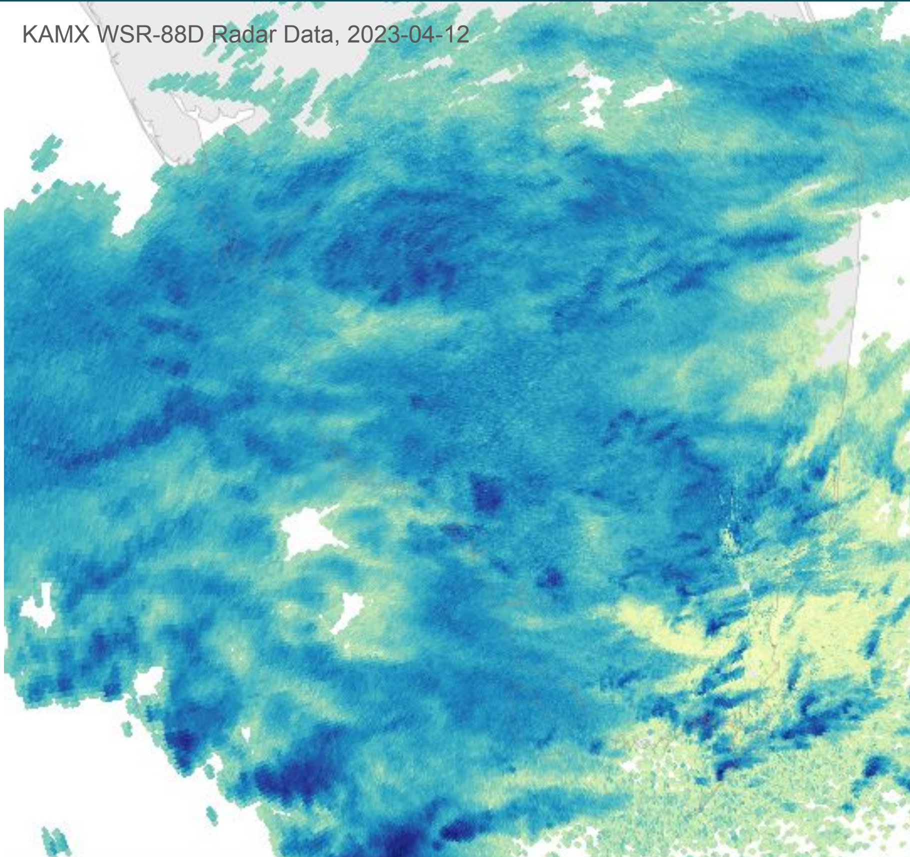
Accurate and Interpretable Radar Quantitative Precipitation Estimation with Symbolic Regression

Brianna Grissom[1], Jonathan He[2], Kenia Munoz-Ordaz[3], Julian Pulido[4], Olivia Zhang[5, 6], Mostafa Cham[7], Haotong Jing[6], Weikang Qian[6], Yixin Wen[6], Jianwu Wang[7]

[1] Department of Statistics and Applied Probability, University of California, Santa Barbara
[2] Atholton High School
[3] School of Computing and Design, California State University, Monterey Bay
[4] Department of Computer Science, California State University, Sacramento
[5] Departments of Statistics, University of Florida
[6] Department of Geography, University of Florida
[7] Department of Information Systems, University of Maryland, Baltimore County

