SAND2021-8598PE

Sandia National Laboratories

Ensemble neural networks for post-processing contaminated Doppler LiDAR spectra to improve wind turbine wake diagnostics

Presented at:

Sandia Machine Learning and Deep Learning Conference 2021

Presented by:

Kenneth Brown, Thomas Herges, and David Maniaci

July 19, 2021

Wind Energy Technologies

Sandia National Laboratories

U.S. DEPARTMENT OF ENERGY NNSA National Nuclear Security Administration

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.

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.

Background

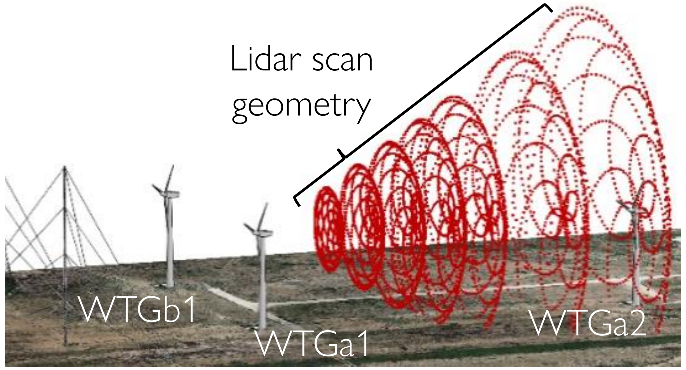
Wind Turbine Wake Diagnostics:

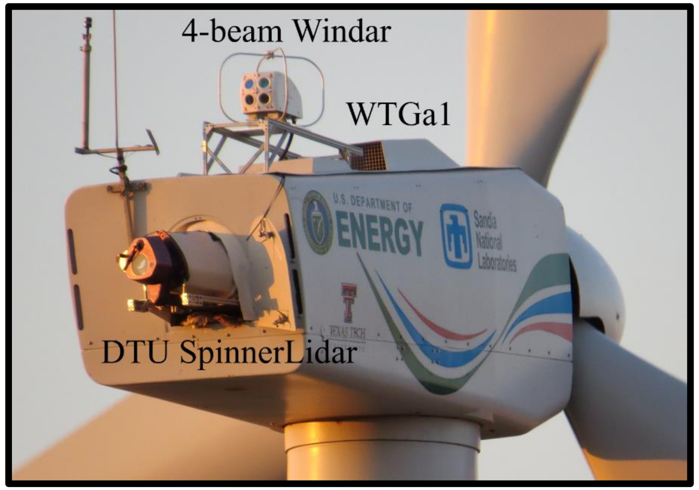
Facility

- Scaled Wind Farm Technology (SWiFT) facility in Lubbock, Texas, USA
  - Characterization of the atmospheric conditions in [1], recent benchmarking activities given in [2]

Lidar

- Continuous-wave DTU SpinnerLidar [3] rear-mounted on WTGa1
- Focus = 105 m from WTGa1 along the axis of the turbine rotor
- A rosette pattern is completed in 2 s and consists of 984 measurement locations, some below ground





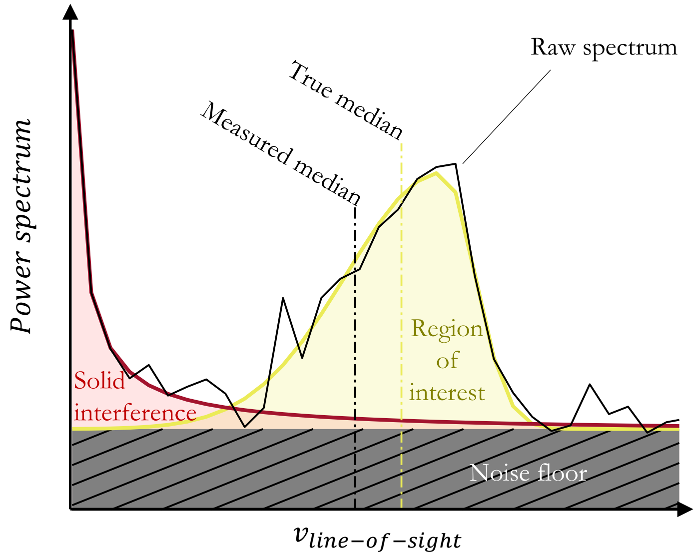
(Images from [9])

1. Kelley, C.L. and B.L. Ennis, *SWiFT site atmospheric characterization*. 2016, Sandia National Lab.(SNL-NM), Albuquerque, NM (United States).
2. Doubrawa, P., et al., *Multimodel validation of single wakes in neutral and stratified atmospheric conditions*. Wind Energy, 2020.
3. Mikkelsen, T., et al., *A spinner-integrated wind lidar for enhanced wind turbine control*. Wind Energy, 2013. 16(4): p. 625-643.

Problem

Modern lidars show biases ~≤ 0.2-m/s and std. dev. ~≤ 0.20-m/s[1] depending mostly on the inhomogeneities in the flow

Largely unquantified errors stem from contamination by solid interference and amplitude noise



Problem: how to extract true statistics of the region of interest (i.e., spectral median and spectral standard deviation) while ignoring contamination from amplitude noise and solid interference?

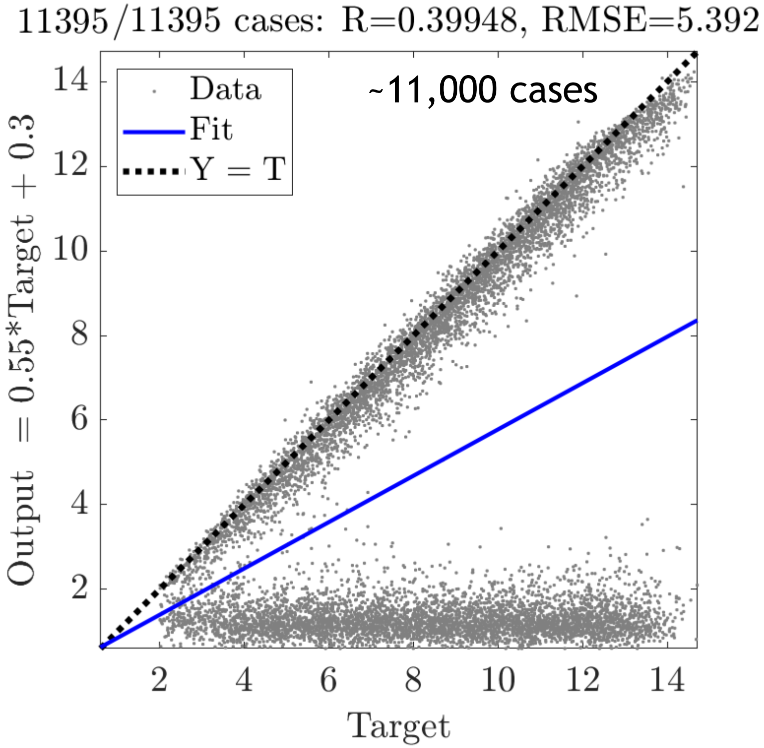
Solution: train a supervised regressive machine learning model based on a large database of synthetic spectra to discern between true signal and interference for any likely spectral shape

1. Courtney, M., R. Wagner, and P. Lindelöw, *Testing and comparison of lidars for profile and turbulence measurements in wind energy.* IOP Conference Series: Earth and Environmental Science, 2008. **1**.

