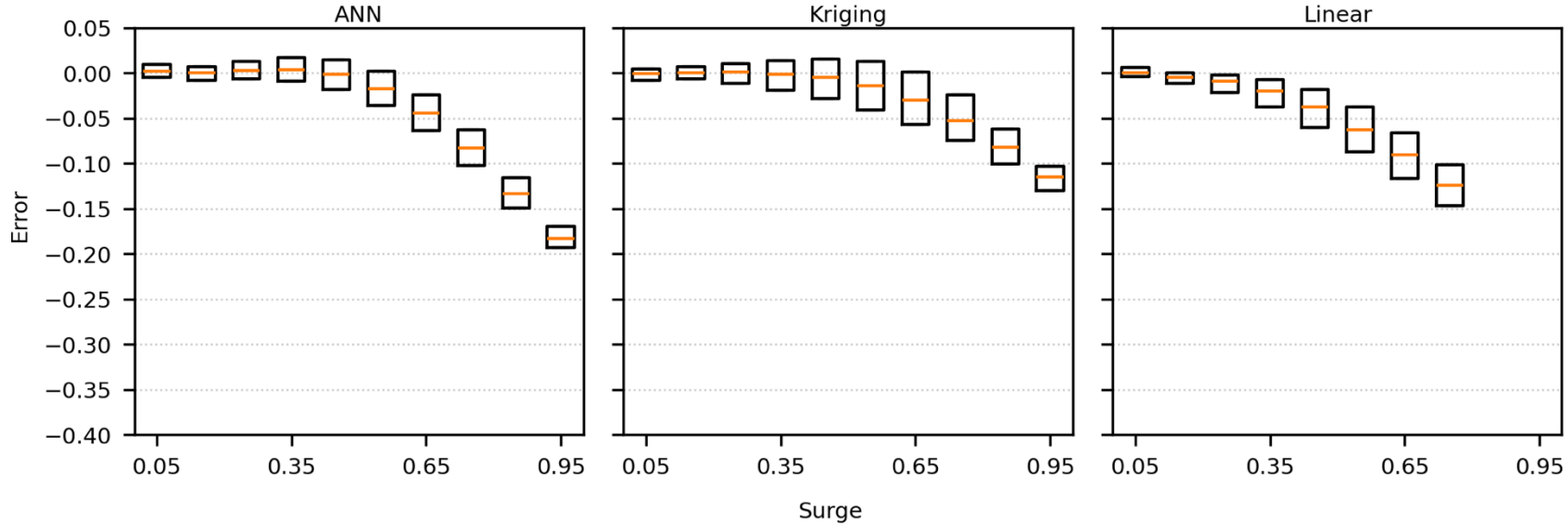
- $-0.016 \leq MNER \leq 0.001$
- $0.005 \leq RMSE \leq 0.038$

Box Plots – Ensemble-Averaged Median & Interquartile Range

STRATIFIED (GRID) - 486



SYSTEMATIC - 486

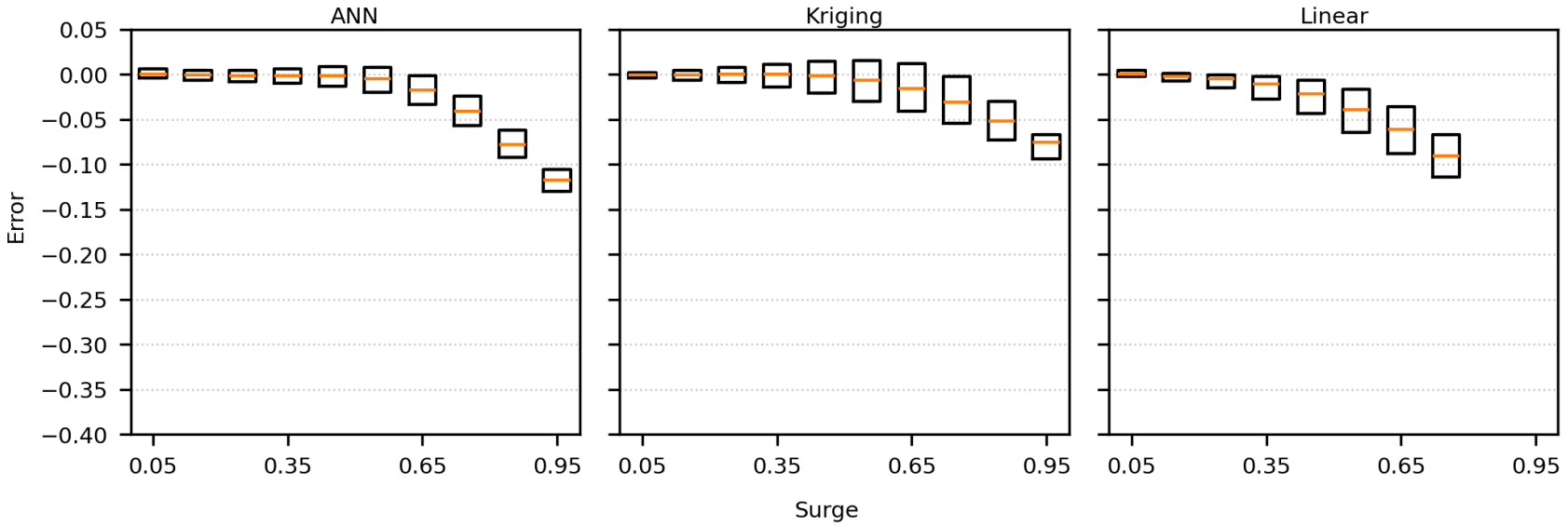


Aggregate Error Statistics for all Models and all Samples

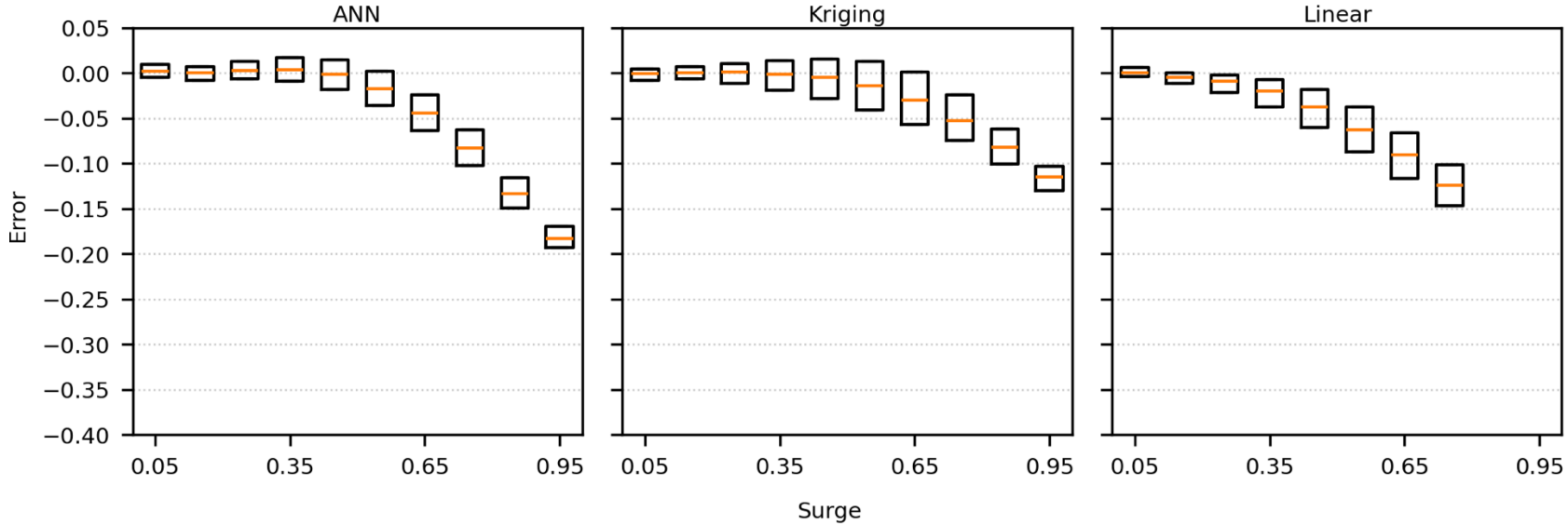
- $-0.016 \leq MNER \leq 0.001$
- $0.005 \leq RMSE \leq 0.038$

