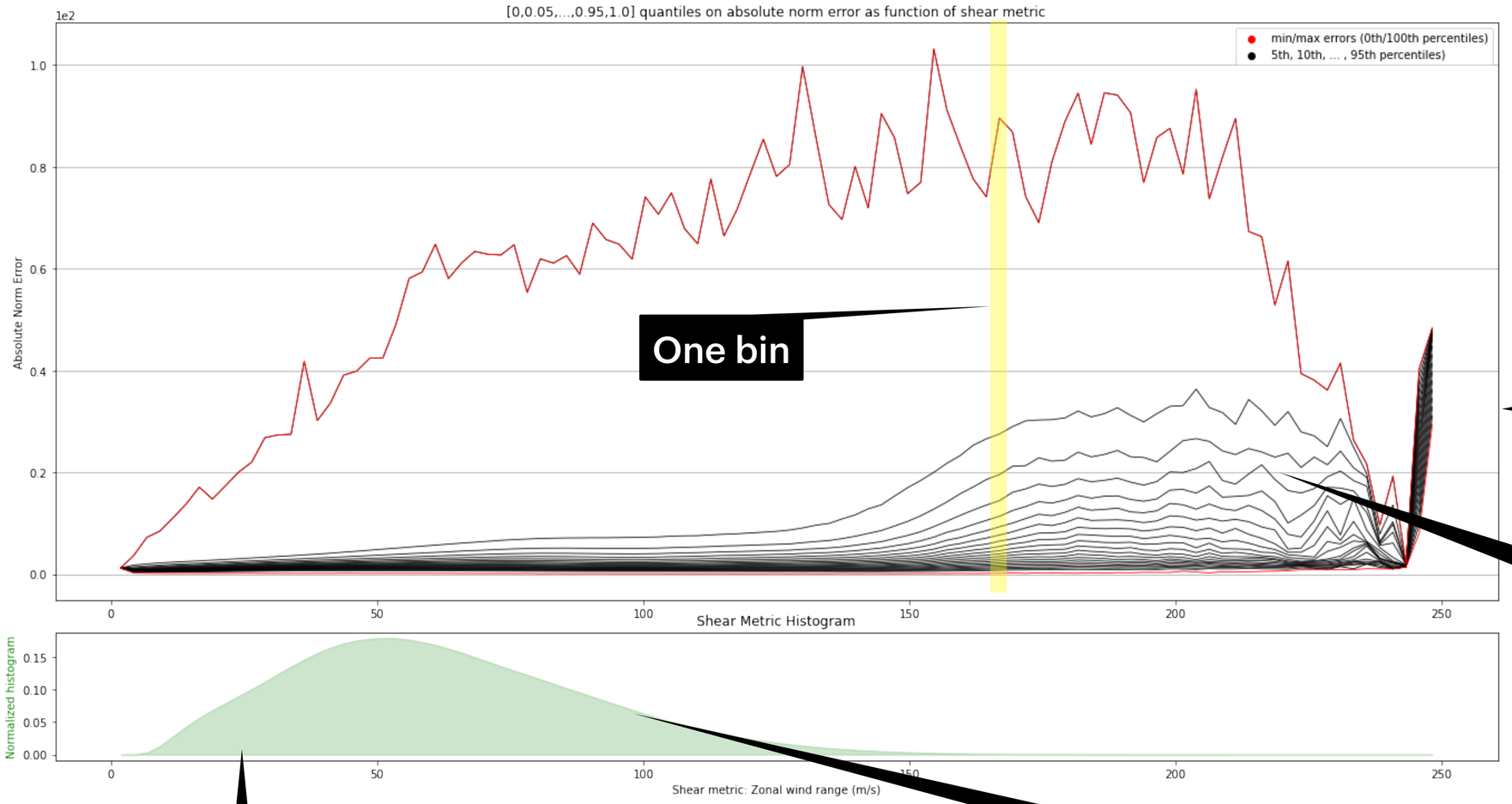
Sampling Strategies for Training Machine Learning Emulators of Gravity Wave Parameterizations

IMSI Machine Learning for Climate and Weather Applications [November 2, 2022]

L. Minah Yang, Ed Gerber
New York University

Imbalanced Data and Regression



1. The 1D projection is designed to put more ``difficult'' samples at its high end.
2. Low frequency samples are associated with larger errors.

Absolute Norm error: $\|y - \hat{y}\|_2^2$

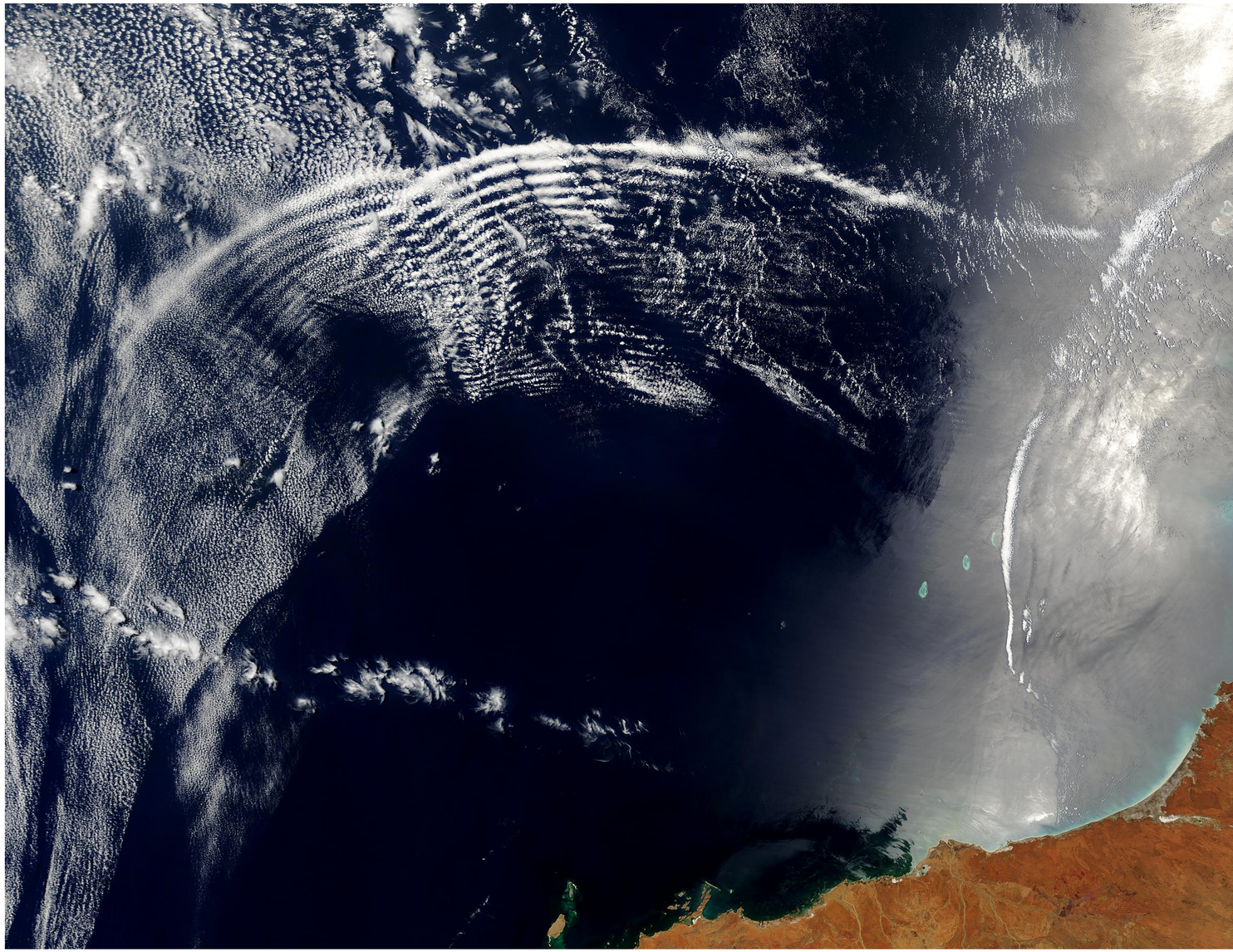
Error quantiles as a function of 1D projection

Projection of high dimensional data onto 1-dimensional space

Histogram of dataset on this 1D projection

Application Background

- **Subgrid-scale parameterization** is necessary in modeling Earth.
  - Dynamical core include resolved equations of motions.
  - Physics package includes:
    - chemistry (atmospheric aerosols),
    - turbulence and convection (cloud microphysics & moisture)
    - **gravity waves.**



NASA Visible Earth

Datawave

- Develop data-driven parameterizations of gravity waves for use in GCMs.
- **Current state of things & challenges**
  - Many sources of gravity waves each require different models.
  - Observations do not capture the entire spectra for GWs.
  - GCMs may partially resolve GWs.
  - Single column models.
  - Difficult to relax assumptions in models.
    => should use **data**!
- First step: Can ML models emulate existing parameterizations?
  - WaveNet [Espinosa 2022 GRL]: Yes, with caveats.
    - Sensitivity to data
  - Offline: ML emulator performance w.r.t. distance metrics
  - Online: reasonable climatology, annual cycle, Quasi-Biennal Oscillation
  - Great offline performance doesn’t always give us great online performance

Gravity wave parameterization in GCMs

- AD99 = Alexander-Dunkerton 1999
  - The parameterization is based on linear theory and adheres closely to fundamental principles of conservation of wave action flux, linear stability, and wave–mean-flow interaction.
  - simple assumption: momentum fluxes carried by monochromatic waves are deposited locally and entirely at the altitude of linear wave breaking.
  - Allows for any desired input spectrum of momentum flux.
- Implementation in MiMA configured with fixed source that varies across latitudes.
- Appears as RHS forcing term.
- We emulate it with ML.

Dataset Description

- MiMA = Model of an idealized Moist Atmosphere, an intermediate complexity GCM.
- MiMA 1.0 (Jucker & Gerber 2017)
    - used in Espinosa GRL 2022
    - Aquaplanet
        - moisture (latent release)
        - Betts-Miller convection (Betts & Miller 1986)
        - Rapid Radiative Transfer Model radiation scheme (Mlawer et al. 1997; Iacono et al. 2000)
- MiMA 2.0 (Garfinkel et al.)
    - Introduction of land-sea contrast:
        - mechanical damping of near-surface winds (rough land, smooth ocean)
        - evaporation between land and ocean,
        - heat capacity
    - Redistributed heat in the ocean to respect zonal asymmetry in ocean
    - Topography
    - WaveNet unsuccessful when coupled to MiMA 2.0.

Disparity in offline and online performance

- Earth system is a chaotic dynamical model.
  - Many, very small errors may nudge the solution away from attractor.
  - Rarely occurring, not small errors may push solution away from attractor.
  - We want to minimize spread and bias of error!