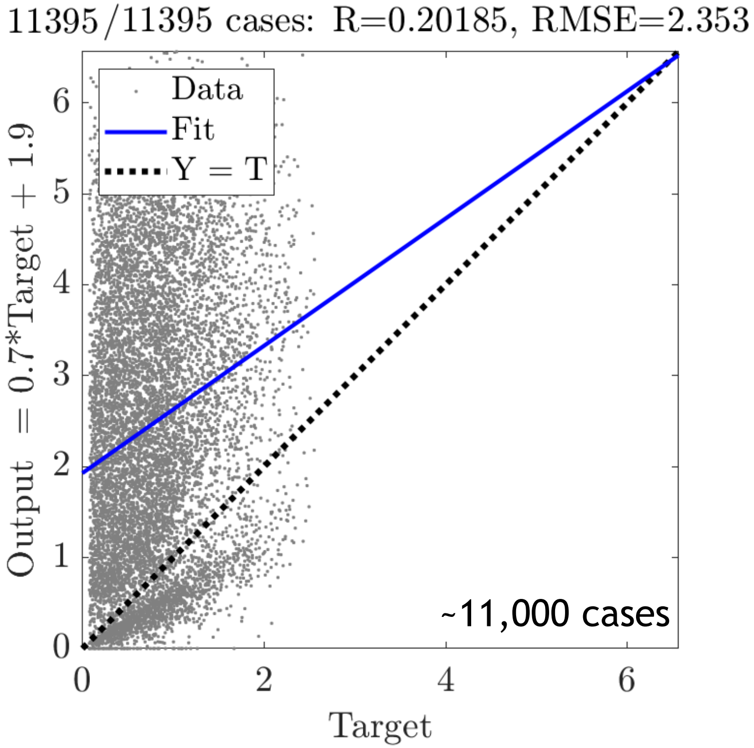
Reference Technique

Raw processing: subtract background noise and calculate quantities of interest (QoIs)

Spectral median



Spectral standard deviation

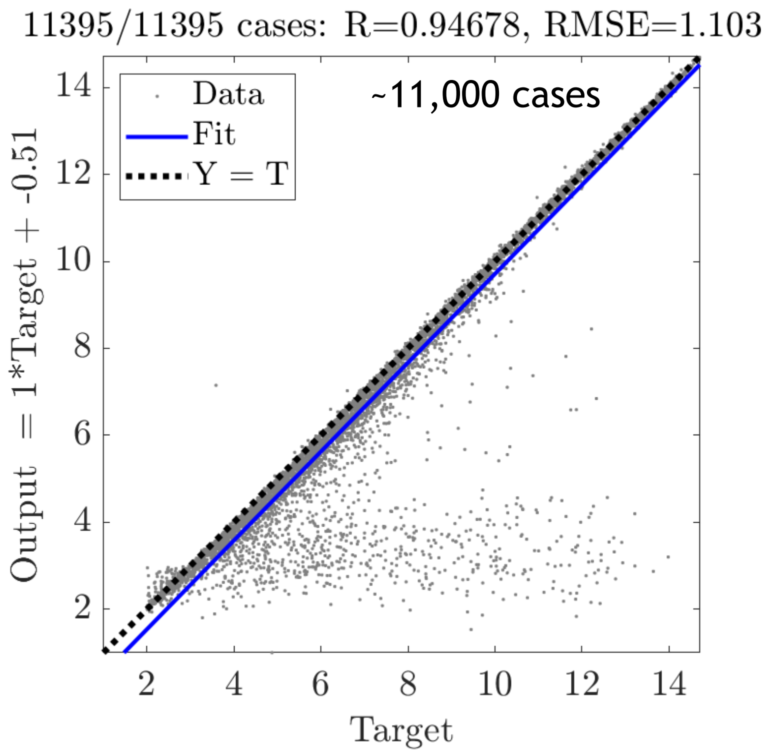


Mean and random errors are substantial

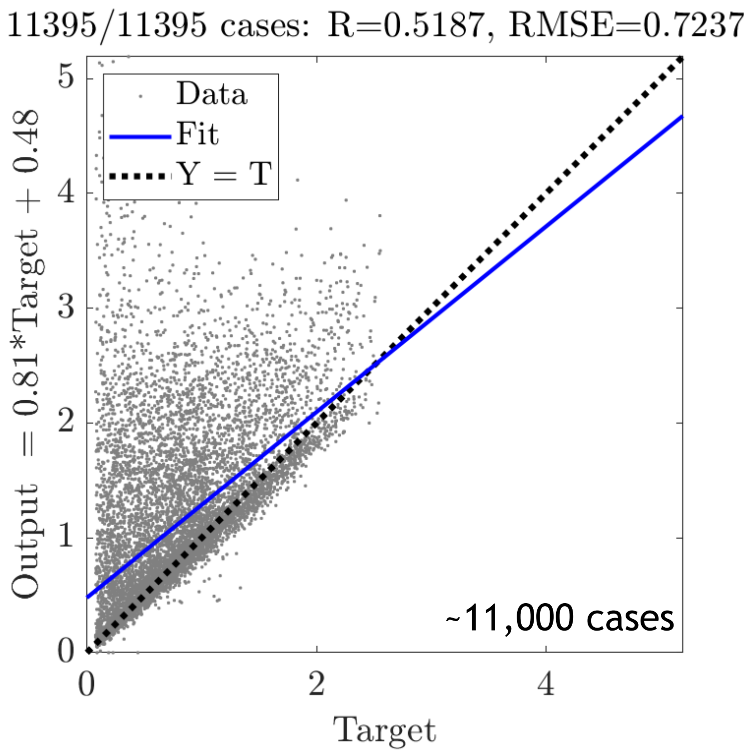
Reference Technique

Quality control processing: subtract background noise, detect and (mostly) filter interference features, and calculate QoIs

Spectral median



Spectral standard deviation



Significant improvement, but we can still do better…

Algorithmic Approach

Individual neural network architecture:

Hidden layer architecture
- Six layers of perceptrons with 48 nodes each

Activation function
- Sigmoid symmetric

Performance function
- MSE

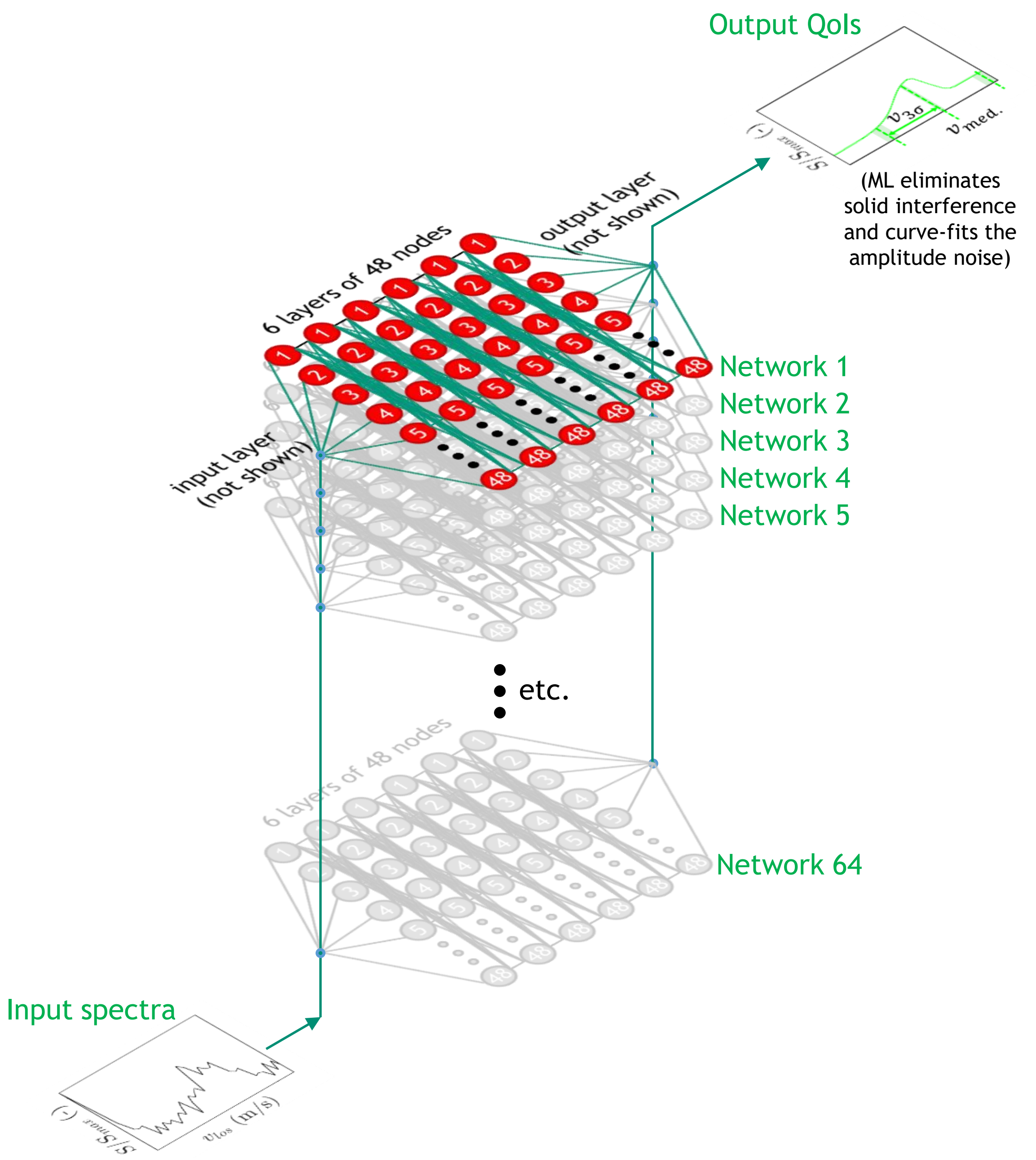
Backpropagation algorithm
- Levenberg-Marquardt

Ensemble neural network architecture:

Number of ensembles
- $N = 64$ (diminishing returns for $N > 32$)

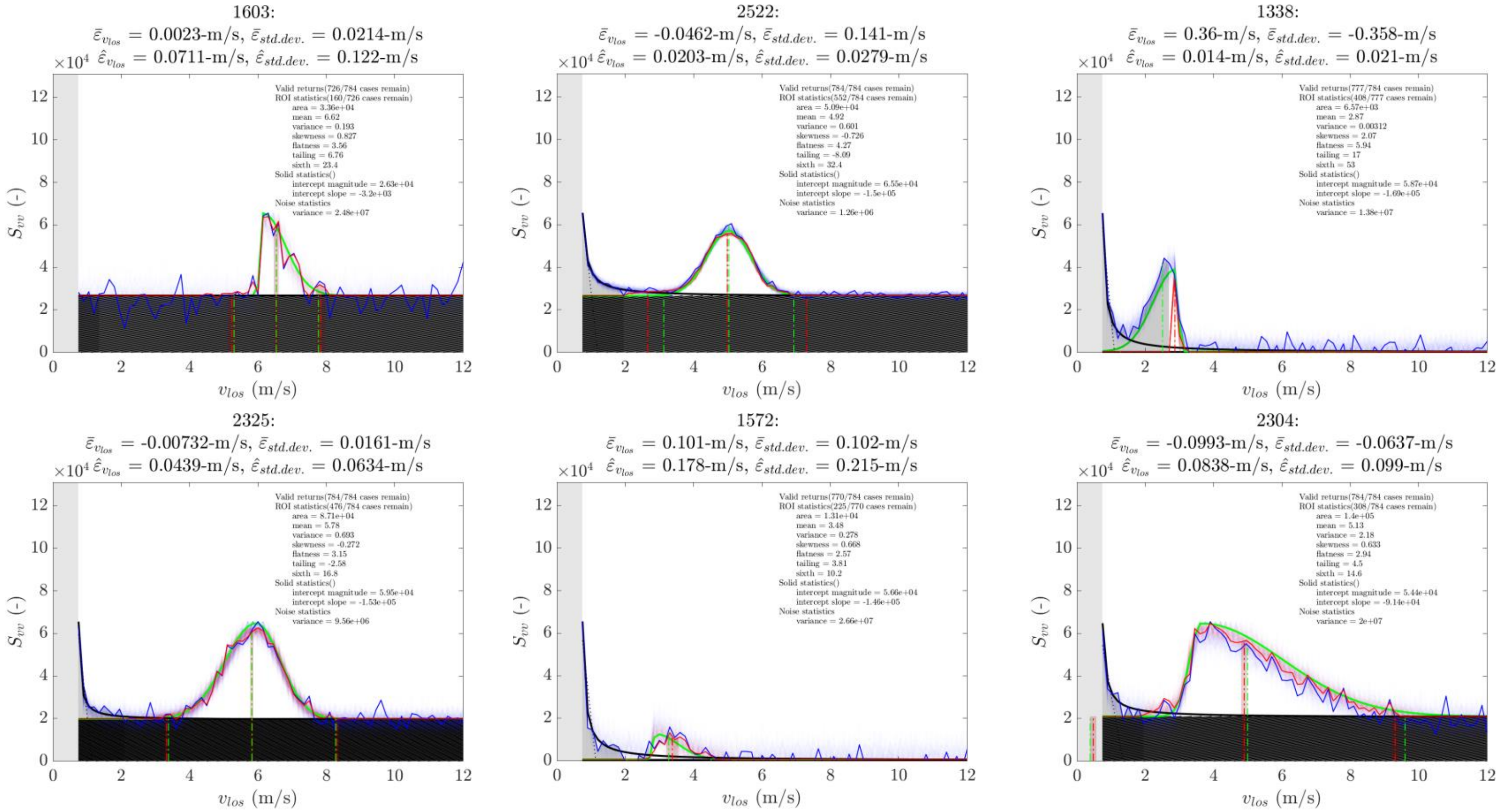
Resampling technique for different ensemble members
- Bootstrapping

Statistics calculated across ensembles for each QoI
- Mean, μ
- Standard error, $\sigma/\sqrt{N-1}$