Acknowledgements: NSF Funded Big Data REU Site, UMBC

Contents



1. Introduction

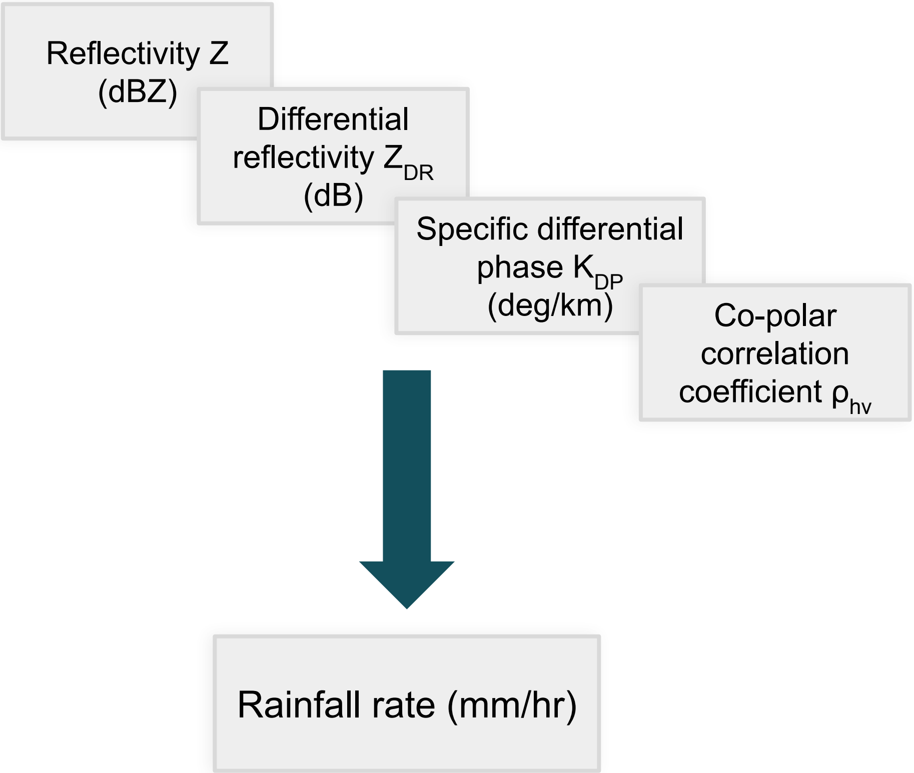
- Accurate estimation of precipitation is crucial for a variety of applications such as extreme weather condition forecasting, flash flood management, and ongoing climate research [9].
- However, quantitative precipitation estimation (QPE) is limited through the following methods:
    - Collecting rainfall from rain gauges has limited spatial coverage.
    - Estimating rainfall from single-polarimetric radar data may fail to account for different precipitation types and intensities [2].
- Our research focuses on improving QPE using dual-polarimetric radar data with symbolic regression.
    - Symbolic regression provides a unique approach by providing interpretable and accurate equations learned from data [1].

2. Background

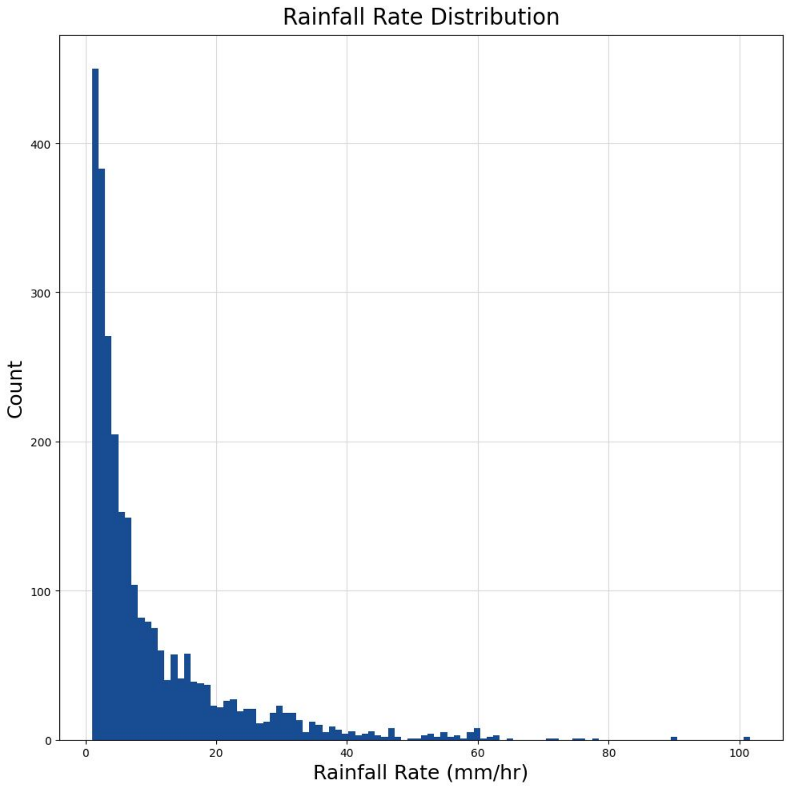
- **Z–R relationships** have been used since 1947 to estimate rainfall rate (R) using reflectivity (Z), and these equations vary slightly based on region and rain type [2].
    - $Z = 300R^{1.4}$ (WSR-88D Convective)
    - $Z = 200R^{1.6}$ (Marshall-Palmer)
- However, the commonly-used Z–R relationships fail to account for nuances in rainfall by precipitation type, region, and season [2].
- **Dual-polarization radar variables** better reflect the size, shape, and orientation of raindrops.
    - Using dual-polarization radar variables as input data, researchers have found that convolutional neural networks [5, 8] and random forest and regression tree methods [7] outperformed conventional Z–R relationships to estimate rainfall rate.
- In a study of deep-learning-based QPE models, rainfall estimates were more accurate when distinguishing rainfall intensity using a K_{DP} threshold [3].