Shear metric

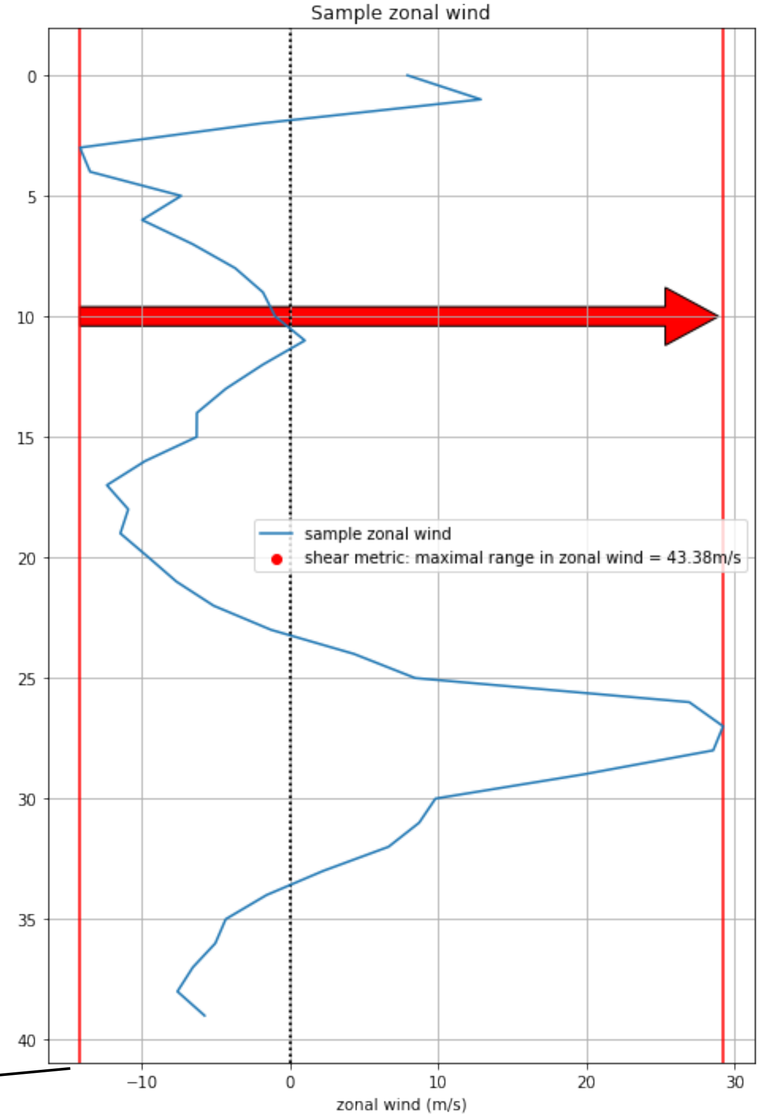
- We consider top level error as a function of a shear-related metric.

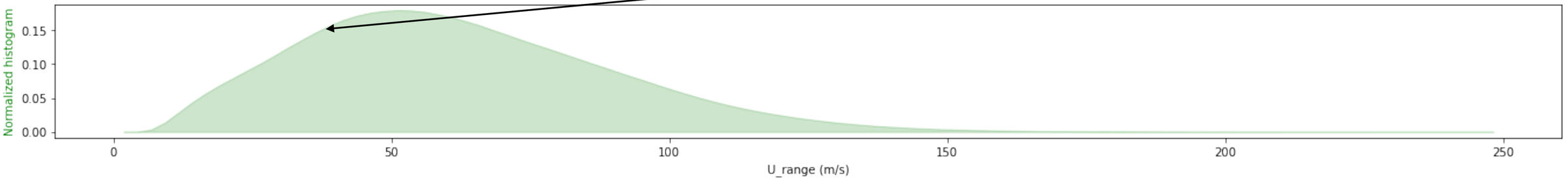
  Given l^0, the level where gravity waves are launched and at level $l*$,

  - the maximum range of zonal wind is computed via

    $$\max_{k,k'} |u_k - u_{k'}|, \quad k, k' \in [l^0, \ldots, , l*]$$

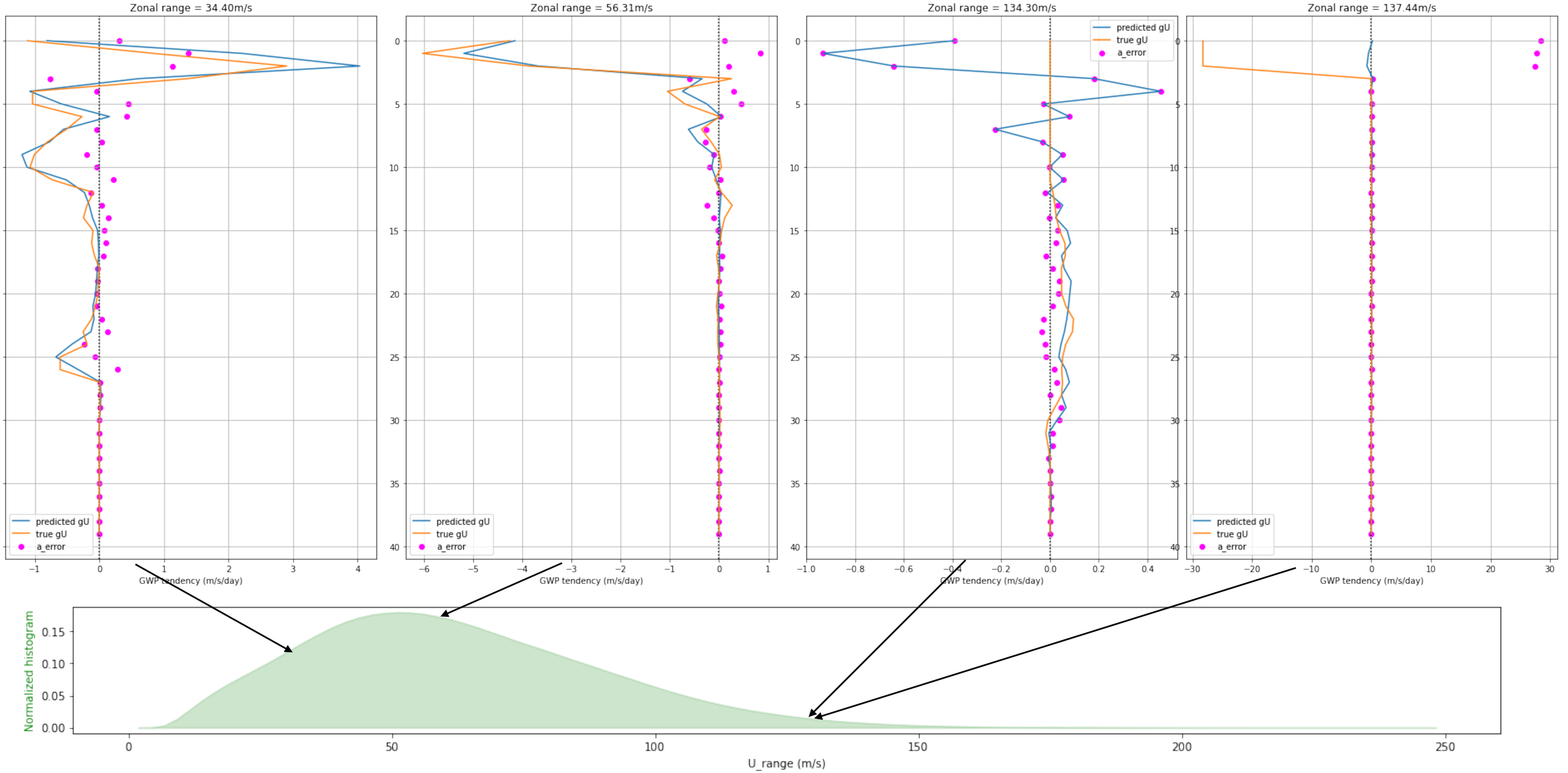
- MiMA dataset yields long-tail distributions this metric.





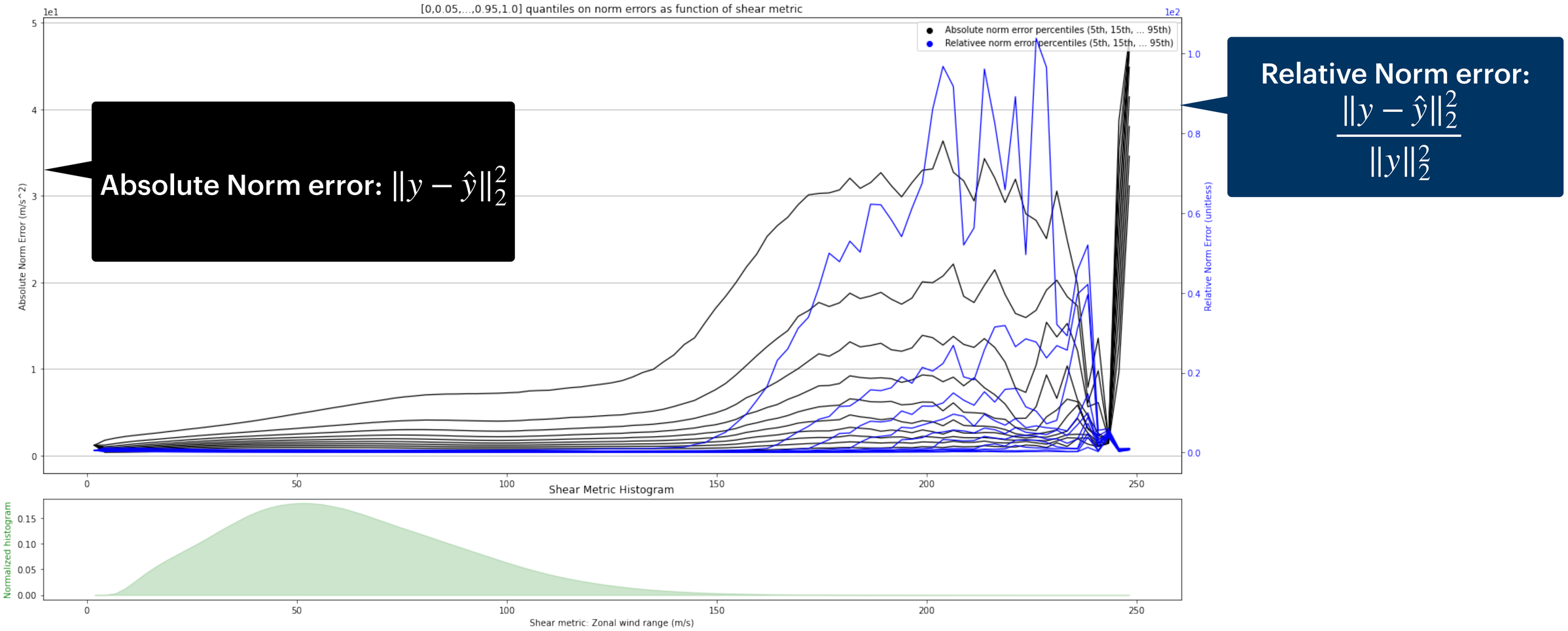
Biased 1D metric: shear-related

GWP profile examples



Shear

- High shears indicate multiple critical level [rare occurrence & difficult task]
- **GOAL: Eliminate error dependence on 1D shear metric**



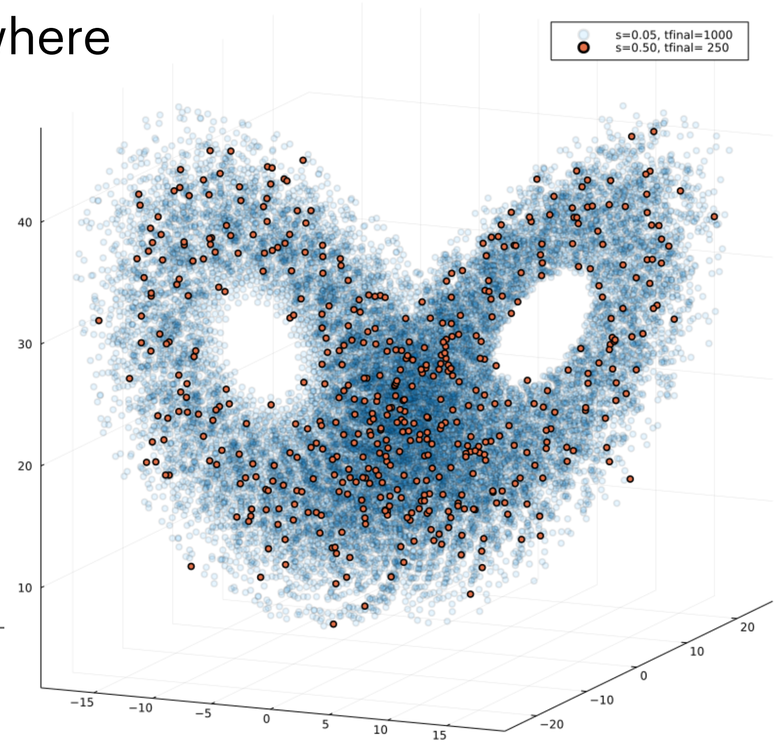
Method Description

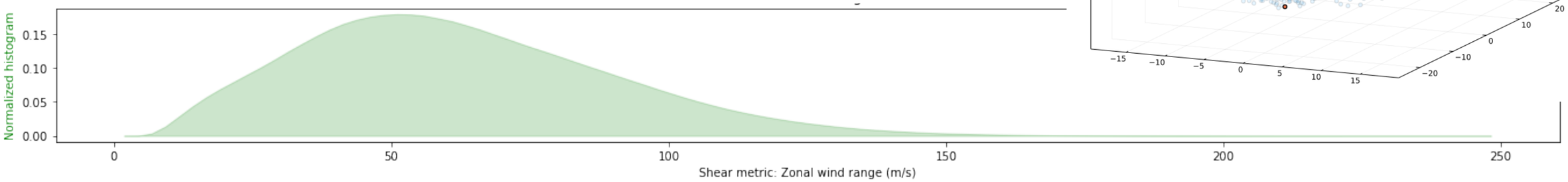
Adjust distribution of data along this metric

- Is it safe to do that? Learning theory tells us sample distribution should be an accurate representation of true distribution (empirical risk minimization).
  - Find $h^* = \arg\min_{h \in \mathcal{H}} R(h)$, the hypothesis that minimizes risk, where

    R(h) = Risk(hypothesis) = $\mathbb{E}[L(h(x), y)] = \int L(h(x), y) \mathrm{d}P(x, y)$

- Are some regions of the phase space / manifold more densely populated or more densely represented?
- Maybe some samples are redundant w.r.t. the interpolating/ generalization capability of ML model.