Data Description

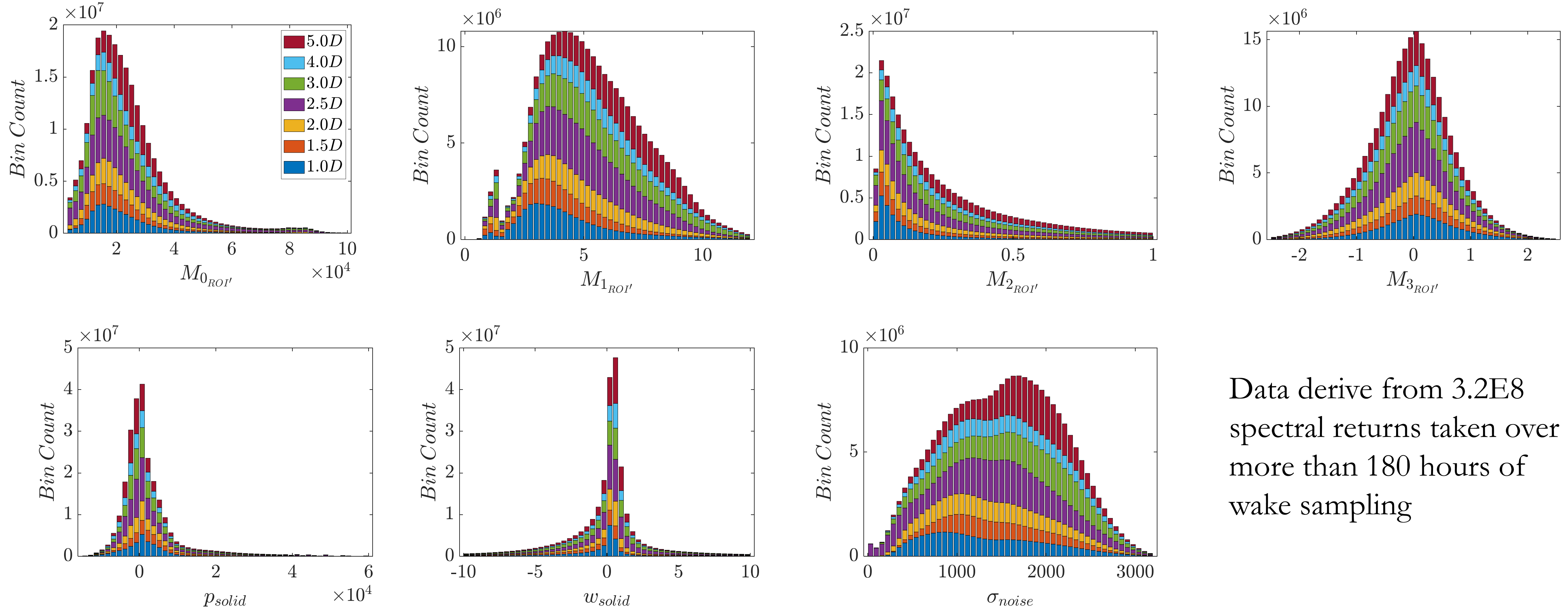
Example Spectra (from field test):



Signal of interest varies widely in shape but falls somewhere along the first 100 bins of the spectrum

Data Description

Histograms of measured spectral parameters:



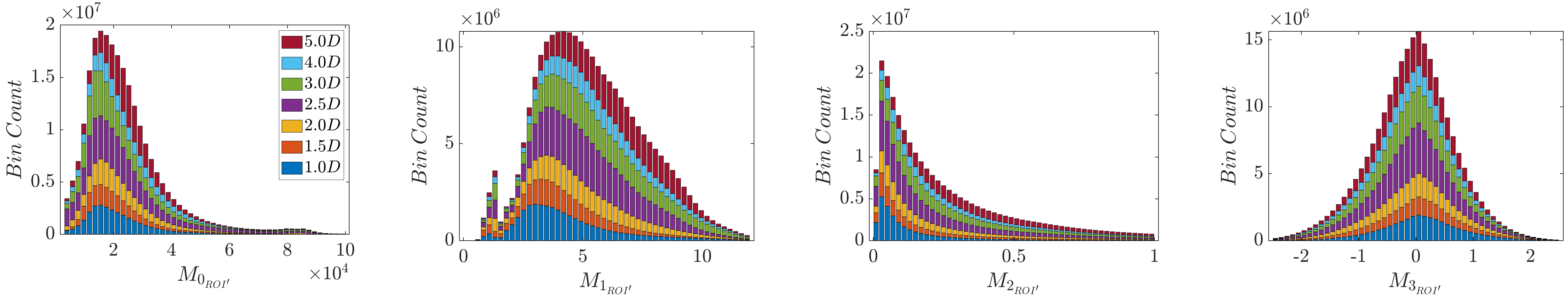
Data derive from 3.2E8 spectral returns taken over more than 180 hours of wake sampling

Measured data used to inform generation of synthetic database*

*Synthetic data has known ground truth parameters with artificial interference added

Data Description

Histograms of measured spectral parameters:



Database of synthetic returns:

Baseline spectra:

$$s_{ROI} = \frac{M_{0_{ROI}}}{M_{2_{ROI}}\sqrt{2\pi}}\, e^{-\frac{1}{2}\left(\frac{v_{los}-M_{1_{ROI}}}{M_{2_{ROI}}\,(1\mp M_{3_{ROI}})}\right)^2}$$

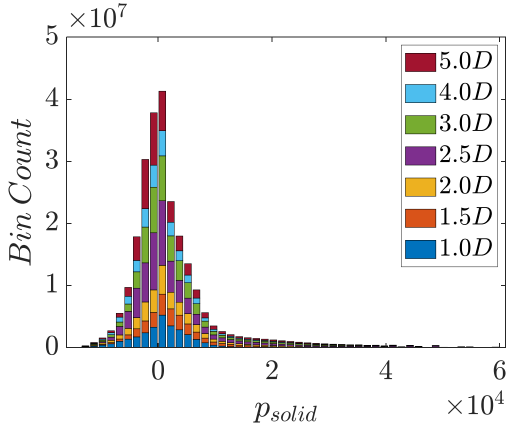
(this is a scaled epsilon-skew-normal distribution[1])

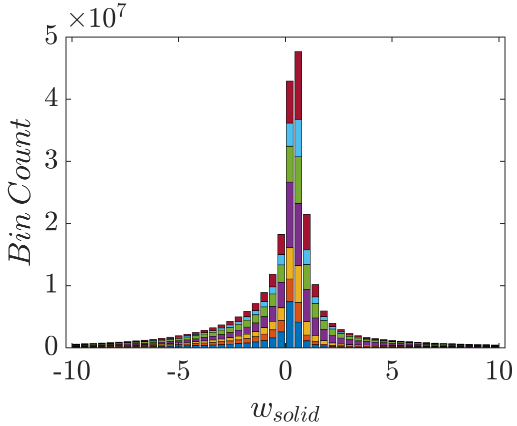
where… $M_{0_{ROI}}$ is a magnitude parameter
$M_{1_{ROI}}$ is a location parameter
$M_{2_{ROI}}$ is a width parameter
$M_{3_{ROI}}$ is a skew parameter whose absolute value is less than one, and the $\mp$ takes the sign opposite of the numerator of the exponent.

1. Mudholkar, G.S. and A.D. Hutson, *The epsilon–skew–normal distribution for analyzing near-normal data*. Journal of statistical planning and inference, 2000. 83(2): p. 291-309.