Box Plots – Ensemble-Averaged Median & Interquartile Range

SYSTEMATIC - 1267



SYSTEMATIC - 486



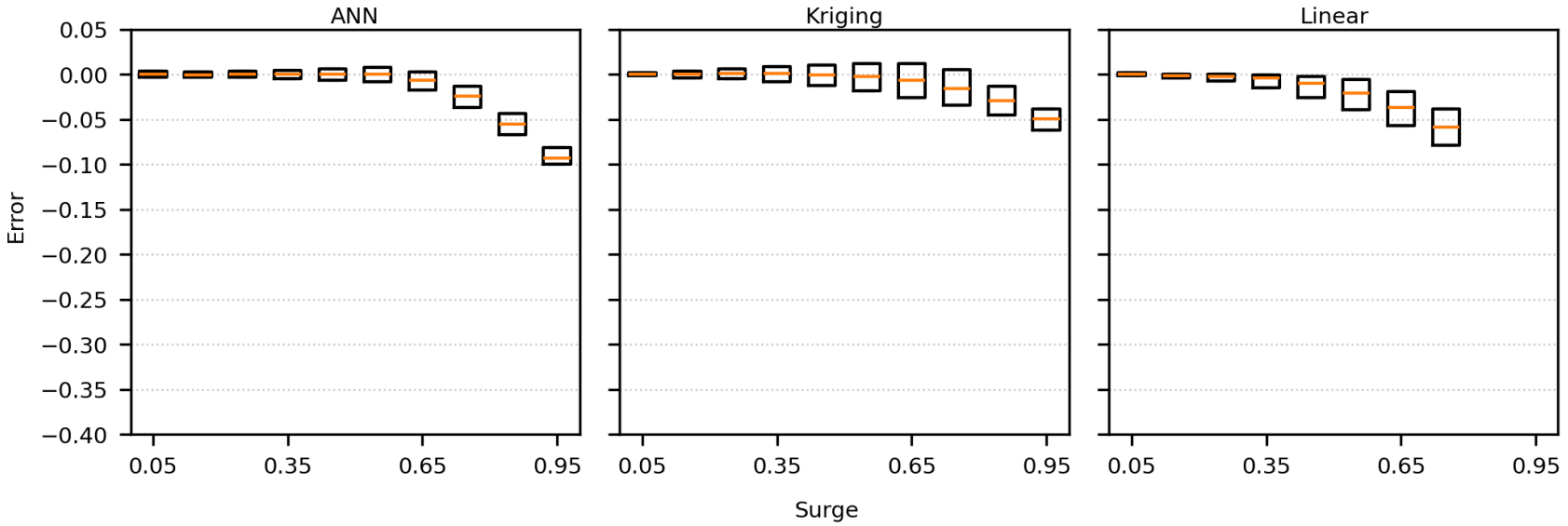
VIRGINIA TECH

Aggregate Error Statistics for all Models and all Samples

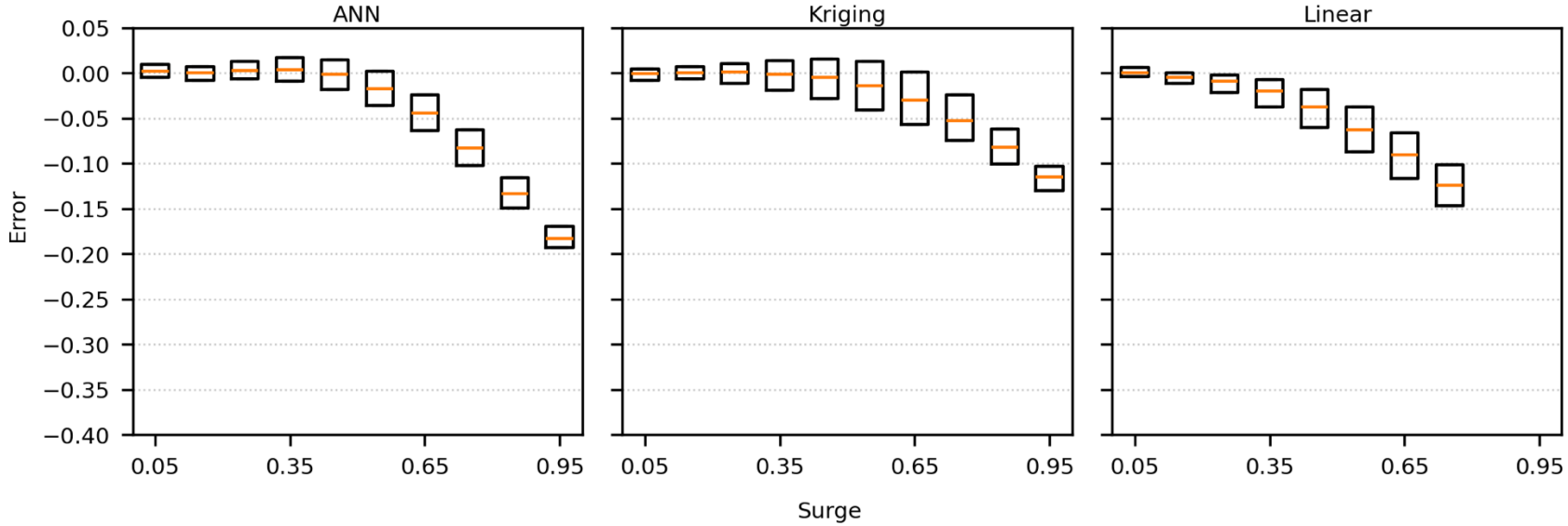
- $-0.016 \leq MNER \leq 0.001$
- $0.005 \leq RMSE \leq 0.038$