Histogram Equalization

- Data-based approach: modify original distribution (S_φ) to be closer to the uniform distribution [Peyré & Cuturi 2019]
- Let x_i represent the original quantile of a sample on S_φ and let $y_{\sigma_i^{-1}}$ represent its quantile on the new (uniform) distribution.
  - This can be parameterized for $t \in [0,1]$ via $(1-t)x_i + t y_{\sigma_1^{-1}(i)}$.

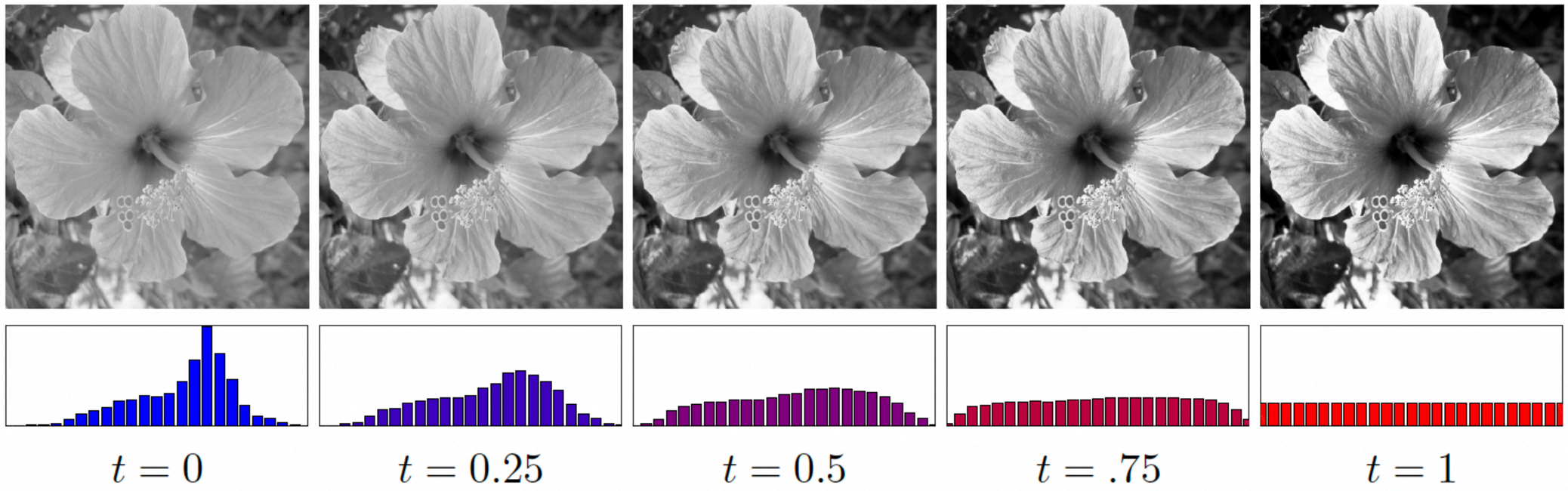


Figure 2.10: Histogram equalization for image processing, where t parameterizes the displacement interpolation between the histograms.

2x2 example

- Original image pixel value: $\begin{bmatrix} x_1 & x_2 \\ x_3 & x_4 \end{bmatrix} = \begin{bmatrix} 0.60 & 0.52 \\ 0.25 & 0.44 \end{bmatrix}$.
- The sorting permutation is $\sigma = [3,4,2,1]$ for a row-wise uncoiling of the matrix.
- Equalized image pixel values coiled: $\begin{bmatrix} y_{\sigma^{-1}(1)=4} & y_{\sigma^{-1}(2)=3} \\ y_{\sigma^{-1}(3)=1} & y_{\sigma^{-1}(4)=2} \end{bmatrix} = \begin{bmatrix} 1 & 2/3 \\ 0 & 1/3 \end{bmatrix}$
- End result: $(1-t)\begin{bmatrix} x_1 & x_2 \\ x_3 & x_4 \end{bmatrix} + t\begin{bmatrix} y_4 & y_3 \\ y_1 & y_2 \end{bmatrix} = (1-t)\begin{bmatrix} 0.60 & 0.52 \\ 0.25 & 0.44 \end{bmatrix} + t\begin{bmatrix} 1 & 2/3 \\ 0 & 1/3 \end{bmatrix}$.

Apply to our problem

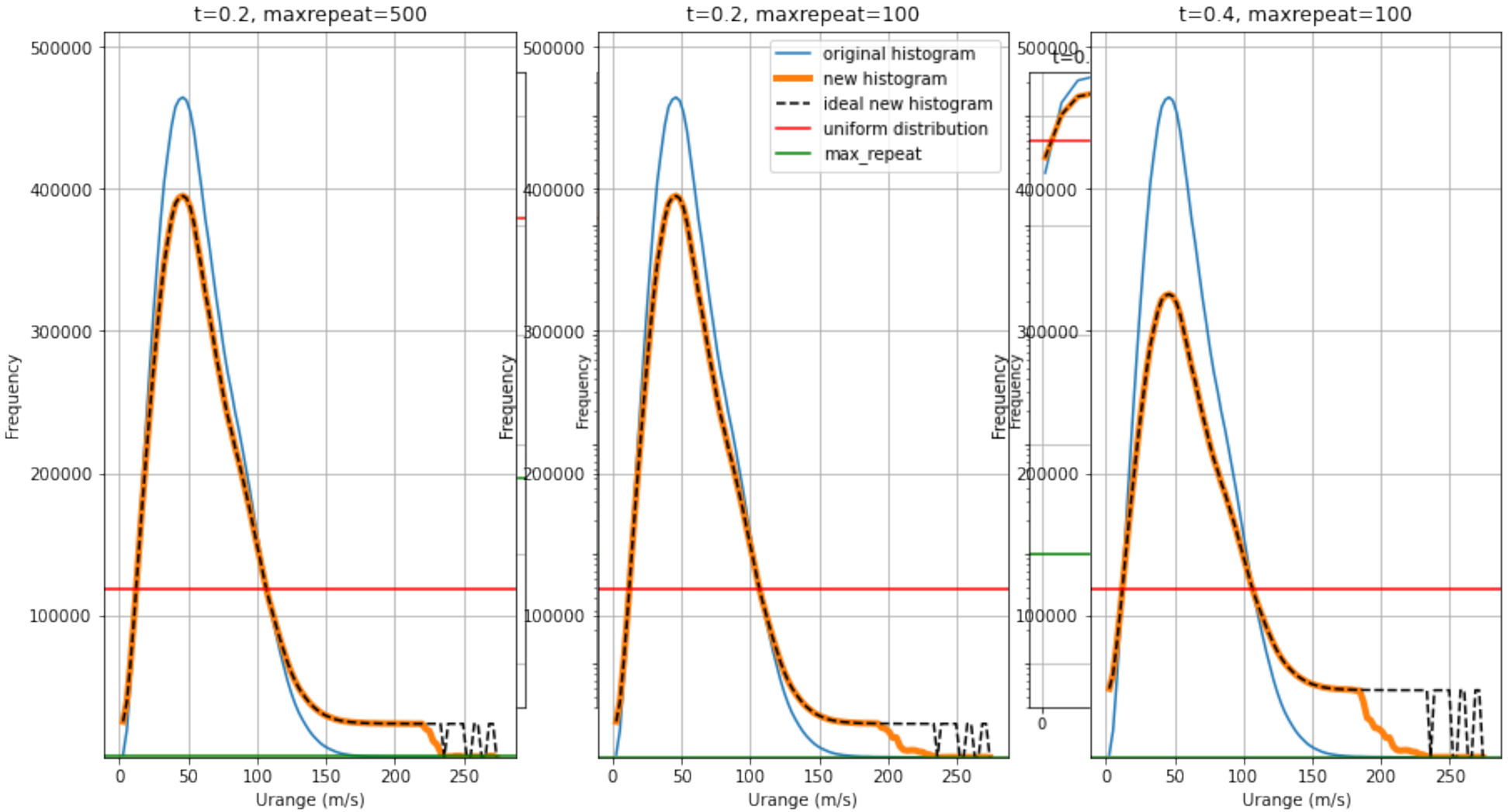
- Given the n^{th} bins of the original and the ideal histograms ($h_n^{(0)}$ & $h_n^{(f)}$), and $t \in [0,1]$ to parameterize the linear interpolation between the two histograms,
- the new bin for the desired histogram is: $h_n^{(t)} = (1-t)h_n^{(0)} + th_n^{(f)}$.
- In practice, we sample from the original bin with probability

$$\alpha_n^{(t)} := \frac{h_n^{(t)}}{h_n^{(0)}} = \frac{\text{desired bin count}}{\text{original bin count}}$$

- We limit the maximal value that α_t because this value can easily be 1e5 or larger for the far end of the tail. (i.e. $\alpha_t \in [0,\text{maxrepeat}]$ instead of $[0,\infty)$)