3. Data

- Data from Central Oklahoma (June 8, 2022 and July 9, 2023) and South Florida (April 12, 2023) with significant rainfall:
  - **Dual-polarimetric radar data** from the Weather Surveillance Radar, 1988 Doppler (WSR-88D) at Level II.
  - **Rain gauge data** from the Oklahoma Mesonet and the South Florida Water Management District's DBHYDRO.



3. Data



Reflectivity Z (dBZ)

Differential reflectivity Z_{DR} (dB)

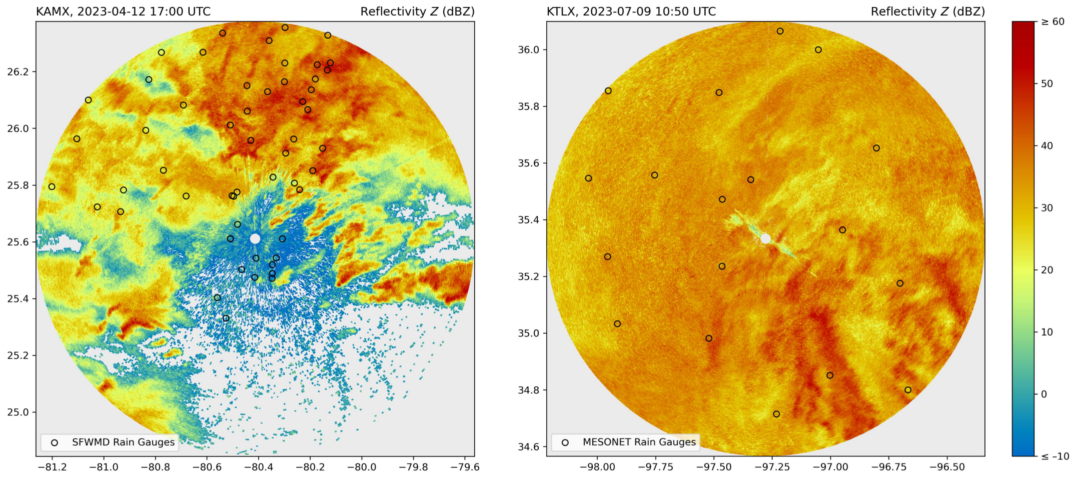
Specific differential phase K_{DP} (deg/km)

Co-polar correlation coefficient ρ_{hv}

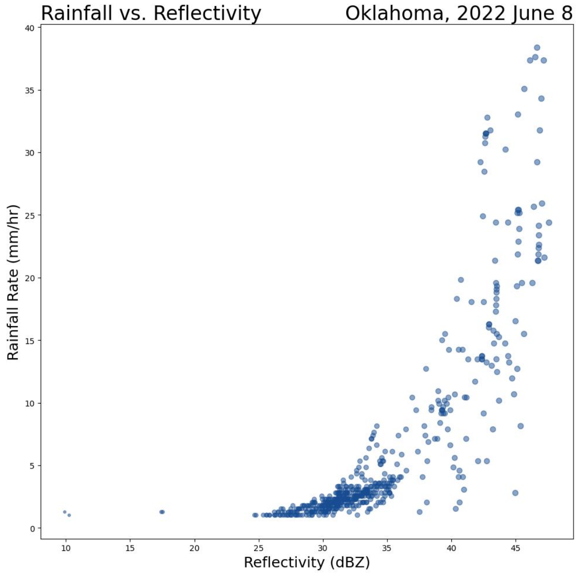
Rainfall rate (mm/hr)