Box Plots – Ensemble-Averaged Median & Interquartile Range

SYSTEMATIC - 4149



SYSTEMATIC - 486



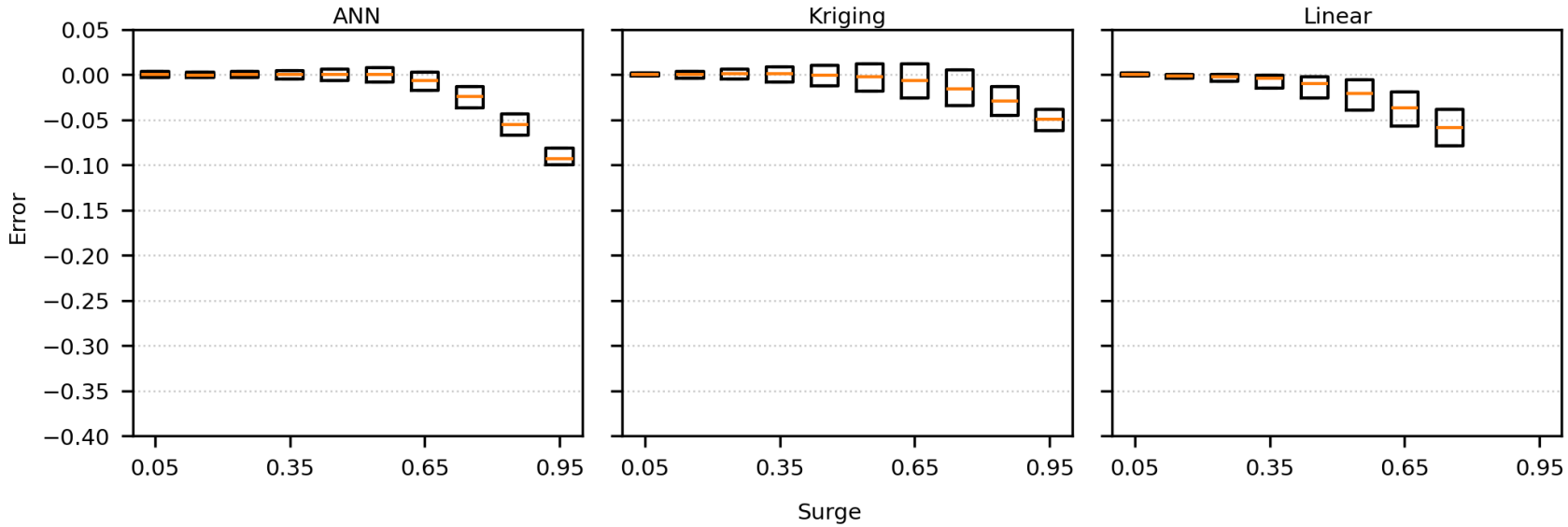
VIRGINIA TECH

Aggregate Error Statistics for all Models and all Samples

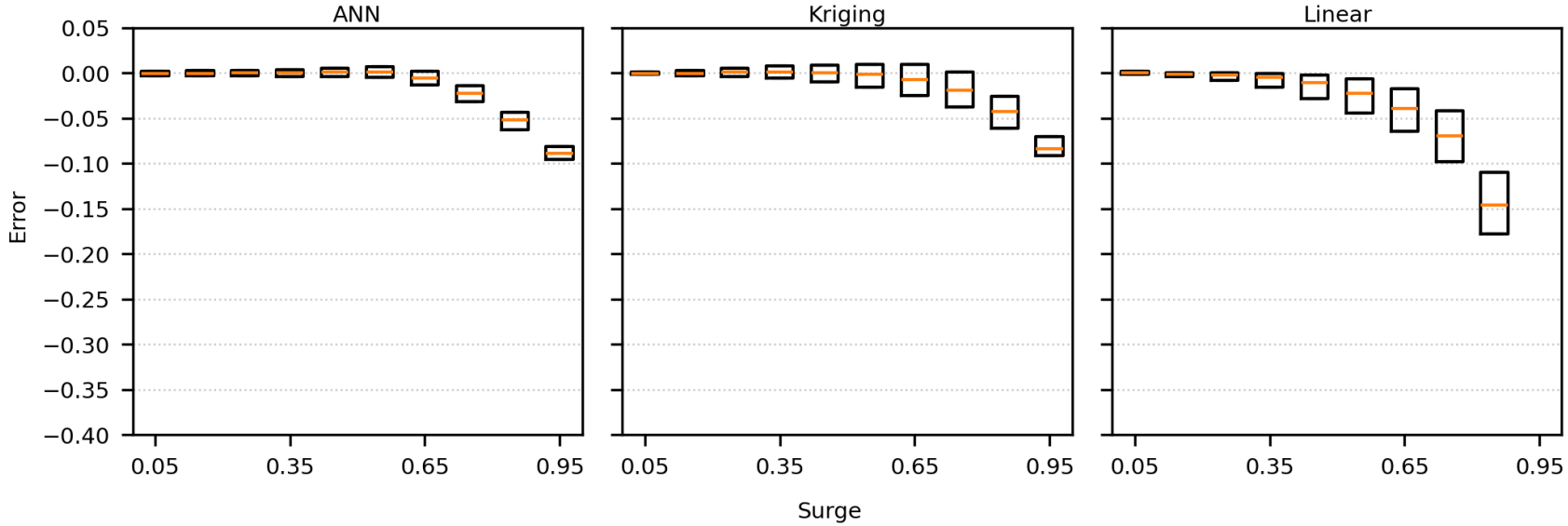
- $-0.016 \leq MNER \leq 0.001$
- $0.005 \leq RMSE \leq 0.038$

Box Plots – Ensemble-Averaged Median & Interquartile Range

SYSTEMATIC - 4149

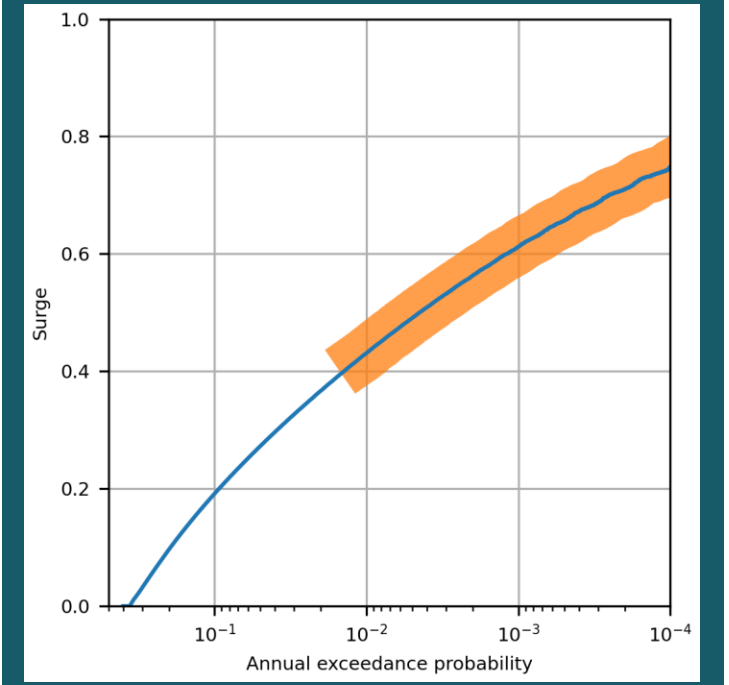


STRATIFIED (GRID) - 4149



VIRGINIA TECH

Illustration of potential impact



Conclusions - Preliminary Findings

- **Systematic** and **stratified (grid)** sets outperform **random** and **stratified (track)** sets

- **Negative** bias at extremes, regardless of set size / type or model type
  - **Extremes underestimated**
  - Bias reduced with larger set size
  - Bias reduced with **systematic** and **stratified (grid)** sets

- Aggregate error statistics overstate performance

- At extremes, Kriging using larger **systematic** or **stratified (grid)** sets outperform ANN

- For moderate surges, ANN outperforms Kriging

Questions?

Coastal Engineering
at Virginia Tech

NSF's NHERI RAPID Facility
supporting post-disaster reconnaissance


U.S. Army Corps of Engineers


Courtesy ASCE


U.S. Army Corps of Engineers

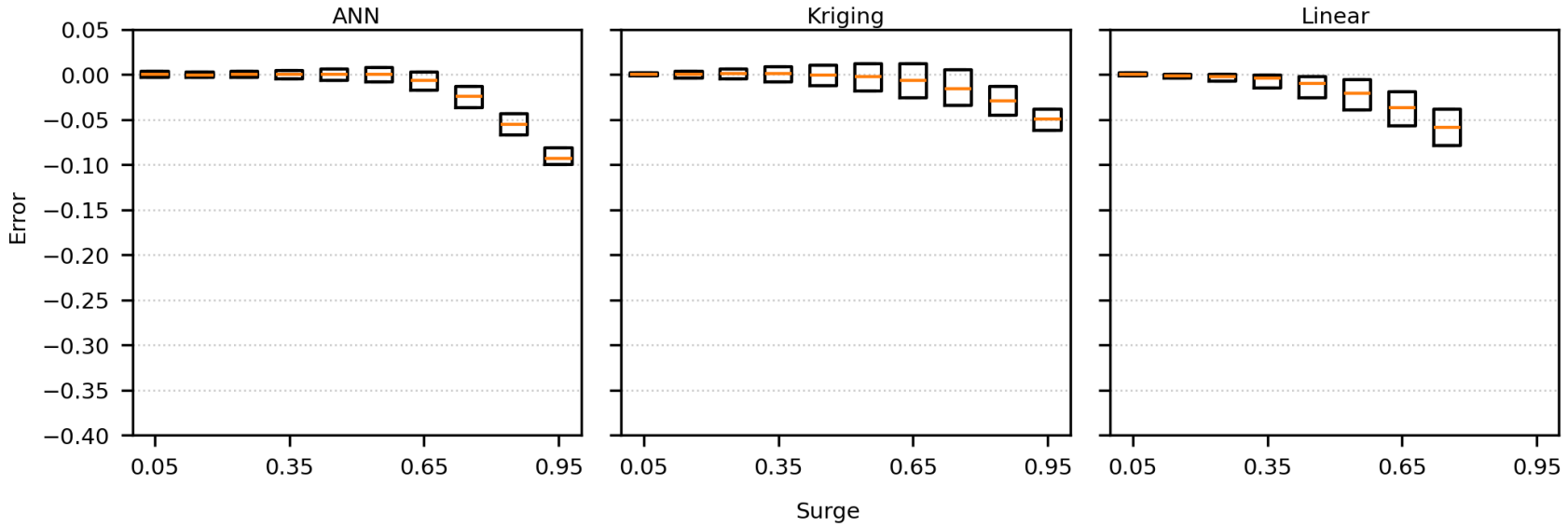
This material is based upon work supported by the U.S. Coastal Research Program, U.S. Army Corps of Engineers.. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of these organizations.