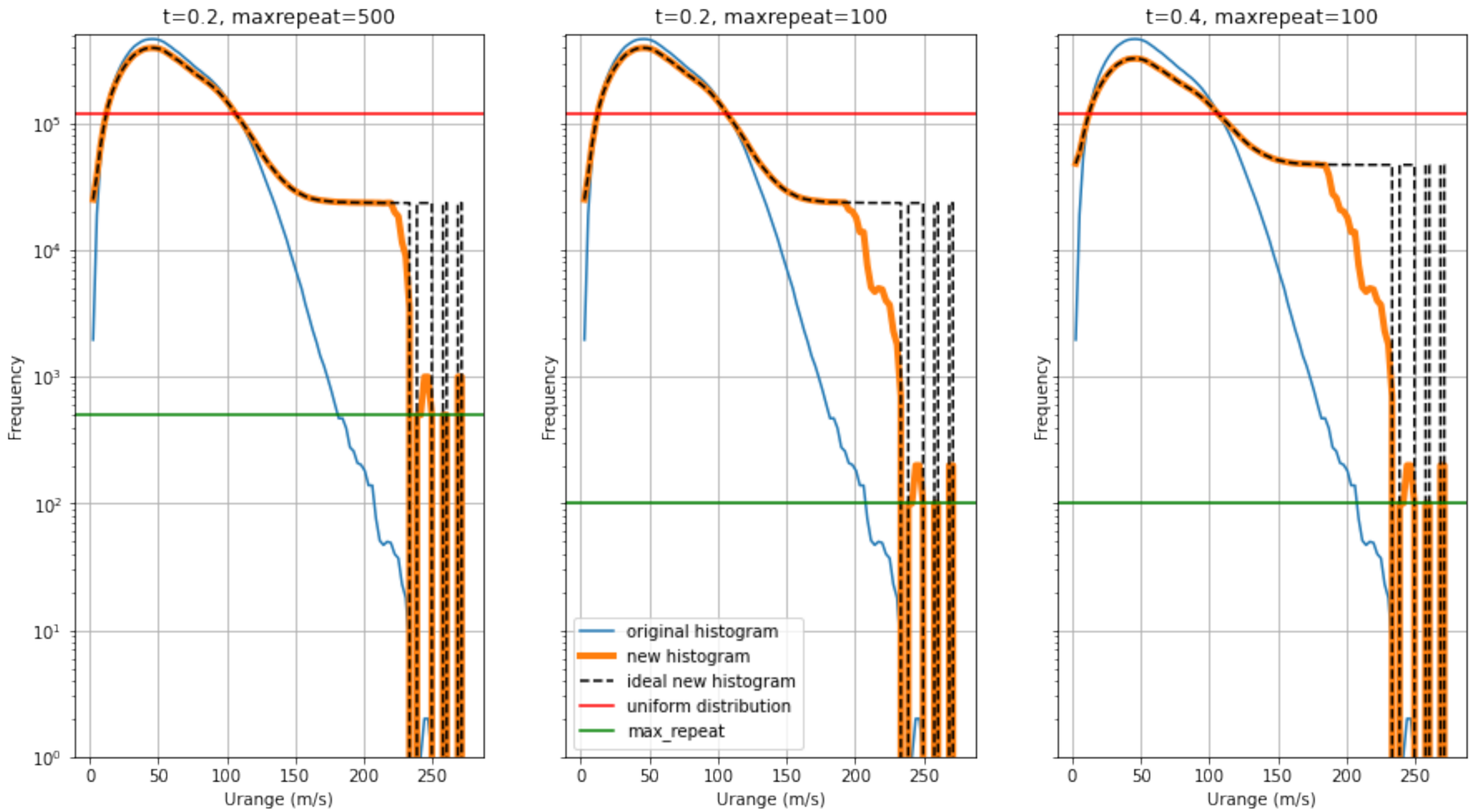
Examples of new histograms

- nbins: number of bins.
- t: recall that t=0: original distribution; t=1: uniform distribution.
- maxrepeat: this is the cap we're putting on α_t.



Examples of new histograms

- nbins: number of bins.
- t: recall that t=0: original distribution; t=1: uniform distribution.
- maxrepeat: this is the cap we're putting on α_t.



Implementation

- Direct sampling: Given t, nbins, and maxrepeat

  1. For each of the bins (for n = 1 .... nbins)

  2.Sample with probability $\min(\alpha_n^{(t)}, \text{maxrepeat})$.

  3.Repeat when learning algorithm saw all of the data in this instance of sampling.

- Weighted loss: Given t, bins, and maxrepeat,
  - Find which bin a sample belongs to, and assign $\min(\alpha_n^{(t)}, \text{maxrepeat})$ as weight for its contribution to the loss in that batch.
- Weighted loss method performs the expected value of direct sampling => less variability.

Bias removal

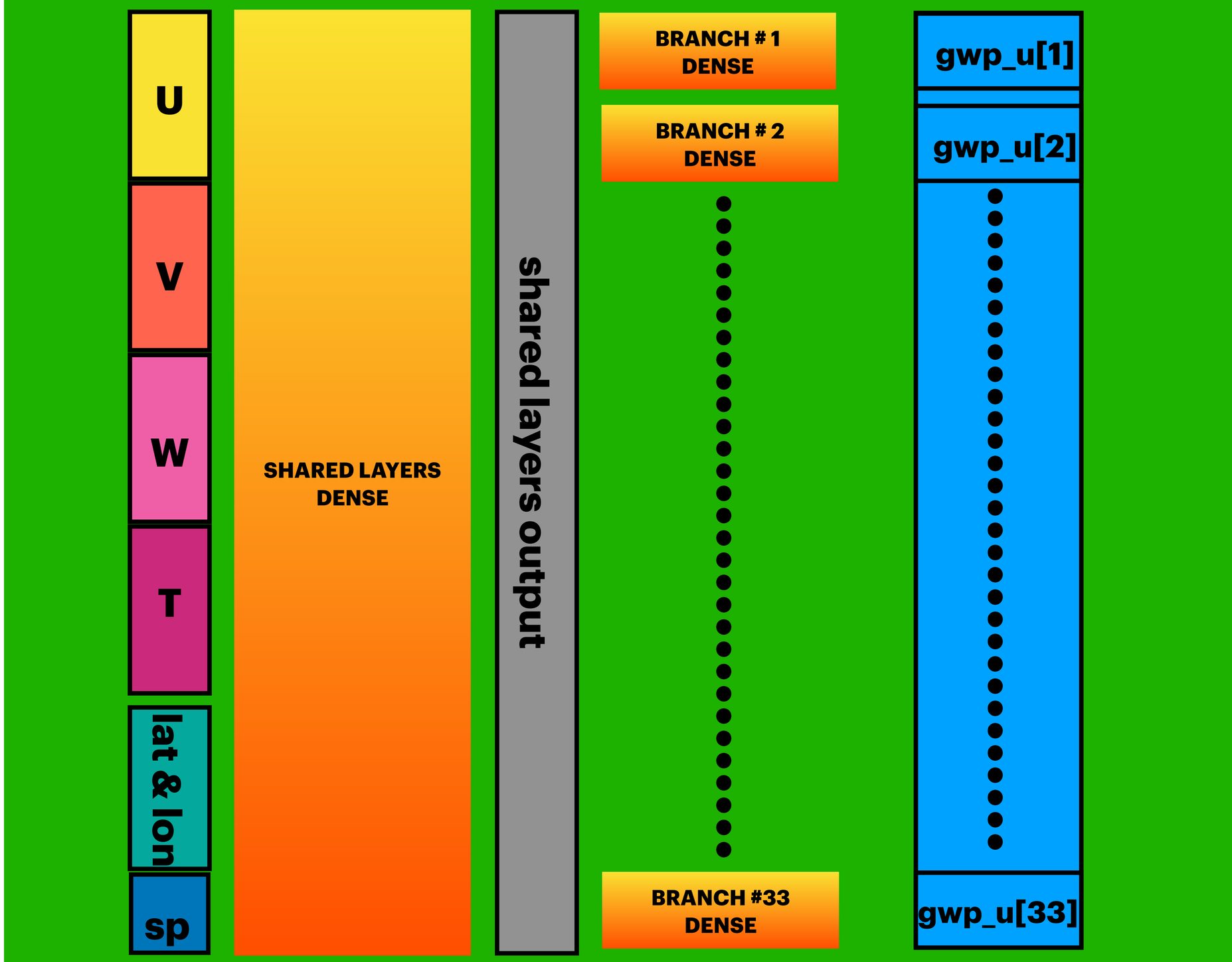
- After training, compute average of componentwise errors for each bin.
- When coupling, remove mean bias profile of the bin each sample belongs to.

Experiment

ML Architecture

WaveNet
[Espinosa 2022]

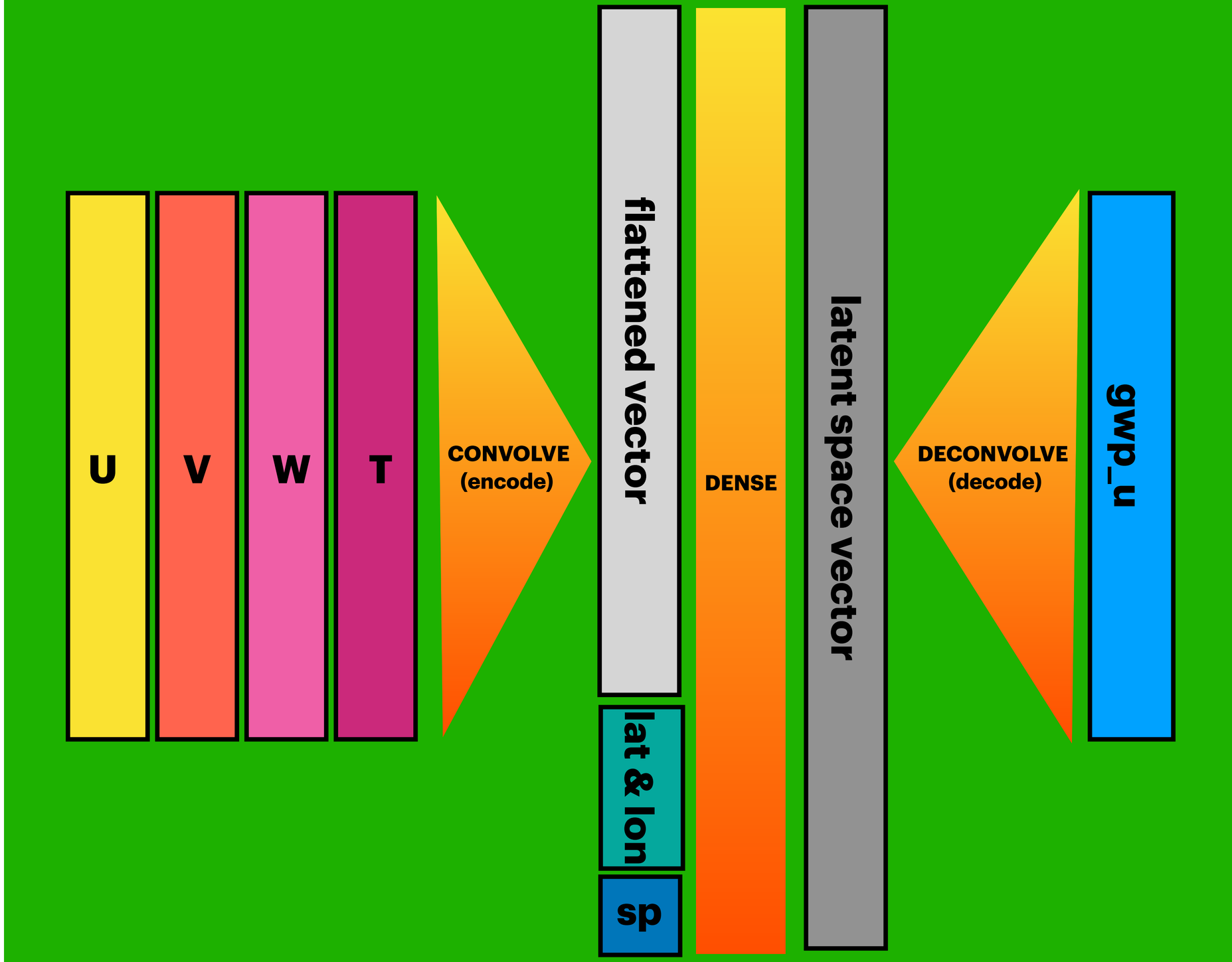
- *Shared layers => global interactions*
- *Branches => specific to each output layer*
- *Small version: 350_000*
- *Large version: 700_000*



ML Architecture

Encoder-Dense-Decoder

- *convolve & deconvolve => local interactions*
- *dense => global interactions*
- *Small version: 350_000*
- *Large version: 700_000*