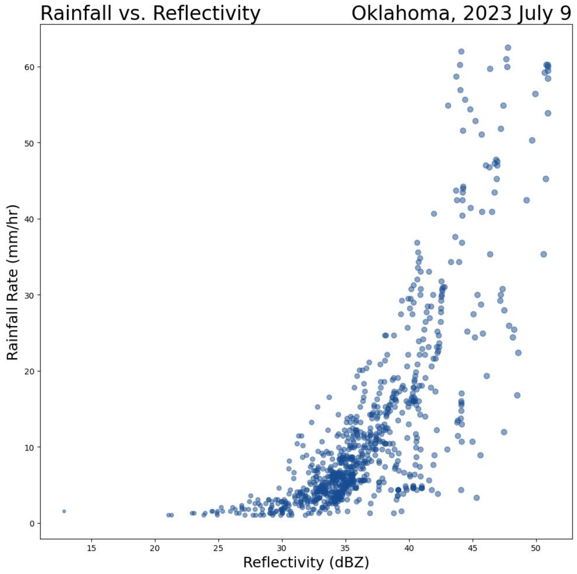
3.1 Reflectivity Z (dBZ)

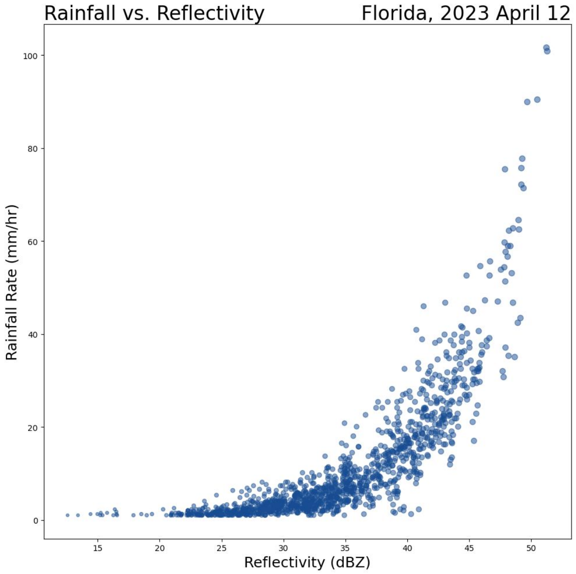
- Measures the amount of energy reflected back to the radar.
- Related to raindrop particle size and generally increases as rainfall rate increases [2].