Aggregate Error Statistics for all Models and all Samples

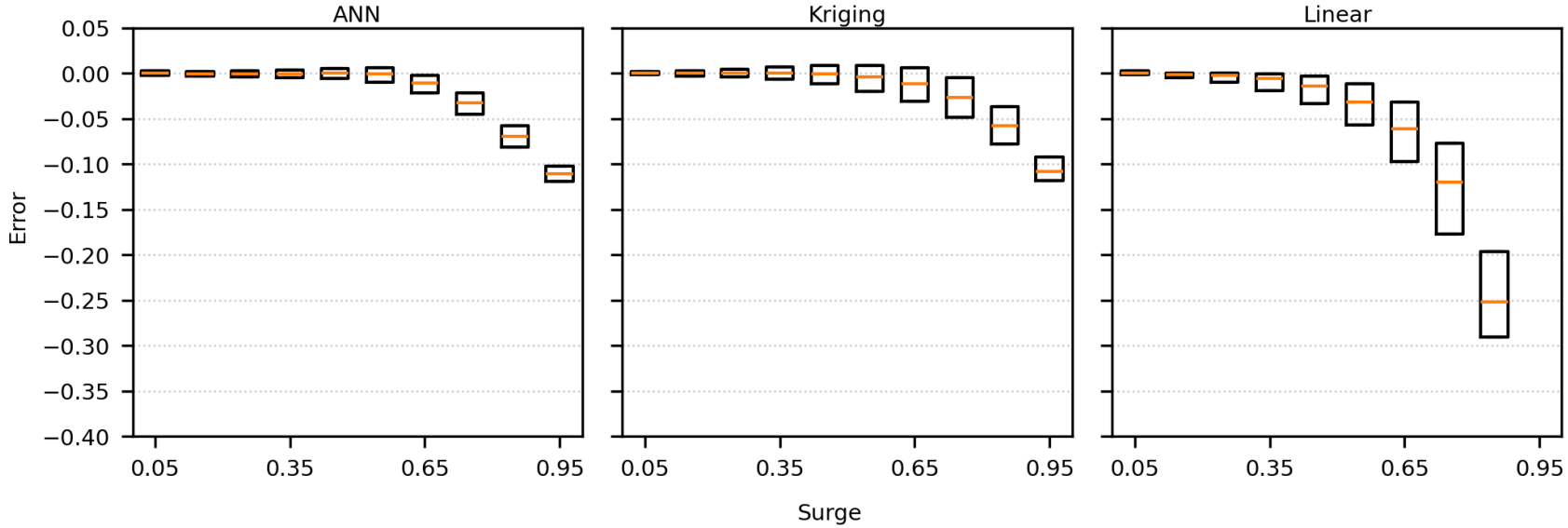
- $-0.016 \leq MNER \leq 0.001$
- $0.005 \leq RMSE \leq 0.038$

Response Surface 2

SYSTEMATIC - 4149



STRATIFIED (TRACK) - 4149



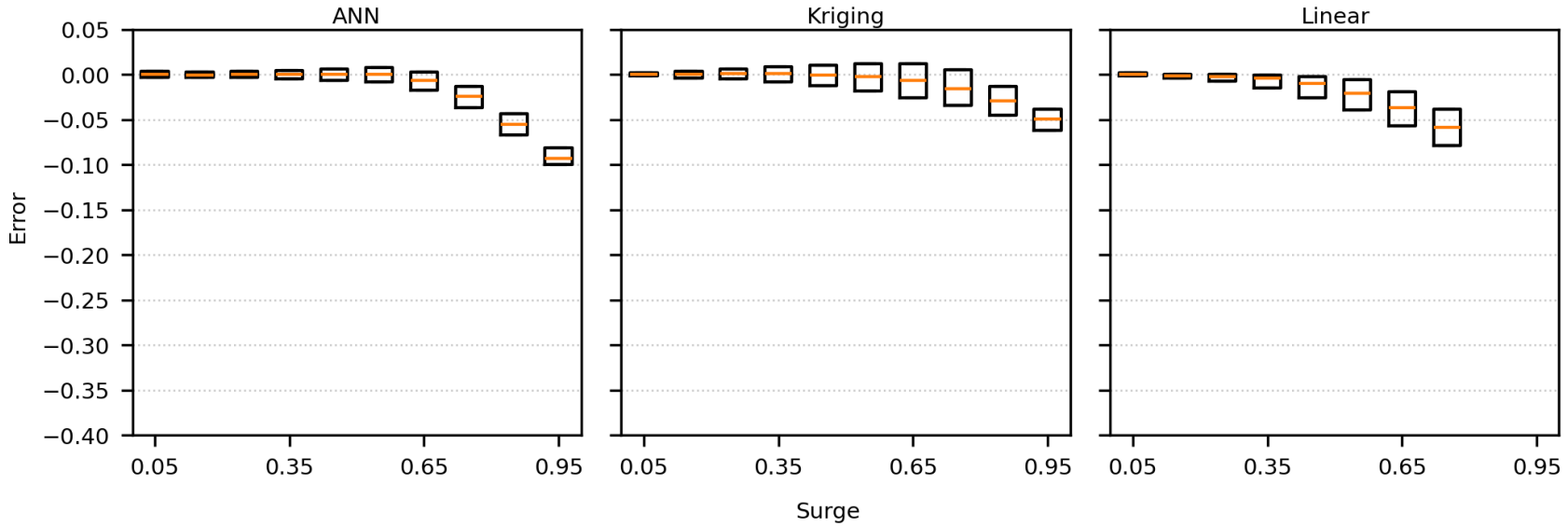
VIRGINIA TECH

Aggregate Error Statistics for all Models and all Samples

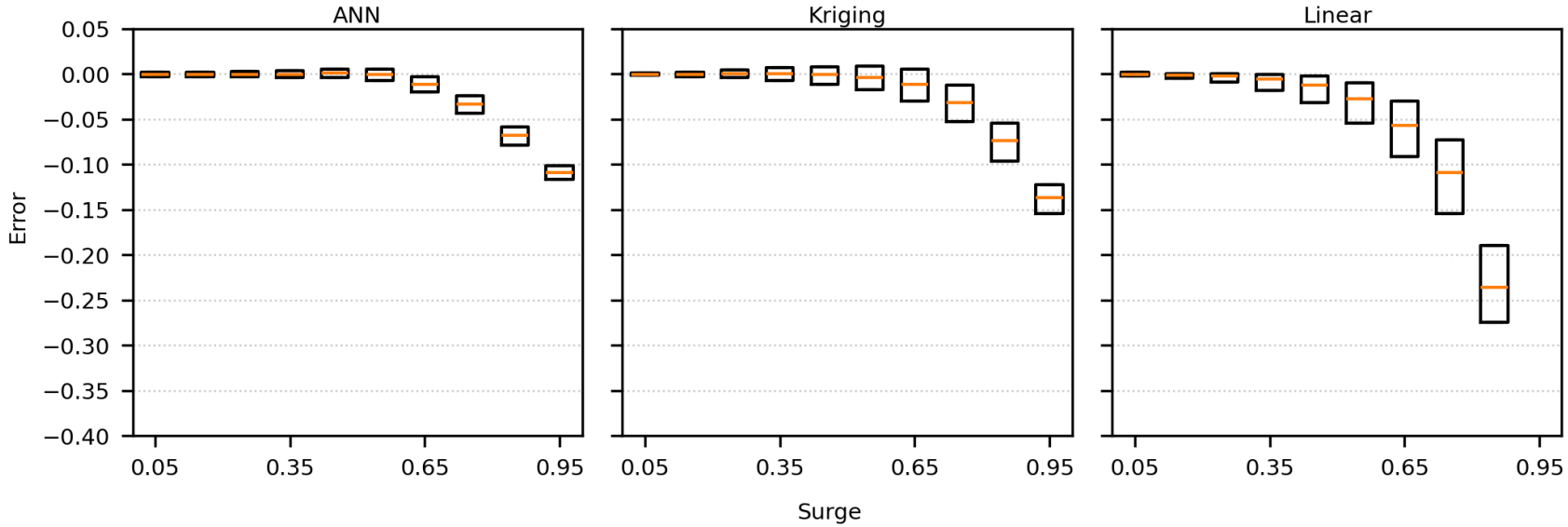
- $-0.016 \leq MNER \leq 0.001$
- $0.005 \leq RMSE \leq 0.038$