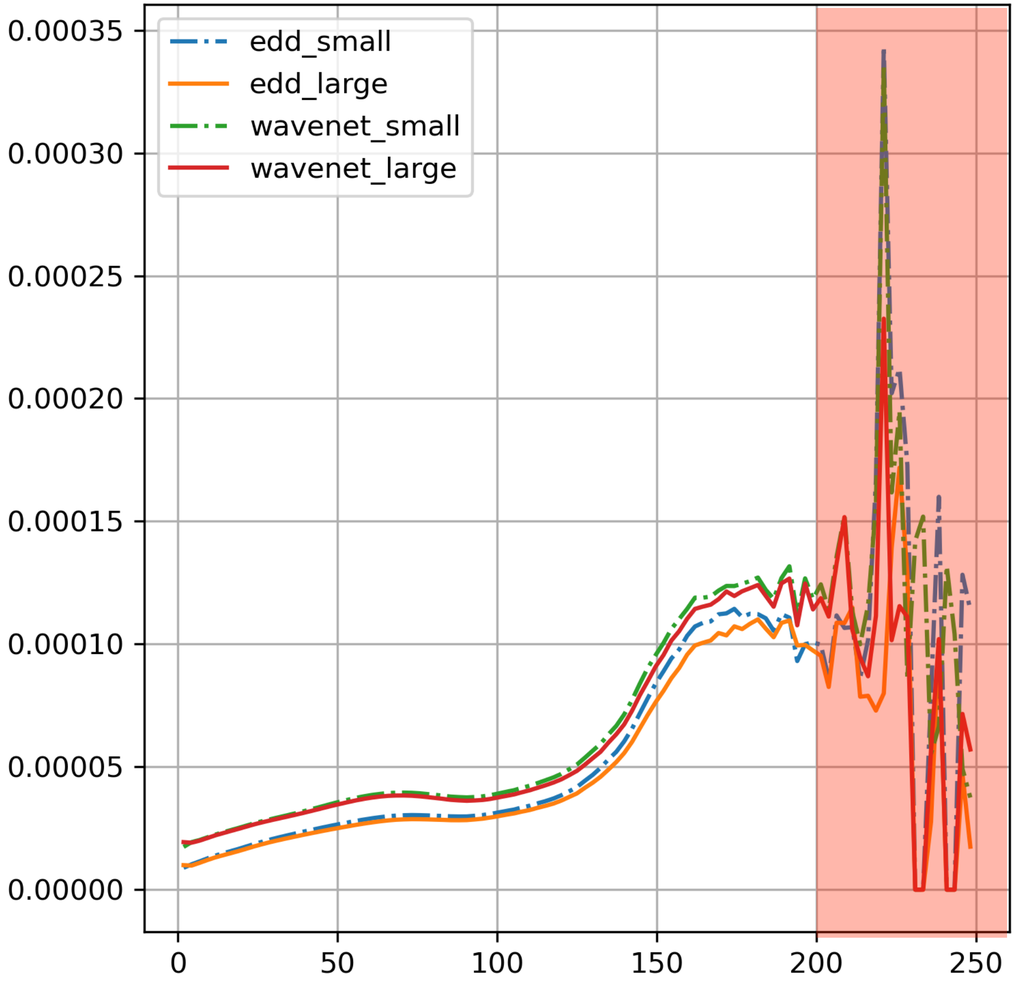


Offline results

Model comparisons

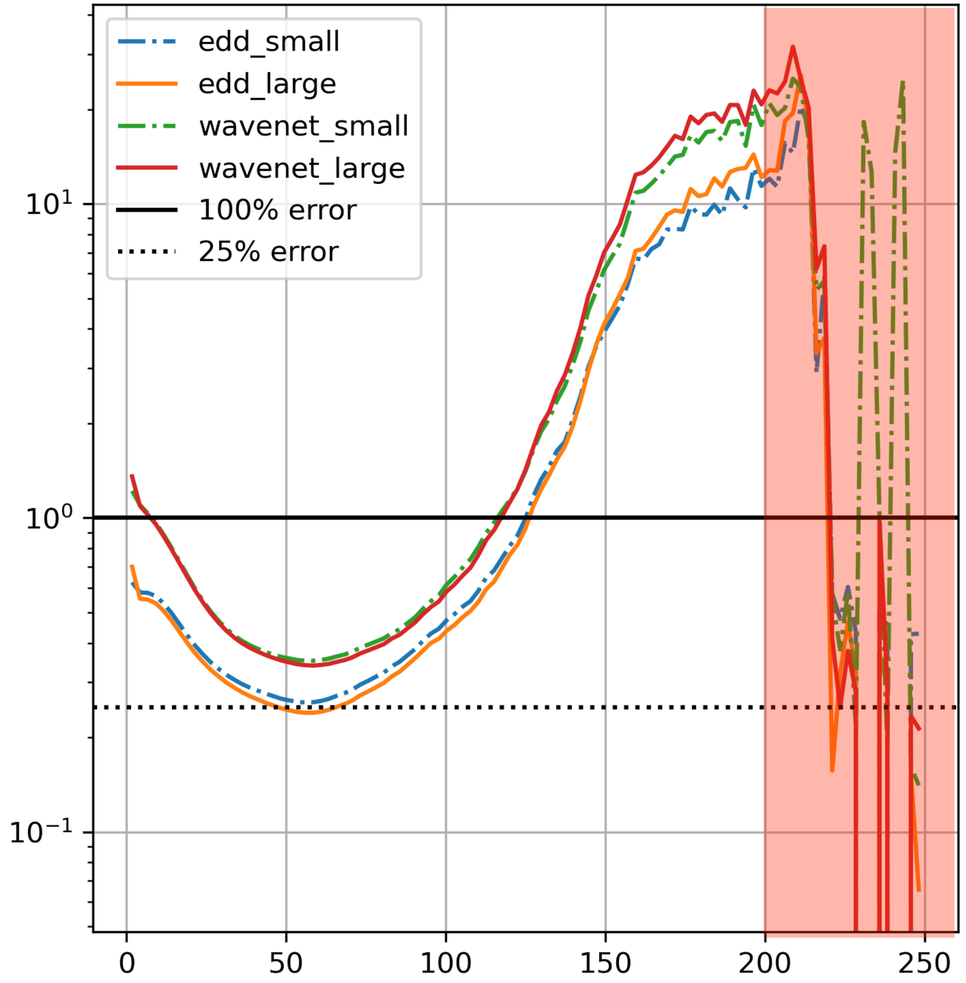
$$\mathbf{ANE}(\mathbf{y}, \hat{\mathbf{y}}) = \|\mathbf{y} - \hat{\mathbf{y}}\|_2$$

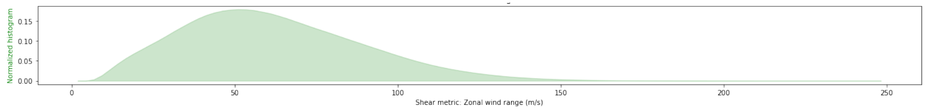
Absolute norm error

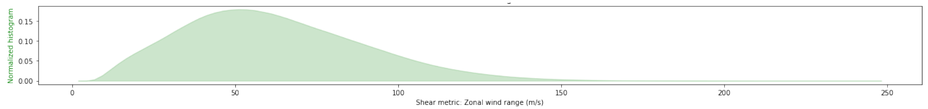


$$\mathbf{RNE}(\mathbf{y}, \hat{\mathbf{y}}) = \frac{\|y - \hat{y}\|_2^2}{\|y\|_2^2}$$

Relative norm error





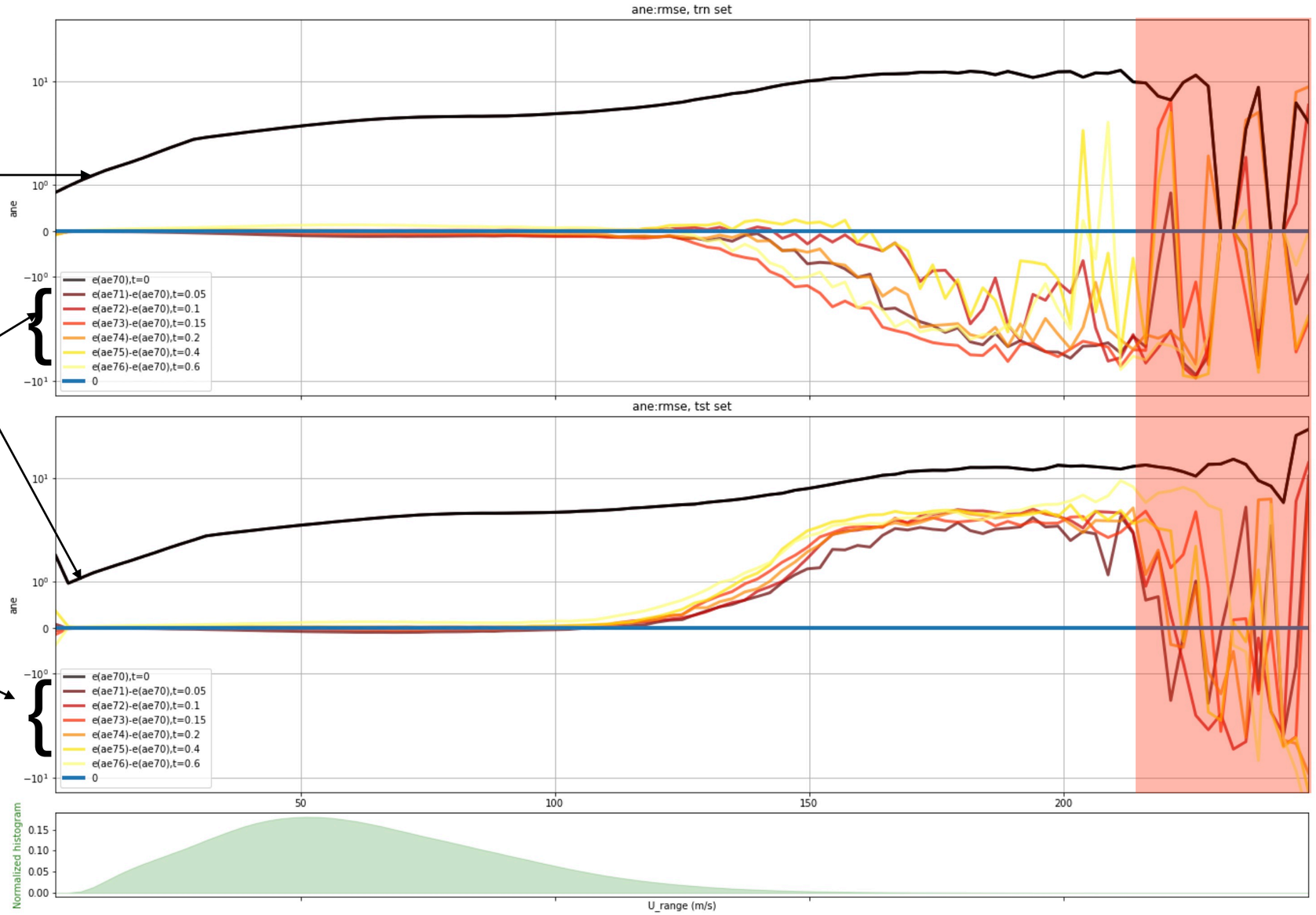


Overfitting enhanced in large models

Benchmark RMSE
(no sampling strategy, t=0)

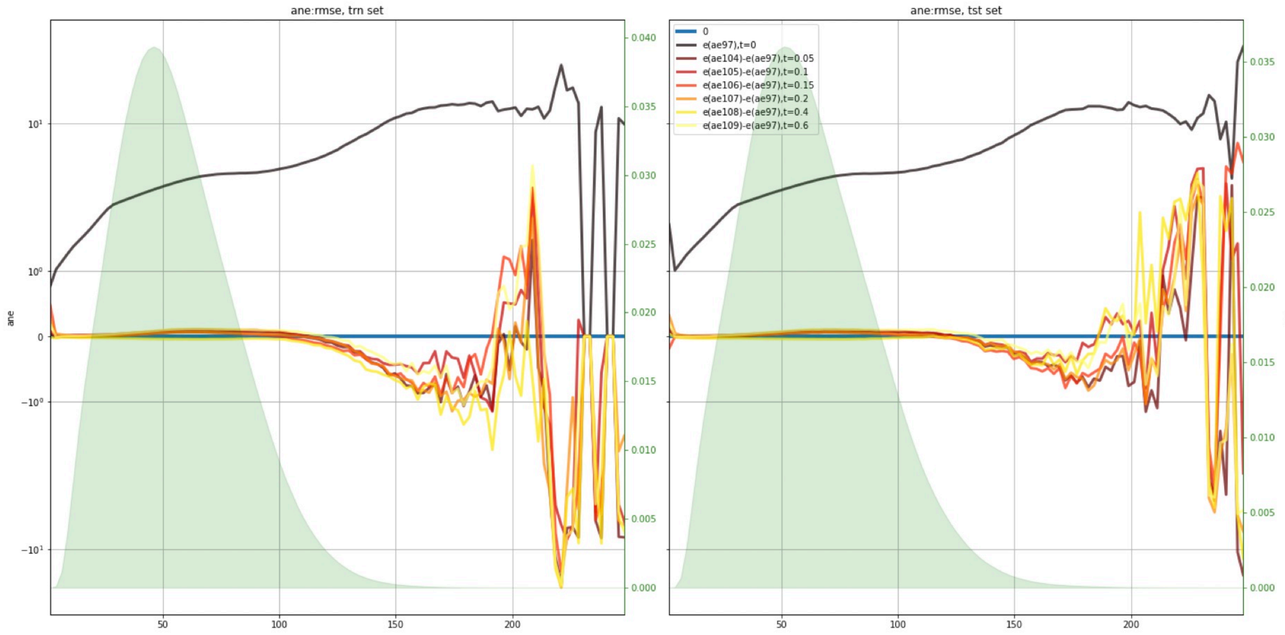
Deviations from benchmark RMSE
(t>0)