Data Description

Histograms of measured spectral parameters:





Database of synthetic returns:

Solid interference:

$$s_{solid} = \frac{p_{solid}}{1 + (v_{los} - v_{solid})/w_{solid}}$$

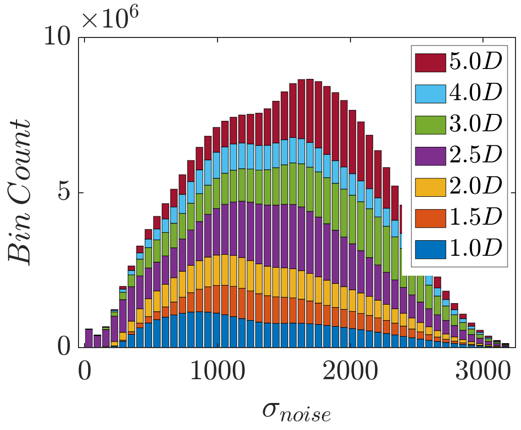
where… w_{solid} is the full-width half-maximum of the solid-interference spectrum

p_{solid} is the prominence of the solid interference

v_{solid} is the velocity at p_{solid}

Data Description

Histograms of measured spectral parameters:



Database of synthetic returns:

Amplitude noise:

s_{noise} is the noise spectrum, which is generated given a variance within each spectral bin. The variance is here taken to be uniform over the spectrum. Each synthetic spectra includes a different randomized distribution of Gaussian noise.

Data Description

Database of synthetic returns:

Baseline spectra:

$$s_{ROI} = \frac{m_{0ROI}}{m_{2ROI}\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{v_{los}-m_{1ROI}}{m_{2ROI}(1\mp m_{3ROI})}\right)^2}$$

(this is a scaled epsilon-skew-normal distribution[1])

where… m_{0ROI} is a magnitude parameter
m_{1ROI} is a location parameter
m_{2ROI} is a width parameter
m_{3ROI} is a skew parameter whose absolute value is less than one, and the $\mp$ takes the sign opposite of the numerator of the exponent.

Solid interference:

$$s_{solid} = \frac{p_{solid}}{1 + (v_{los} - v_{solid})/w_{solid}}$$

where… w_{solid} is the full-width half-maximum of the solid-interference spectrum
p_{solid} is the prominence of the solid interference
v_{solid} is the velocity at p_{solid}

Amplitude noise:

s_{noise} is the noise spectrum, which is generated given a variance within each spectral bin. The variance is here taken to be uniform over the spectrum. Each synthetic spectra includes a different randomized distribution of Gaussian noise.

The combined synthetic PSD:

$$s = s_{ROI} + s_{solid} + s_{noise}$$

1. Mudholkar, G.S. and A.D. Hutson, *The epsilon–skew–normal distribution for analyzing near-normal data*. Journal of statistical planning and inference, 2000. 83(2): p. 291-309.

Data Description

Database of synthetic returns:

Can generate millions of cases easily with the cluster

We'll use a subset (~76,000 cases) here to evaluate different parameter estimation approaches
note: subset presented here includes oversampling of unusual/outlier spectral shapes (to give better "recall")

Division of cases:

- Training data (70% of total data, or ~53,000 cases) – parametrically varied inputs; parameter ranges (but NOT distributions) are matched to that of the population of observed returns

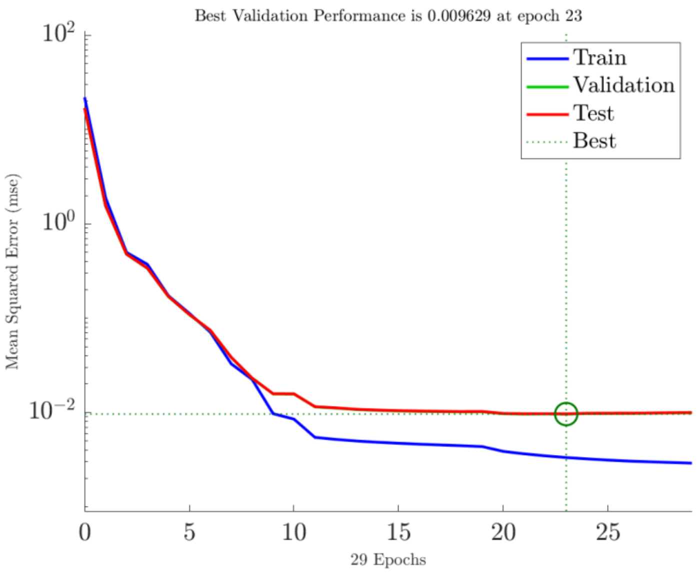
- Validation data (15% of total data, or ~11,000 cases) – randomized uniform distribution within the population of observed returns

- Testing data (15% of total data, or ~11,000 cases) – randomized uniform distribution the population of observed returns

Retained data: first 129 bins of spectrum:

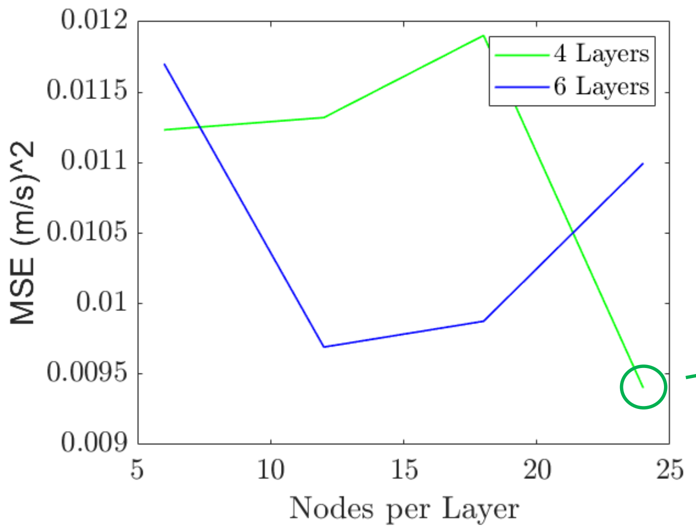
Results – Individual Network Performance