Response Surface 2

SYSTEMATIC - 4149



RANDOM- 4149



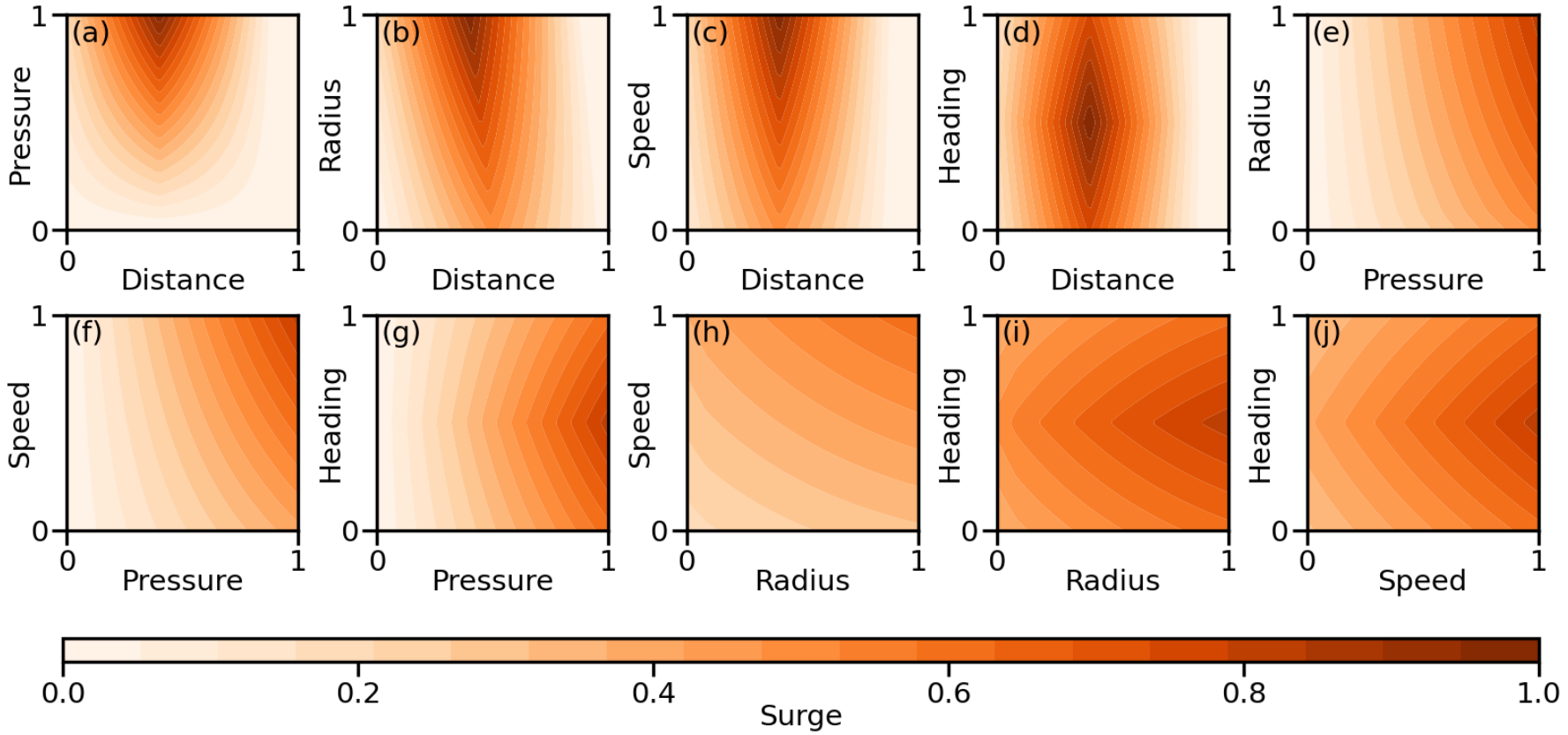
VIRGINIA TECH

Surge Response Surface Basis

- Single geographic location
- Dimensionless "unit" surge
- Follows Irish et al. (2008, *J Phys Ocean* & 2010, *Ocean Eng*)
- Five-dimensional "unit" track parameter space:
  - Distance (landfall location)
  - Pressure (pressure deficit)
  - Radius
  - Heading
  - Speed (forward speed)