- large EDD, direct sampling
- Benchmark RMSE doesn't overfit
- Eliminate large model



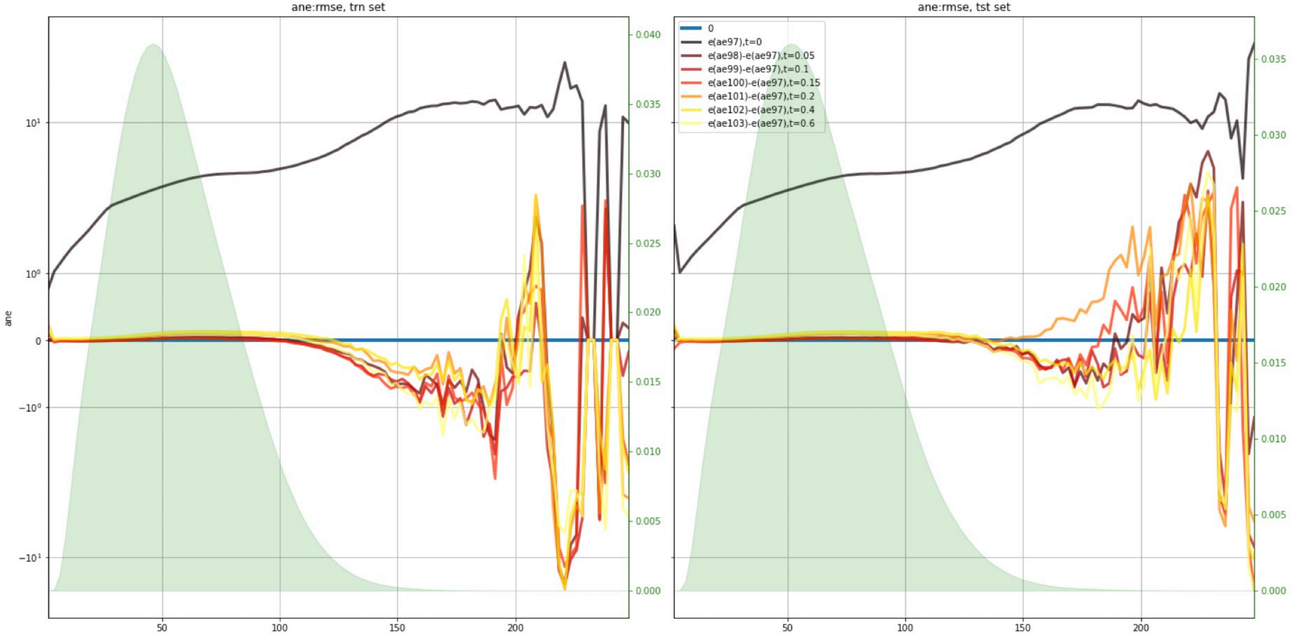
Comparison of implementation of sampling strategy for EDD small

Direct sampling



- Moderate t values yield least difference between training & test.
- All yield improvement in moderate tail region.

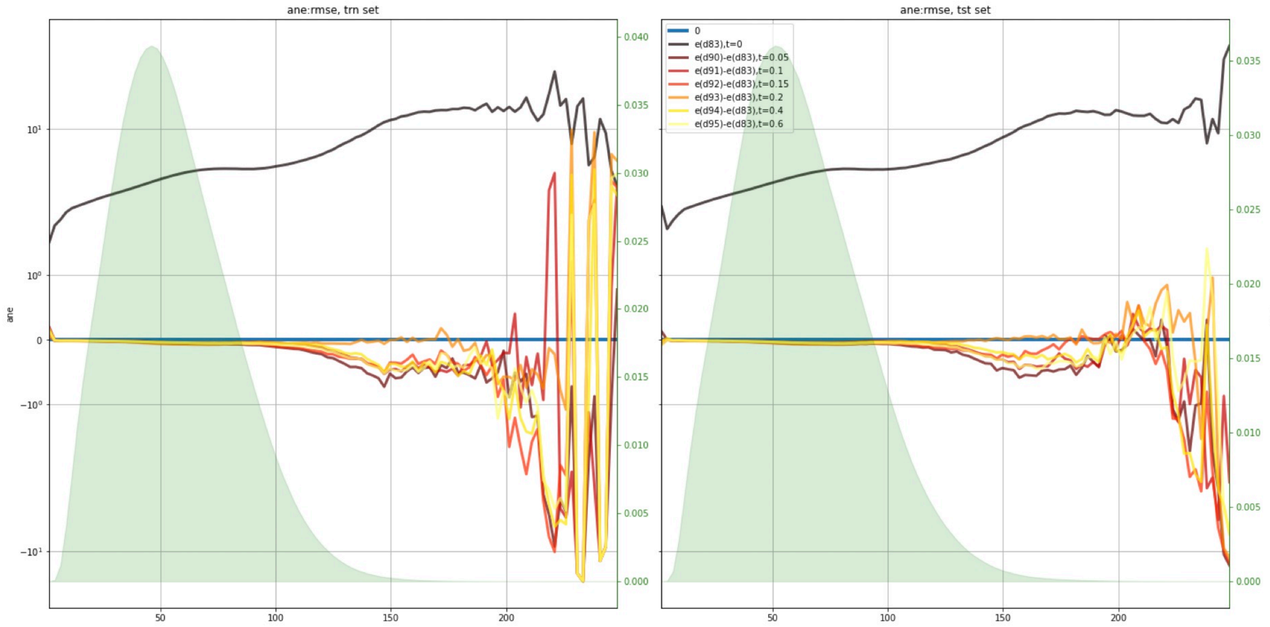
Weighted loss



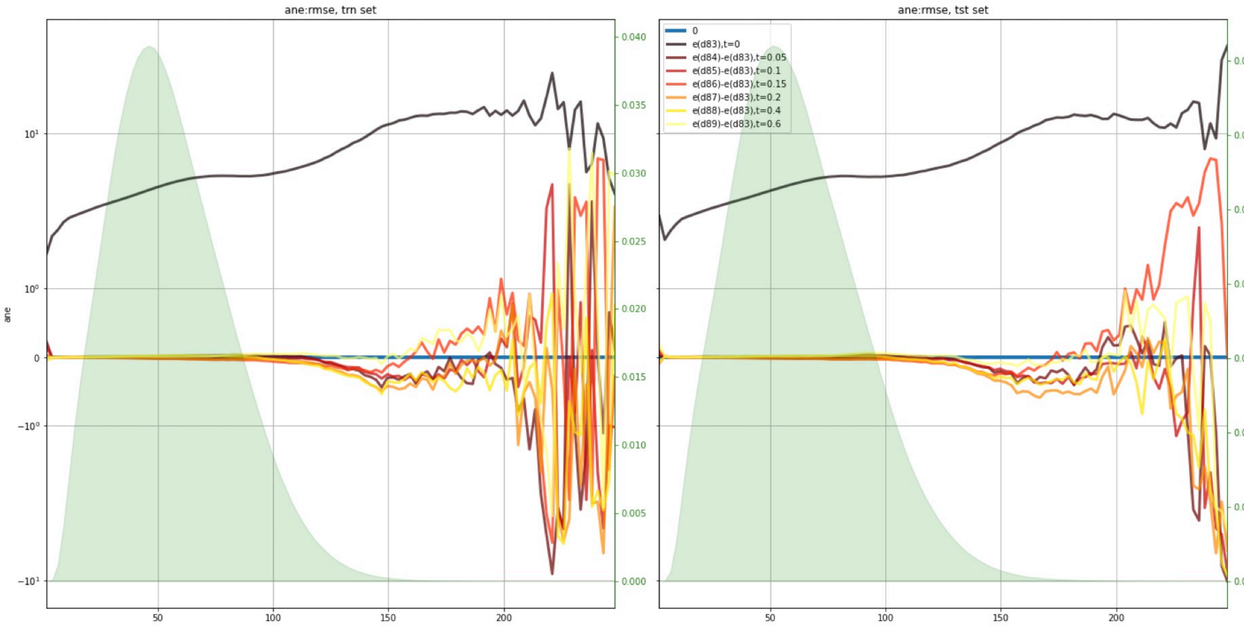
- Extreme t values yield least difference between training & test .
- All but one yield improvement in moderate tail region.

Comparison of implementation of sampling strategy for WaveNet small

Direct sampling



Weighted loss



- Hard to distinguish between two methods
- Noisier training errors at extreme tail
- Less gains than for EDD small