Hyperparameter sweep

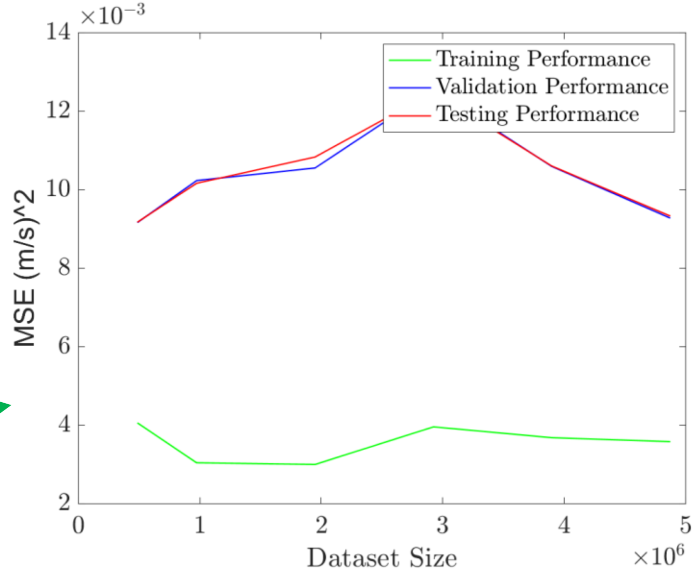


Results – Individual Network Performance

Architecture selection



Dataset size selection

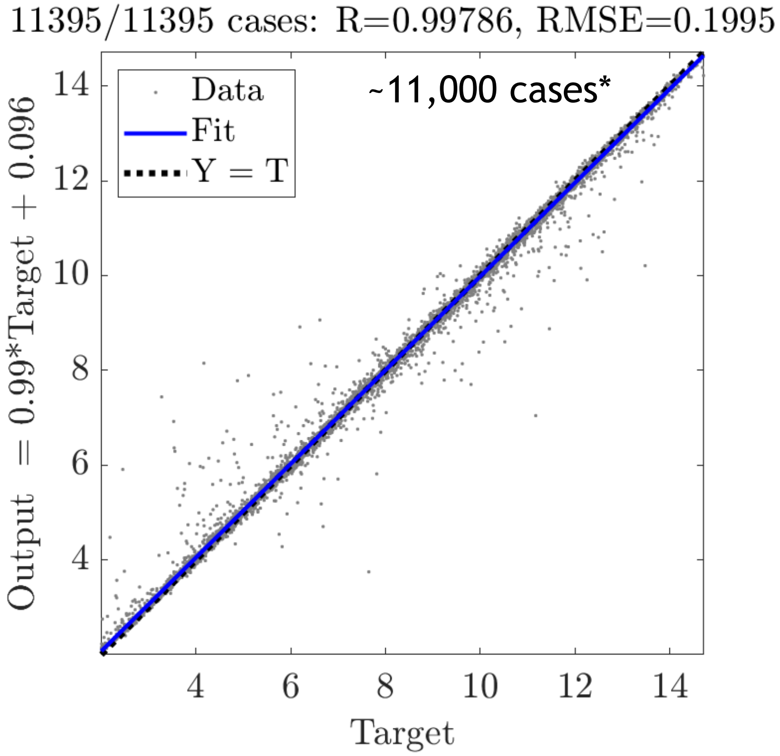


Update: architecture has since been modified to be six layers of perceptrons with 48 nodes each

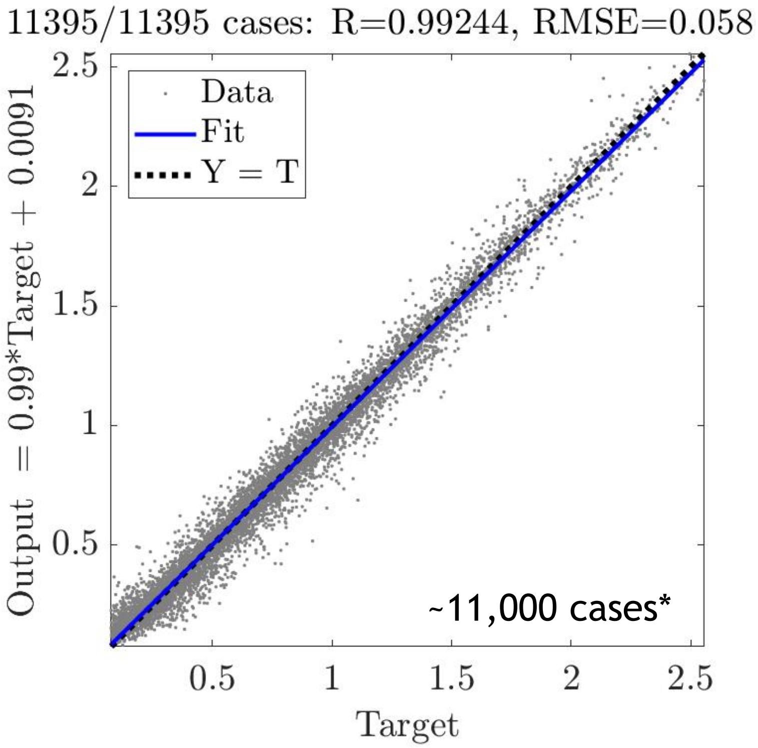
Results – Ensemble Network Performance

Baseline approach: input raw spectra directly into neural network ensemble

Spectral median



Spectral standard deviation



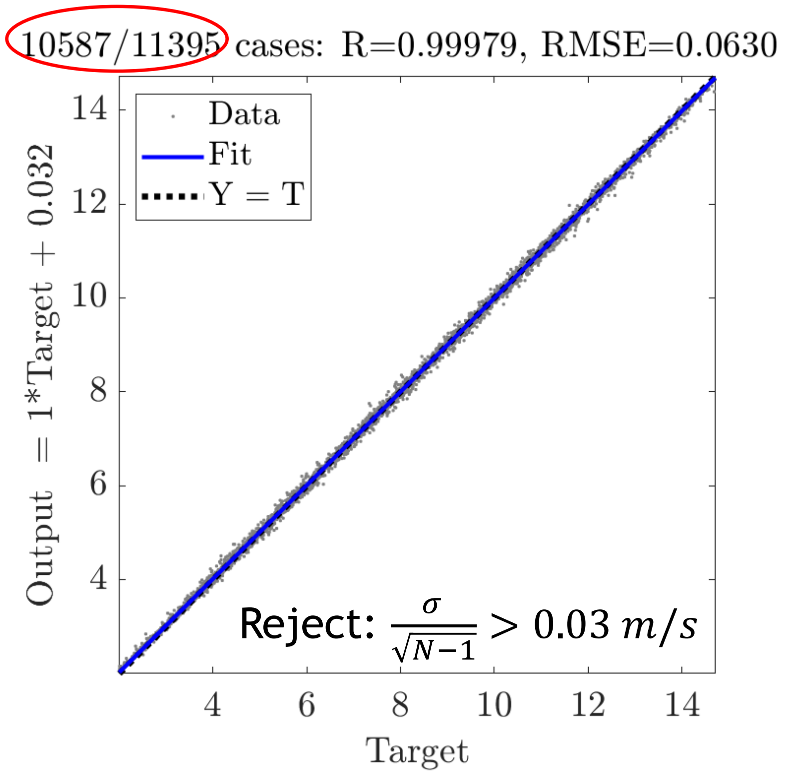
Results are good, but there are still a few outliers

*results show test data only (not training or validation data)