"True" Oceanic Surge from Analytical Model

RESPONSE SURFACE 1

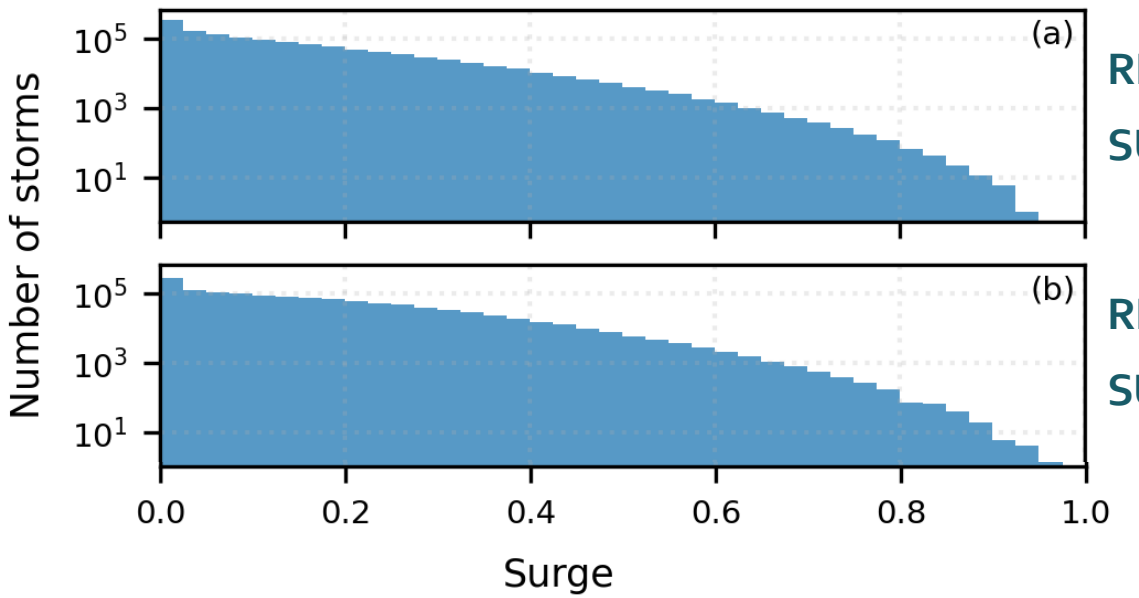


- Symmetric triangular on Distance and Heading
- Linear on all other parameters

BASIS

- 1.3-million storm systematic triangular set
- Error = Surrogate - Analytical

ERROR CHARACTERIZATION



RESPONSE SURFACE 1

RESPONSE SURFACE 2

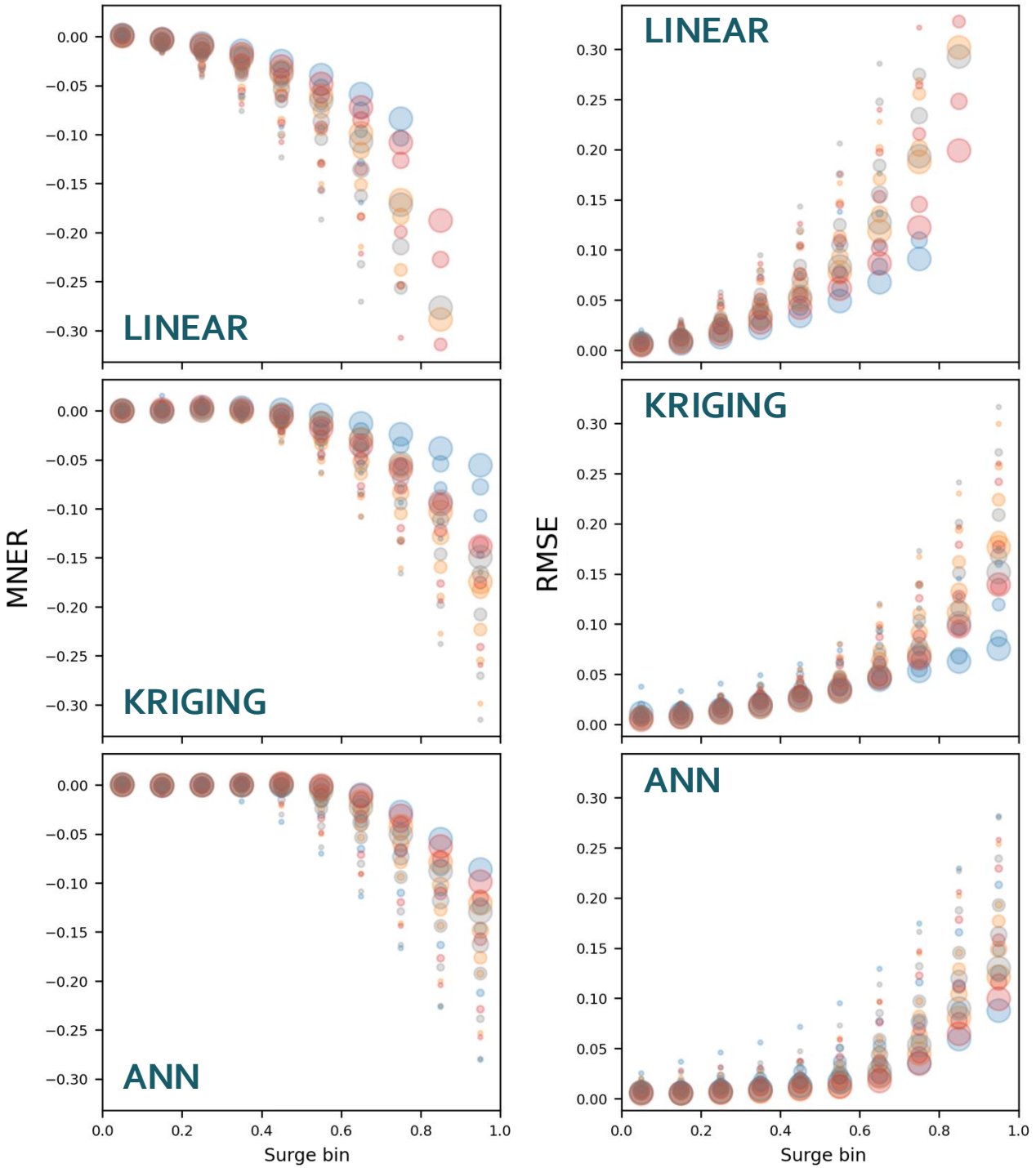
STATISTICS

- Root-mean-square error (RMSE)
- Mean error (MNER)
- Standard deviation of errors (STDE)
- Quartiles: Q1, Q2 (median), Q3
  - Interquartile range: IQR = Q3-Q1
- Considered:
  - **Aggregate** for entire 1.2-million storm set
  - **Binned by surge magnitude** (0.1-unit intervals)

Aggregate Error Statistics for all Models and all Samples

- $-0.020 \leq MNER \leq 0.003$
- $0.006 \leq RMSE \leq 0.048$

Response Surface 1

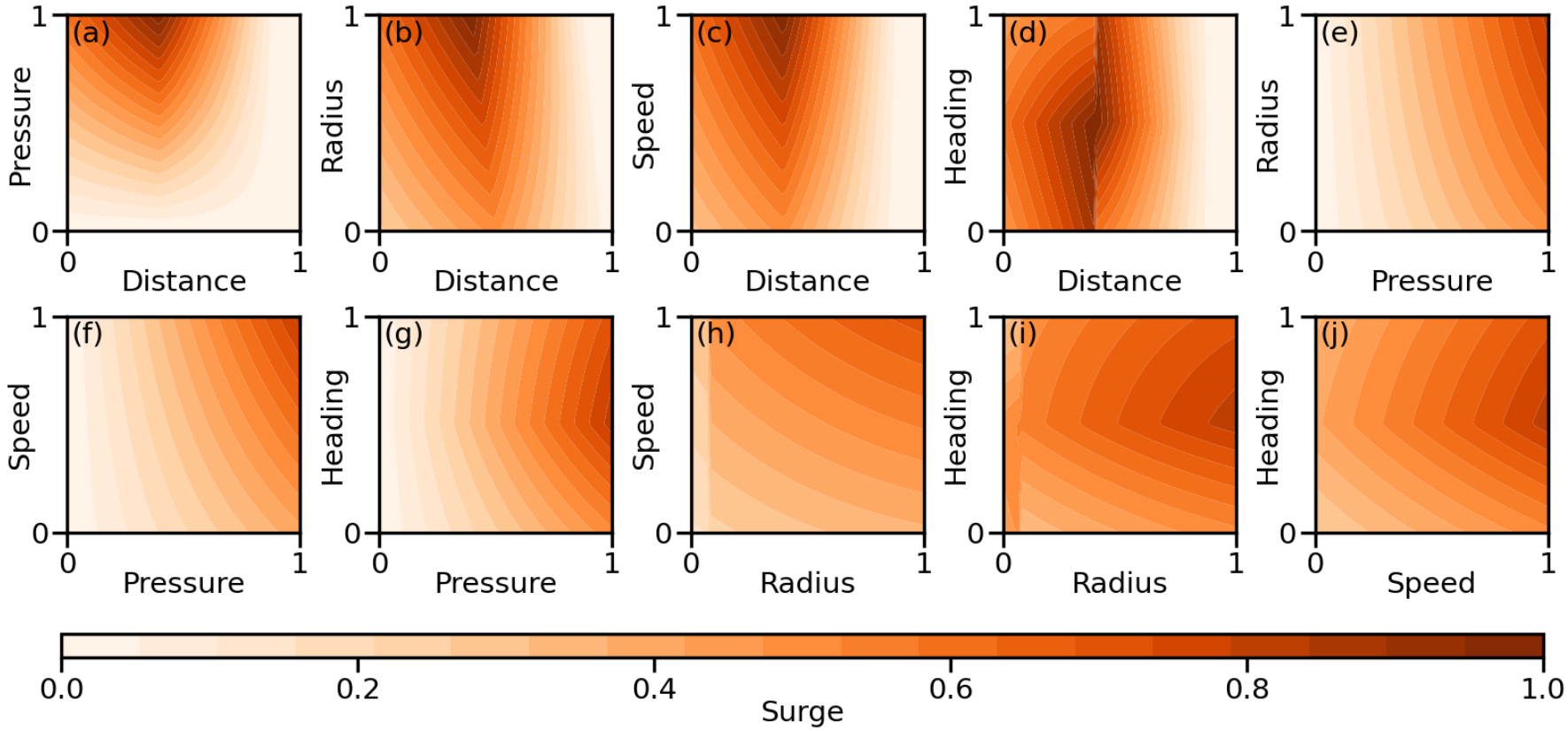


VIRGINIA TECH

Future Work

ADDITIONAL RESPONSE SURFACES

- Additional analytical surfaces
- Real-world computational simulation sets with thousands of storms