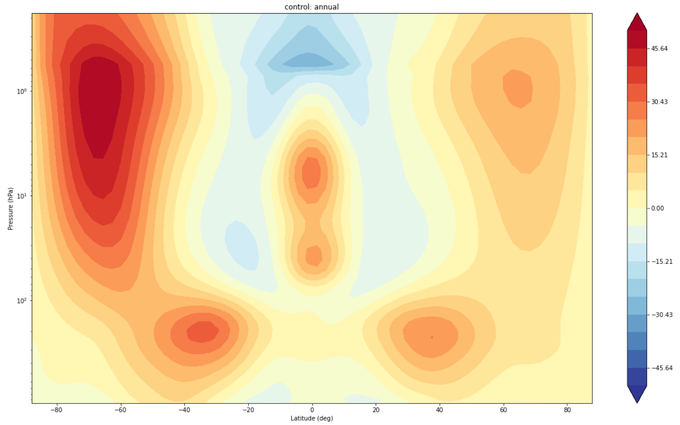
Online results

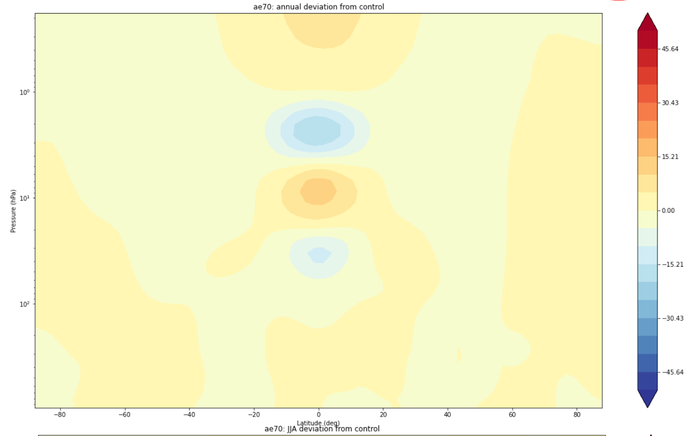
Climatology

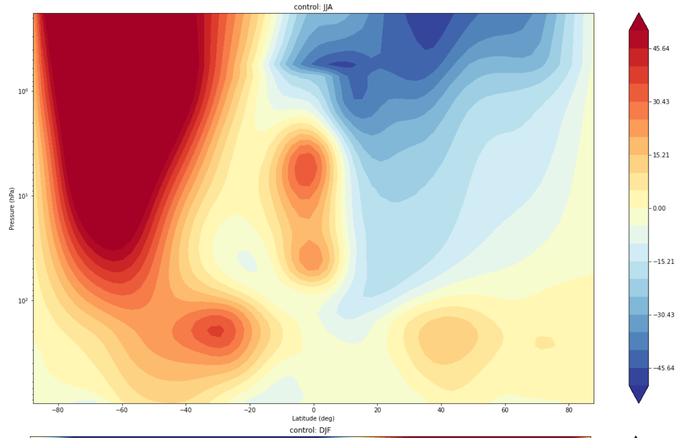
Control Run

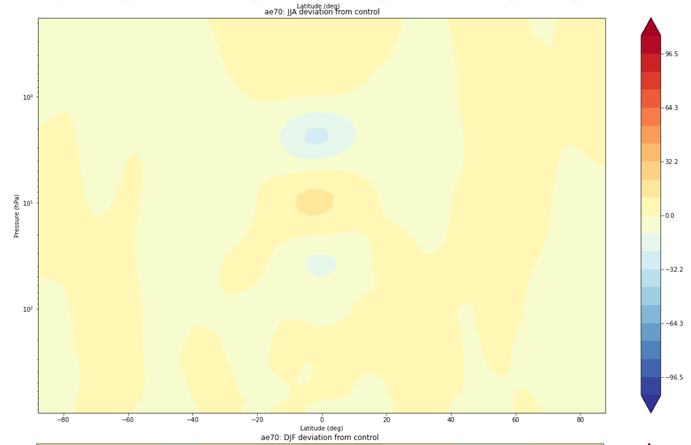
Large EDD

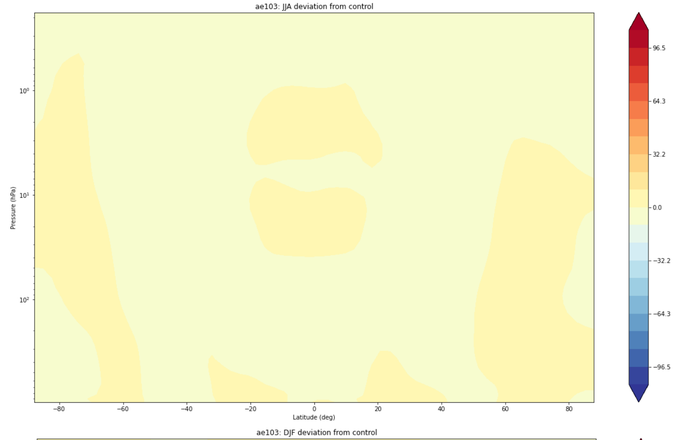
Small EDD

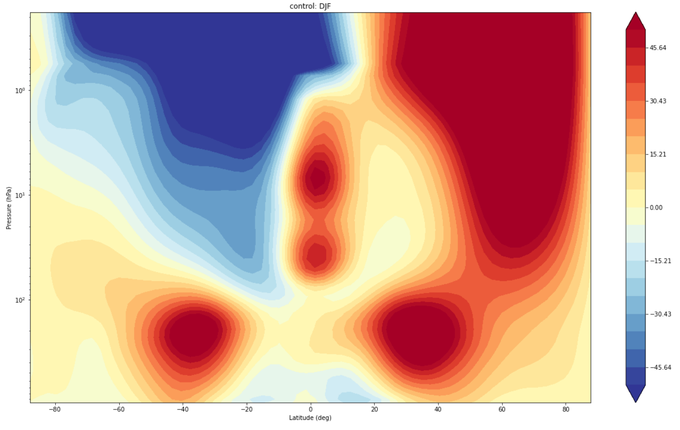


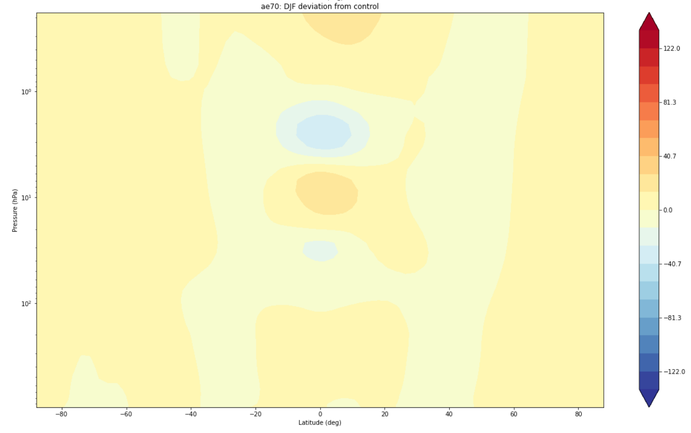


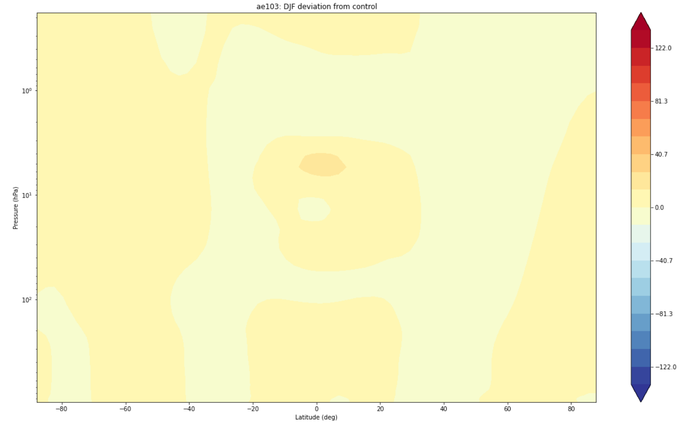








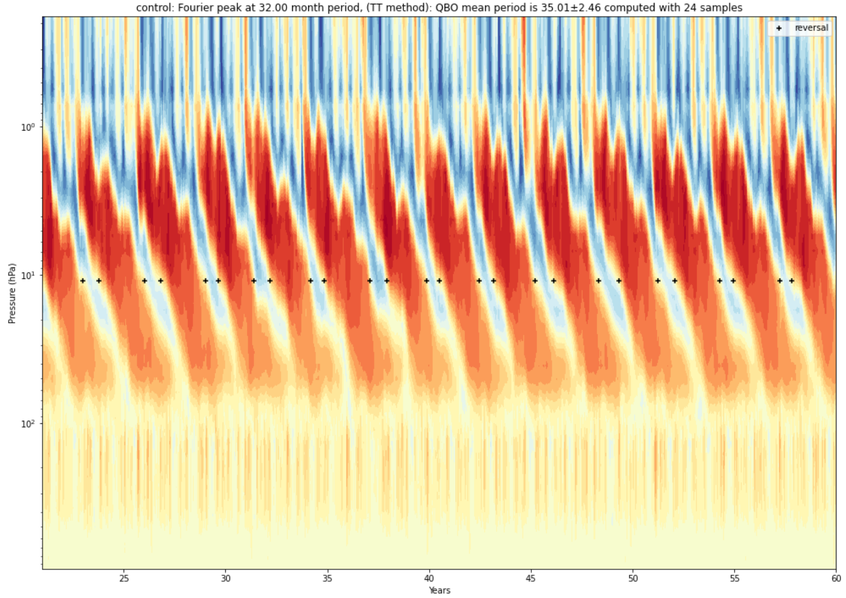


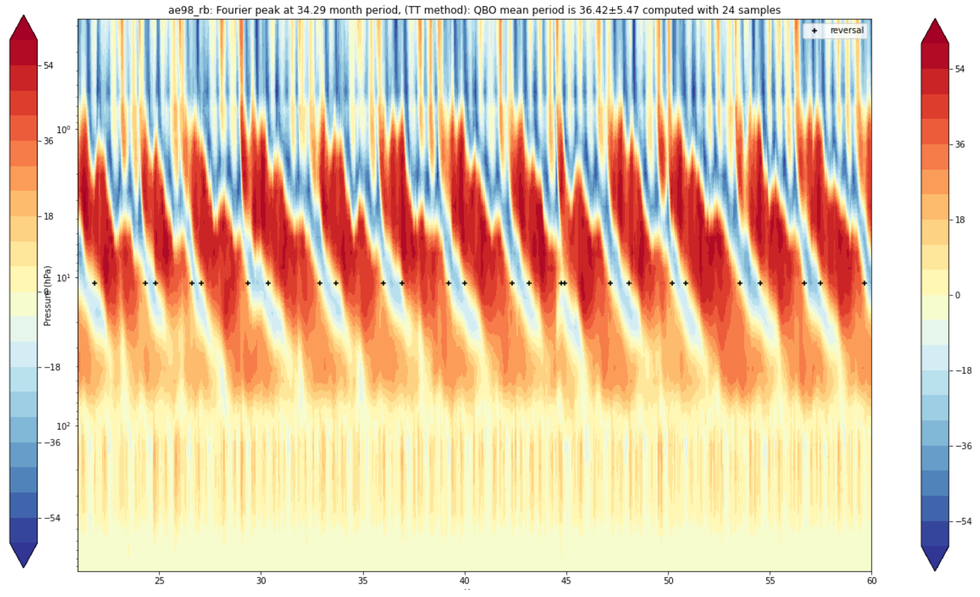


QBO

Emulator Description	Fourier	Transition Time
control	32	35.01 ± 2.46
large EDD, $t = 0$	34.29	49.99 ± 13.73
large EDD, $t = 0$, bias removed	34.29	40.81 ± 5.20
large EDD, $t = 0.05$, direct sampling	40	39.61 ± 6.99
large EDD, $t = 0.05$, bias removed, direct sampling	40	41.69 ± 5.93
large EDD, $t = 0.60$, weighted loss	53.33	60.11 ± 3.94
large EDD, $t = 0.60$, bias removed, weighted loss	53.33	59.62 ± 7.21
small EDD, $t = 0$	36.92	38.37 ± 6.59
small EDD, $t = 0$, bias removed	36.92	37.01 ± 7.70
small EDD, $t = 0.05$, direct sampling	34.29	37.05 ± 2.98
small EDD, $t = 0.05$, bias removed, direct sampling	34.29	38.12 ± 3.69
small EDD, $t = 0.05$, weighted loss	34.29	39.66 ± 8.97
small EDD, $t = 0.05$, bias removed, weighted loss	34.29	36.42 ± 5.47

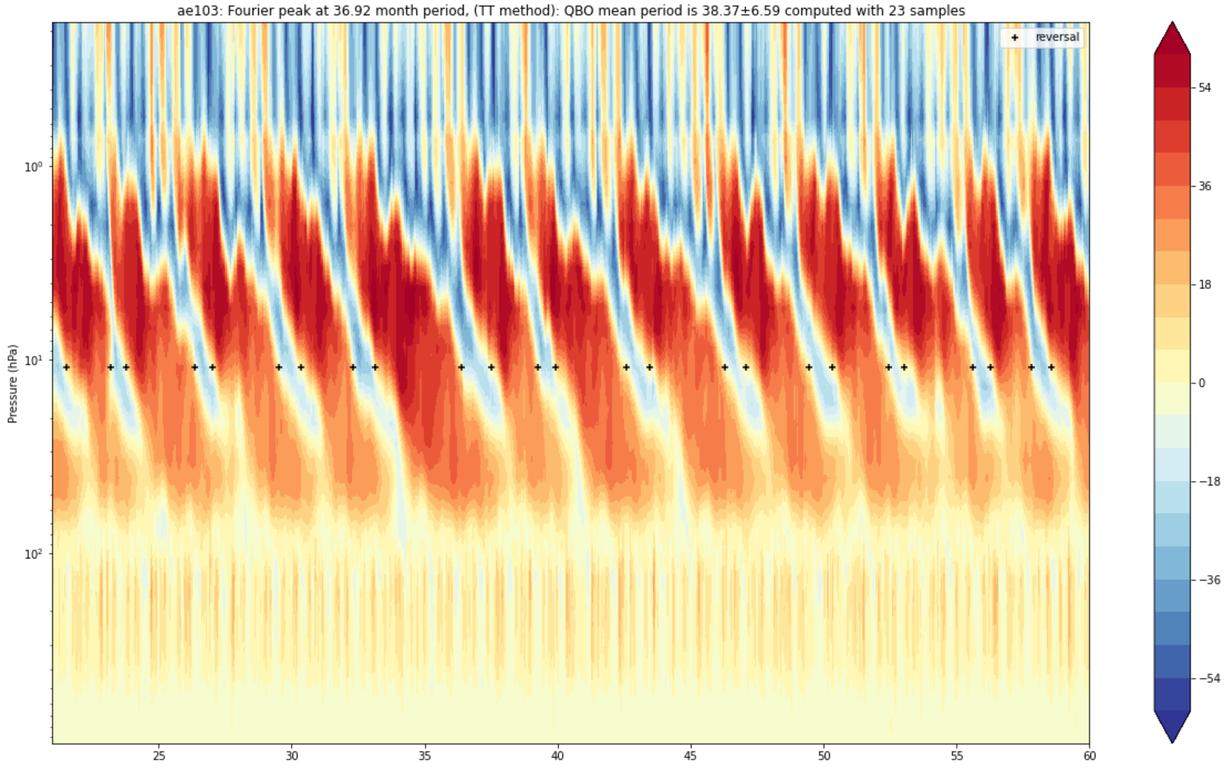
- Zonal mean zonal wind in tropics
- Fourier method picks out zonal mode w/ largest coefficient around 10hPa
- Transition Time averages length between phase transitions at around 10hPa
- Large models yield QBO periods far from control run QBO
- Small models yield QBO periods within sample error.