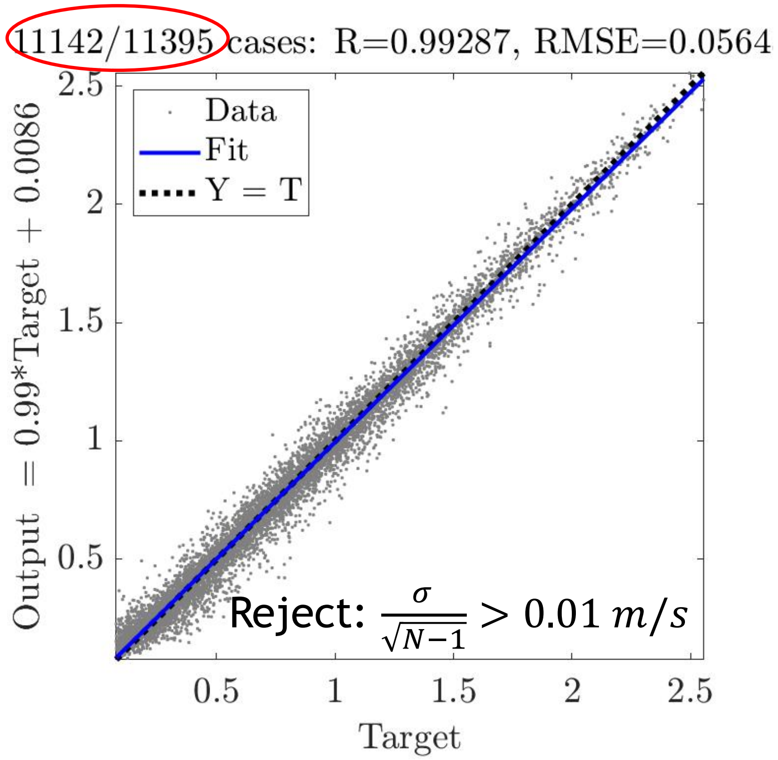
Results – Ensemble Network Performance

Uncertainty threshold approach: input raw spectra directly into neural network ensemble, reject cases where variance of ensemble estimates is large