- Many people live near coast
- Death tolls large
- Direct damage from tropical cyclones is nearly half of all damage in U.S.
- Infrastructure is aging

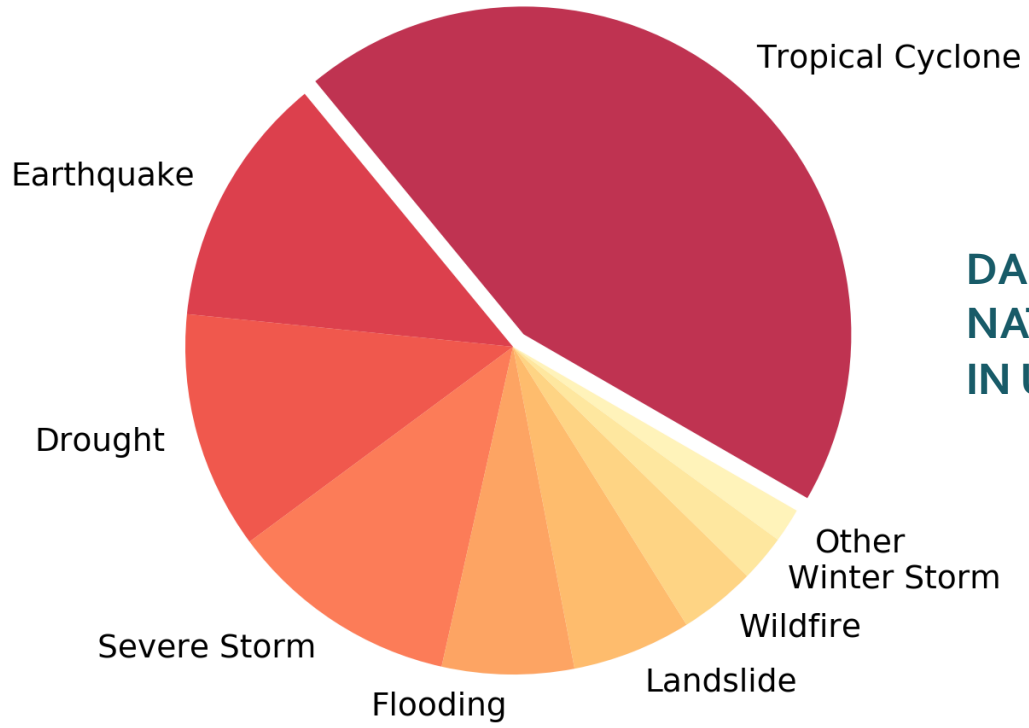
Death Tolls and Damage from Tropical Cyclones

TROPICAL CYCLONE DEATH TOLLS IN ASIA

- Typhoon Haiyan (2013): 6,300
- Typhoon Odisha (1999): >9,500
- Typhoon Bhola (1970): 300,000

TROPICAL CYCLONE DEATH TOLLS IN UNITED STATES

- Hurricane Maria (2017): 2,981
- Hurricane Sandy (2012): 159
- Hurricane Ike (2008): 112
- Hurricane Katrina (2005): 1,833
- 1900 Galveston Hurricane: 6,000

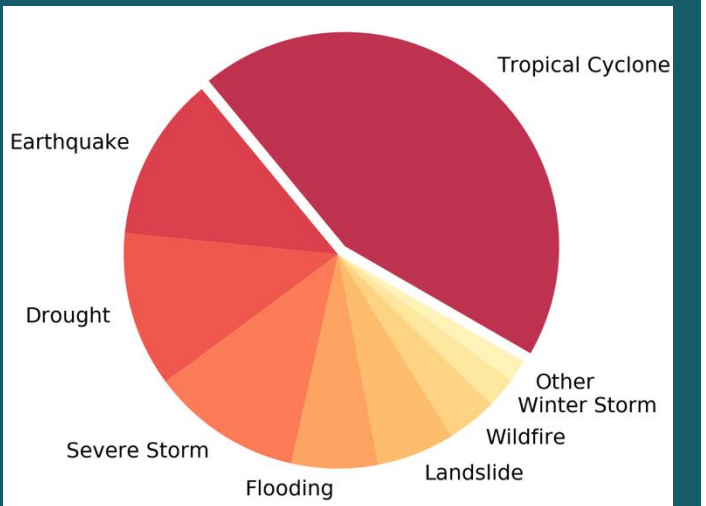


DAMAGE FROM NATURAL HAZARDS IN UNITED STATES

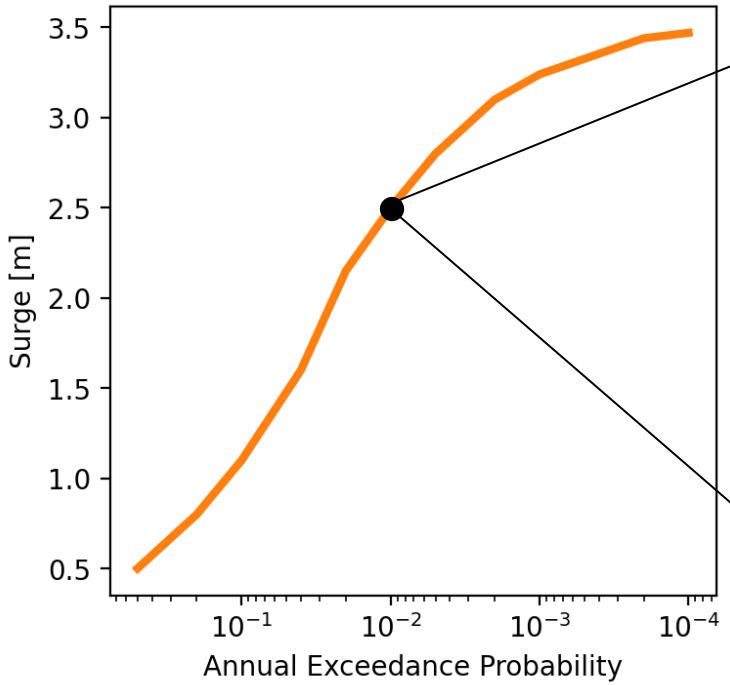
Damage based on Smith & Katz 2013, NOAA 2019, FEMA et al. 2017, Tuck et al. 1992, Tuck & Huckey 1994, Fleming & Taylor 1980, Dunbar & Weaver 2015

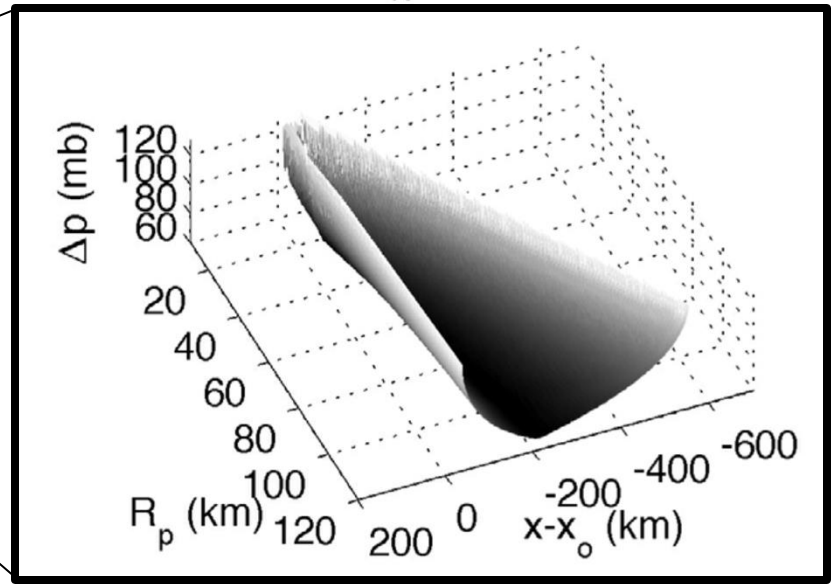
- Probabilistic Surge Hazard Assessment (PSHA) essential for robust engineering, planning and design