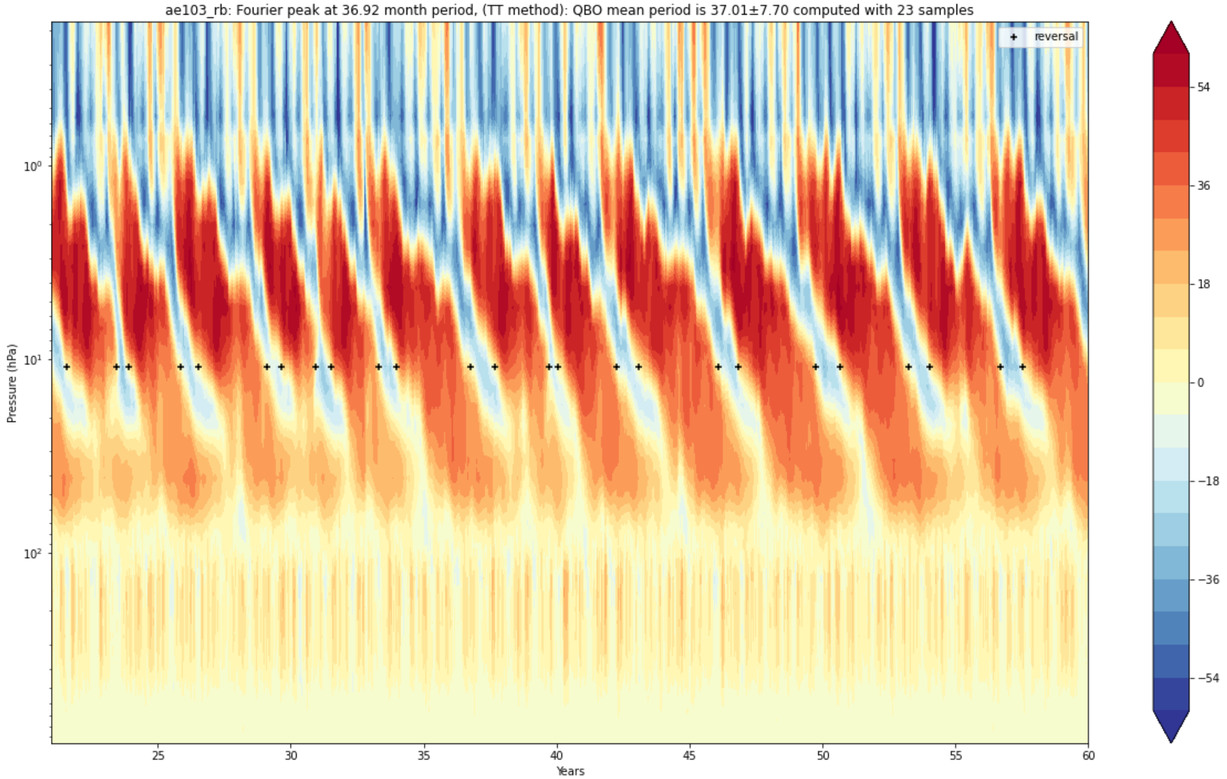


Bias removal Ex #1: EDD small

Bias not removed

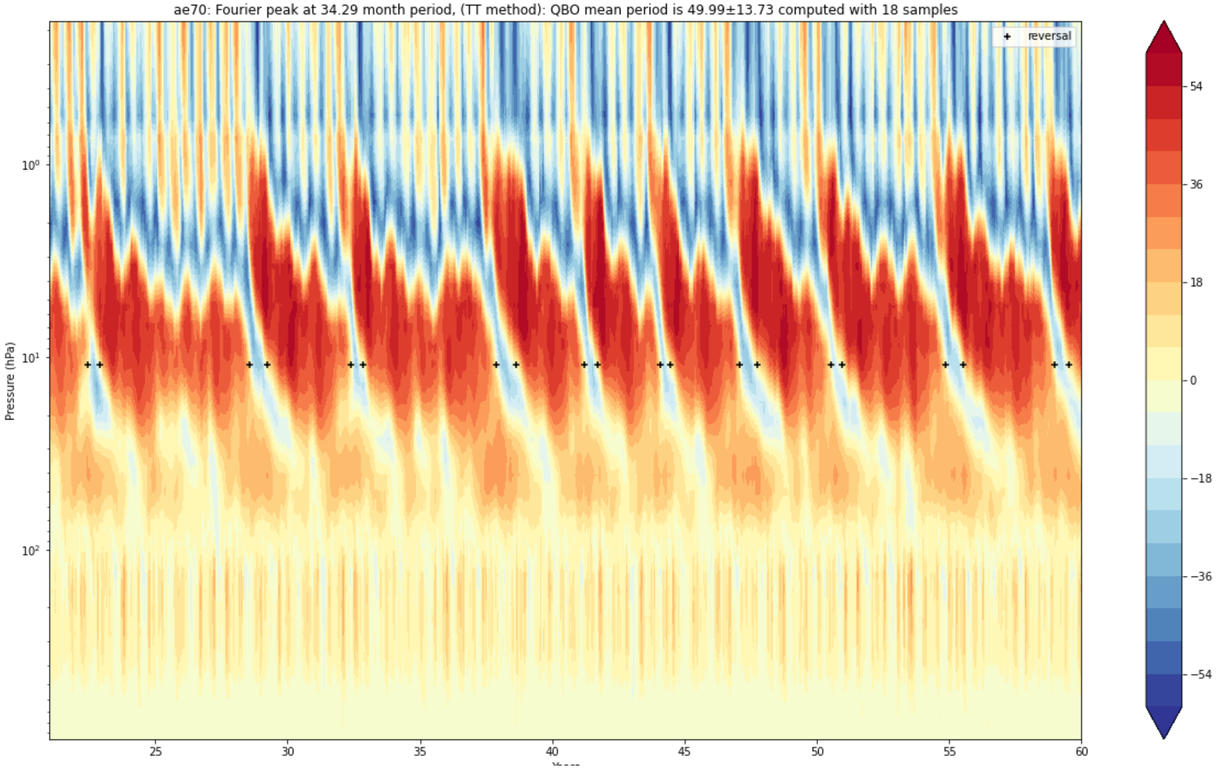


Bias removed

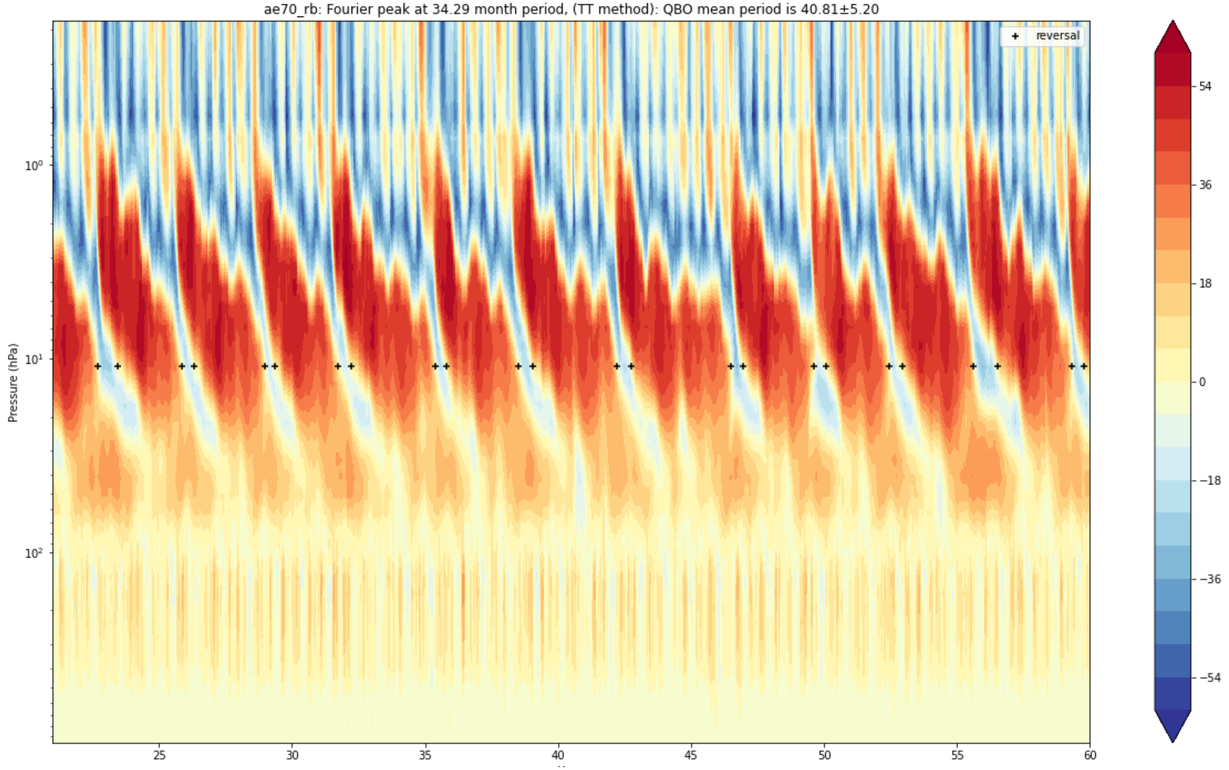


Bias removal Ex #2: EDD large

Bias not removed



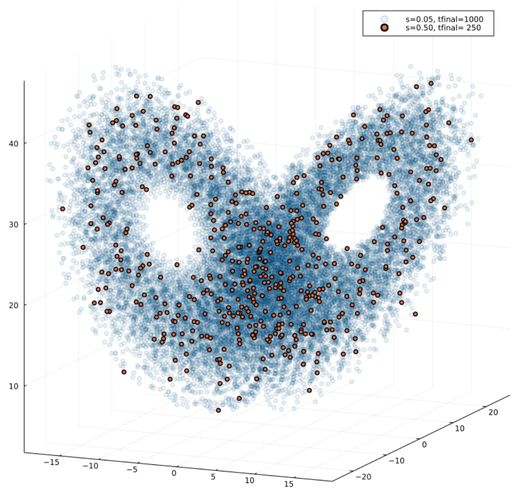
Bias removed



Conclusions

- This work is similar to treating imbalanced datasets in regression (instead of classification).
- At a minimal loss to the samples in the peak portion of the distribution, we achieve 5-10% improvements in the first ~half of the tail. The extreme end of the tail may be a lost cause.

- When using ML on data from dynamical systems, it may be important to consider the distribution of data over the phase space.
    - We ended up with this shear related metric because of prior knowledge.
    - Can we find another way to identify regions of the phase space that are not densely sampled?



References

1. Alexander, M. J., & Dunkerton, T. J. (1999). A Spectral Parameterization of Mean-Flow Forcing due to Breaking Gravity Waves. *Journal of the Atmospheric Sciences*, *56*(24), 4167–4182. https://doi.org/10.1175/1520-0469(1999)056<4167:ASPOMF>2.0.CO;2
2. Espinosa, Z. I., A. Sheshadri, G. R. Cain, E. P. Gerber, and K. J. DallaSanta, 2021: A Deep Learning Parameterization of Gravity Wave Drag Coupled to an Atmospheric Global Climate Model, Geophys. Res. Lett., accepted.
3. Peyré, G., & Cuturi, M. (2019). Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, *11*(5-6), 355-607.