Spectral median



Spectral standard deviation

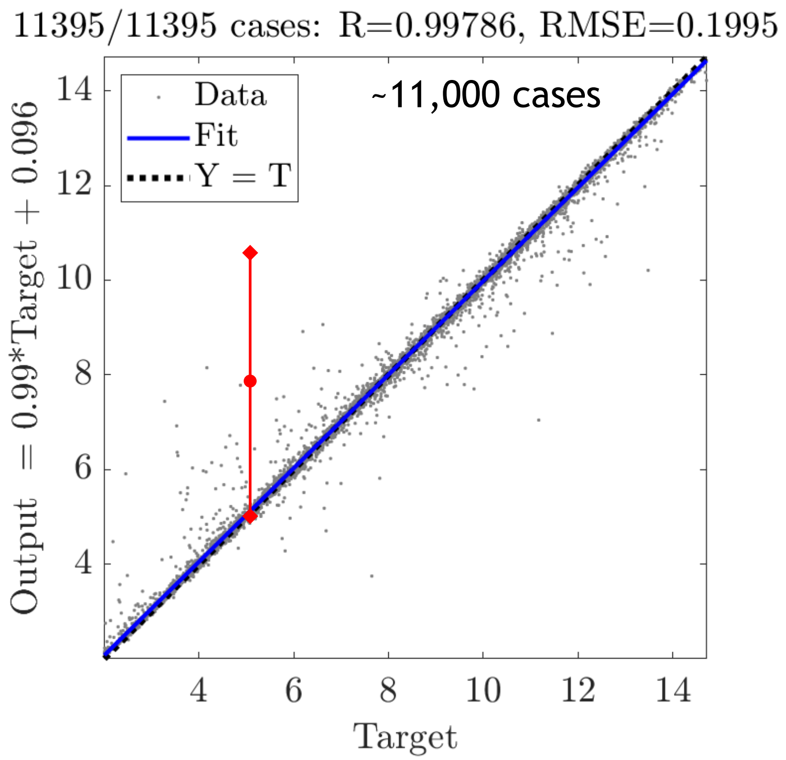


The results are good, although we have reduced data availability

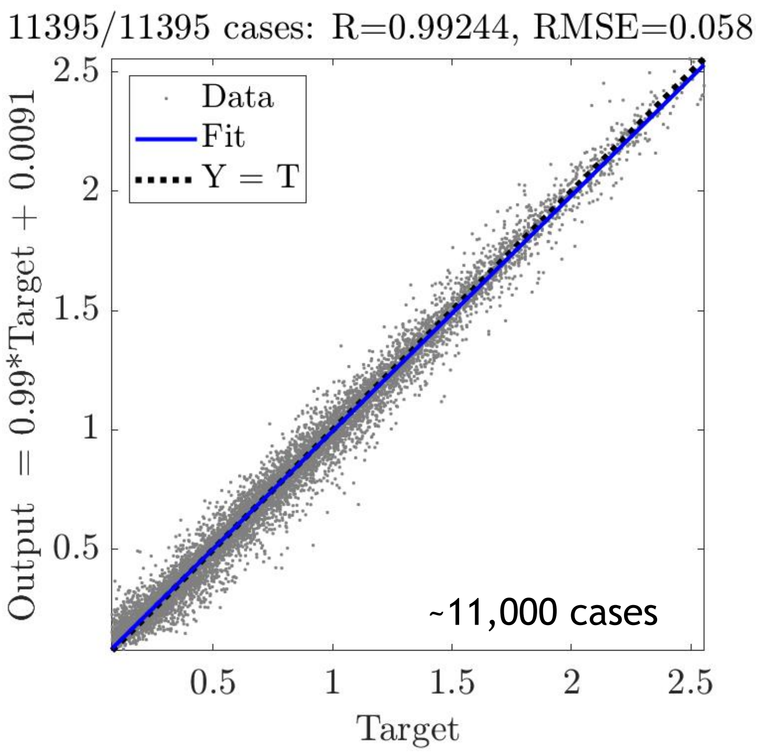
Results – Ensemble Network Performance

Baseline approach: input raw spectra directly into neural network ensemble

Spectral median



Spectral standard deviation

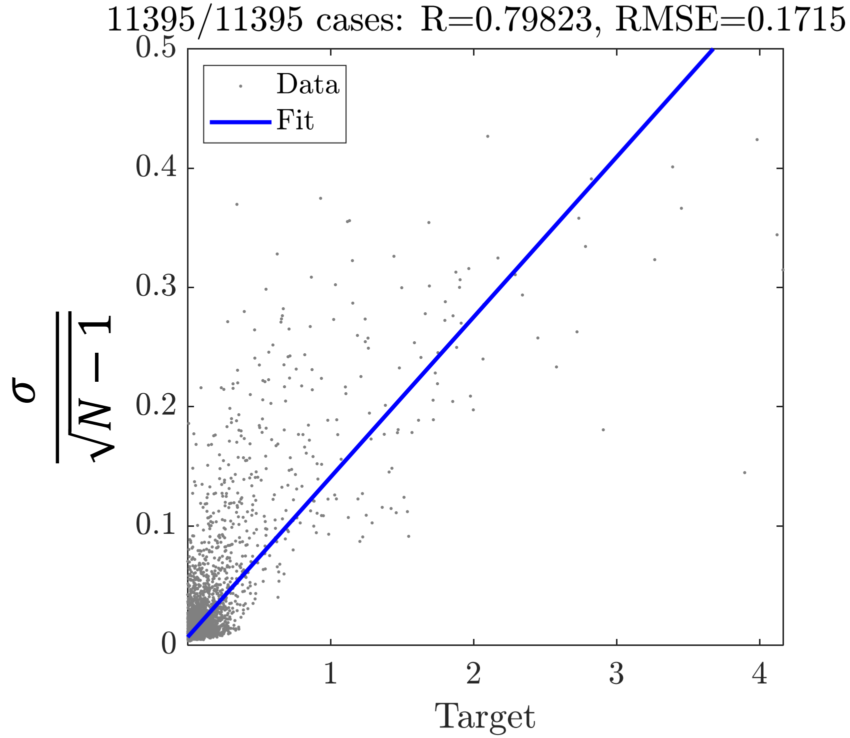


Results are good, but there are still a few outliers

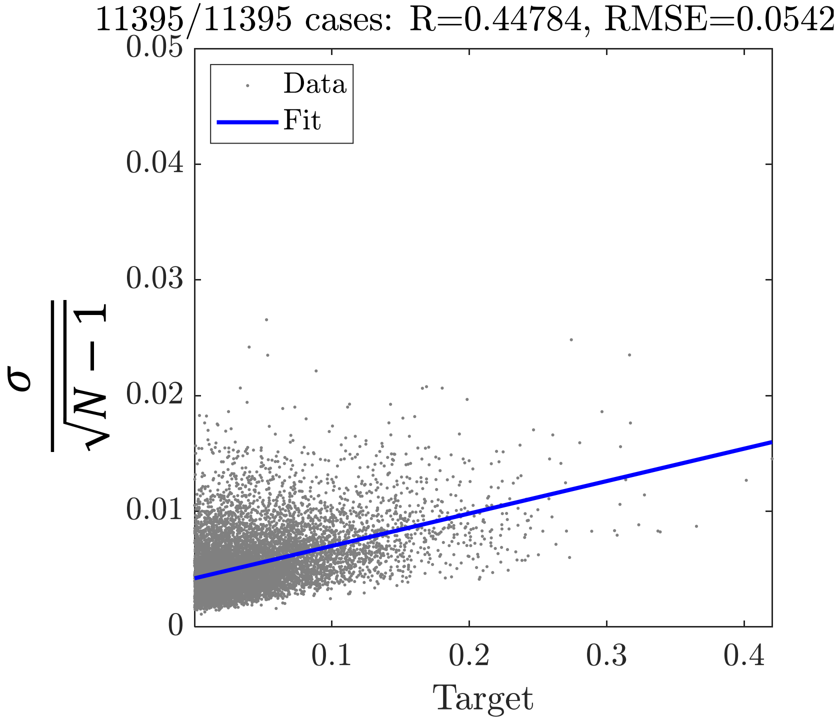
Results – Ensemble Network Performance

Uncertainty quantification approach: input raw spectra directly into neural network ensemble, use variance of ensemble estimates to quantify uncertainty

Error in spectral median



Error in spectral standard deviation



UQ results have a LOT of scatter

Conclusions

Machine learning techniques are being applied to solve UQ problems related to wind turbine wake diagnostics

Ensemble neural networks are effective to reduce mean and random errors during post-processing of Doppler lidar spectra

- Ensemble approach leveraging 64 networks to reduce uncertainty has been developed for lidar post-processing
- Initial results show less than 0.03 m/s (~0.3%) standard error of the ensemble mean for $v_{med.}$ with marginal change in data availability

Ongoing work to investigate how to get individual UQ estimates for each input spectrum

Initial Ideas for Improvement of "UQ Approach"

Post-processing:

Use higher-order statistics than just standard error to estimate uncertainty

Network setup:

Training data:

- Generate more training data
- Stop oversampling outliers

Training methodology:

- Use different backpropagation algorithm
- Use different machine learning technique
- Use different ensemble configuration

Network type:

- Use different Bayesian inference or Monte-Carlo dropout

ML type:

- Random forest, auto encoder, others?

Thank you!

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.

Sandia National Laboratories

Backup Slides

ML processing scheme

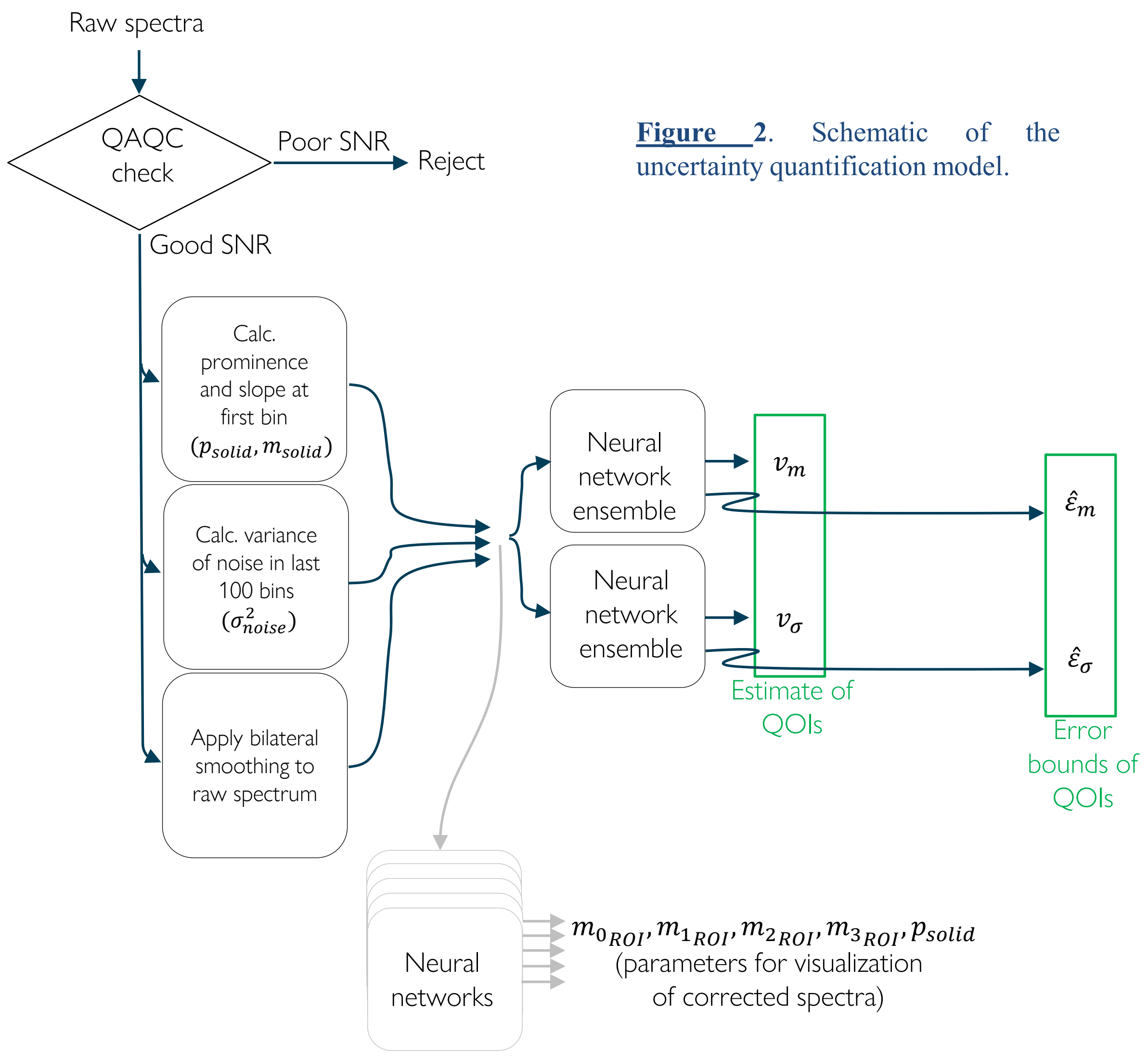


Figure 2. Schematic of the uncertainty quantification model.