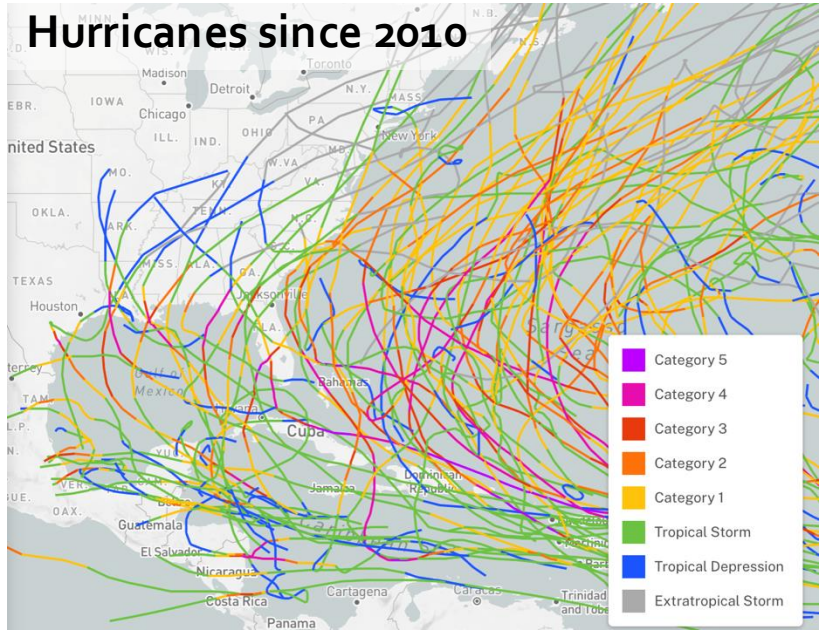
Damage from Natural Hazards



Probabilistic Surge Hazard Assessment with Surrogate Models







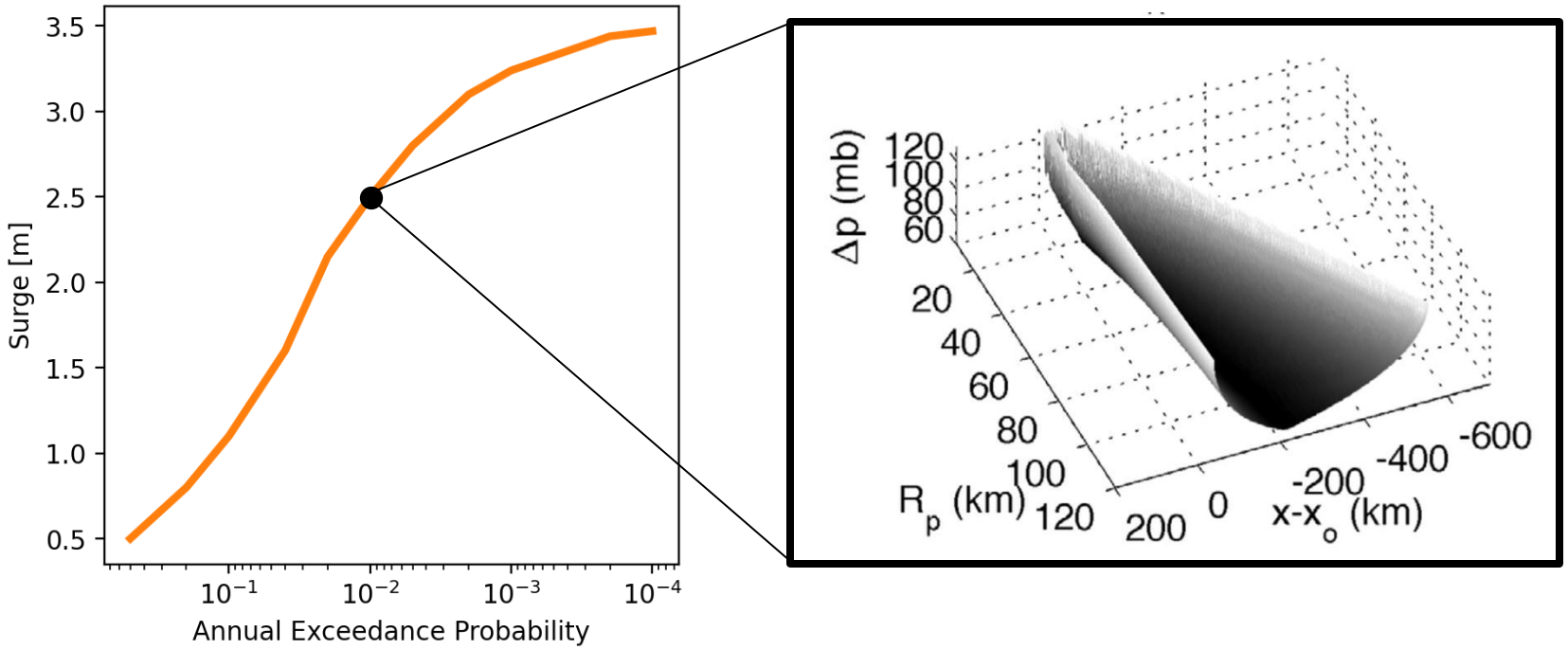
Left figure: Damage based on Smith & Katz 2013, NOAA 2019, FEMA et al. 2017, Tuck et al. 1992, Tuck & Huckey 1994, Fleming & Taylor 1980, Dunbar & Weaver 2015

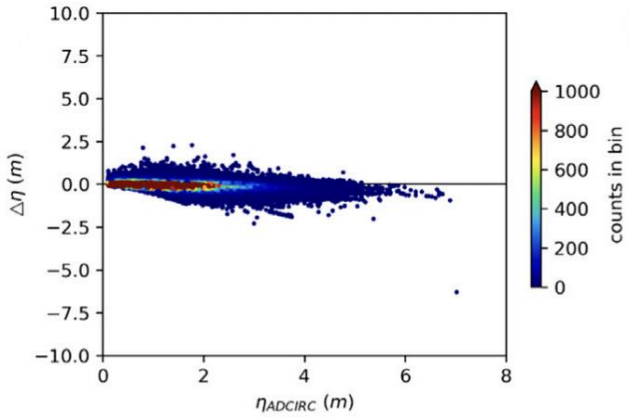
Top right figure from Irish et al. 2011, JGR. Bottom right figure from NOAA 2024.

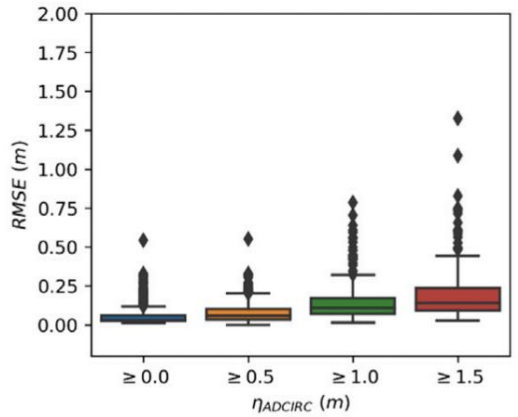
Probabilistic Surge Hazard Assessment with Surrogate Models

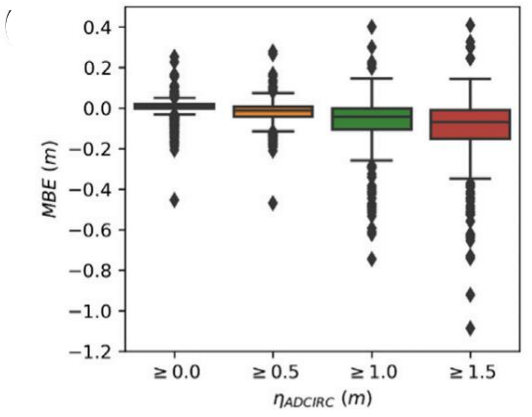
CHALLENGE

Published high-fidelity surge model sets may under-resolve storms in the extremes, leading to poorer surrogate model performance in the extremes









Top right figure from Irish et al. 2011, JGR. Lower figure from Lee* et al., 2021, Coast Eng.