3.1 Reflectivity–Rainfall (Z–R) Relationships

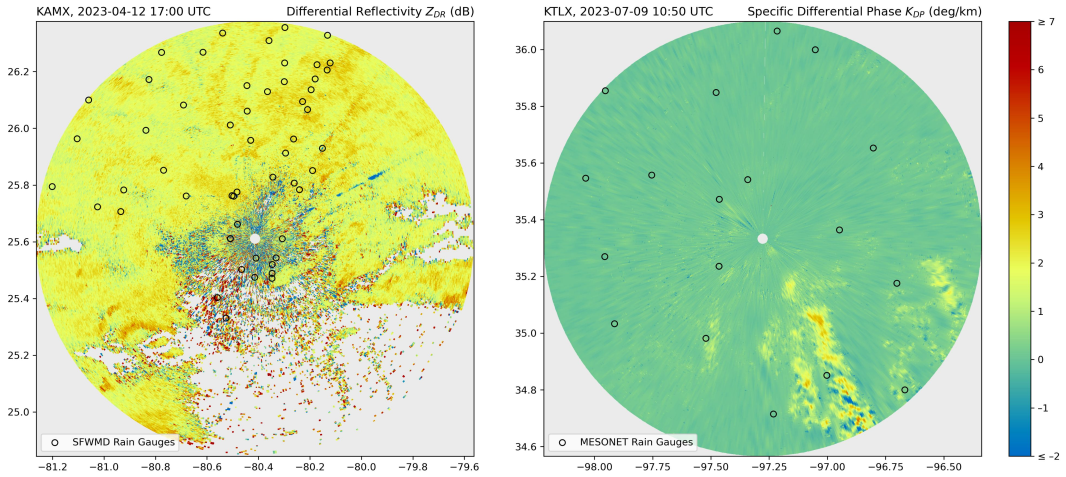






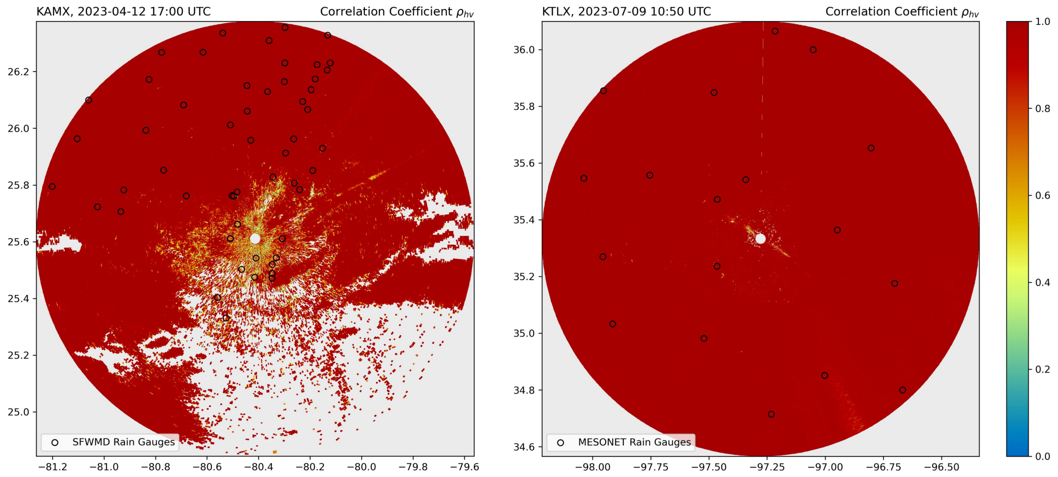
3.1 Differential reflectivity Z_{DR} (dB)

- Impacted by the composition or density of raindrops, helping differentiate water drops from ice pellets and snow [4].
- The ratio between reflectivity factors at horizontal and vertical polarizations.



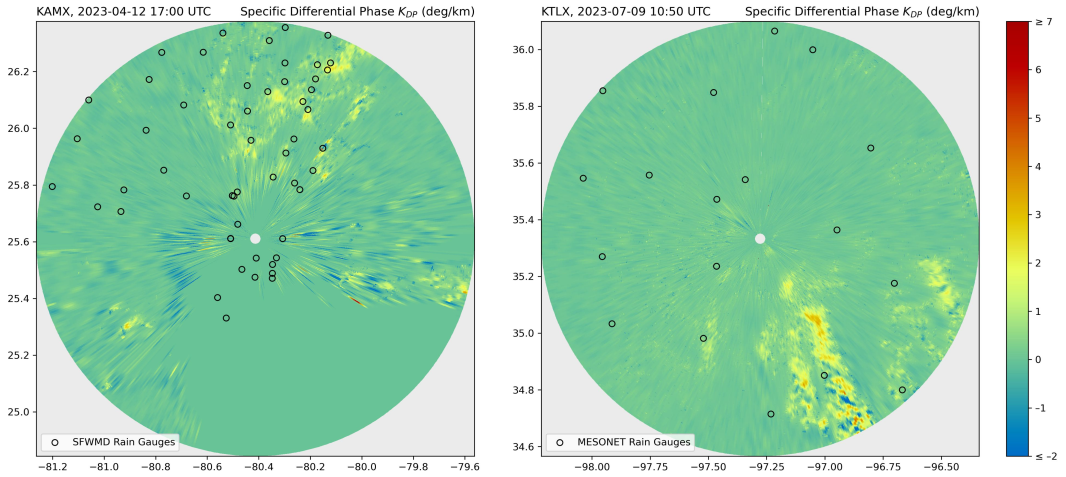
3.1 Co-polar correlation coefficient ρ_{hv}

- Measures variation in particle shapes and orientations [2, 4].
- Close to 1.0 during uniform rainfall and decreases with more variability in the types, shapes, and orientations of particles [4].



3.1 Specific differential phase K_{DP} (deg/km)

- Derived variable that represents the change in differential phase shift Φ_{DP} [2, 4].
- Useful for identifying heavy precipitation and when hail is mixed with rain, but can be noisy for light rain [4].



4. Methodology & Results

4.1 Benchmarking Symbolic Regression Algorithms

4.2 Symbolic Regression on Subsets of the Data Using Feyn

- Clusters (K-Means, Bisecting K-Means, Agglomerative Hierarchical)
- Decision Tree Leaf Nodes
- Grouping by Radar Variable Mean

4.3 Exploring New Symbolic Regression Models with gpg

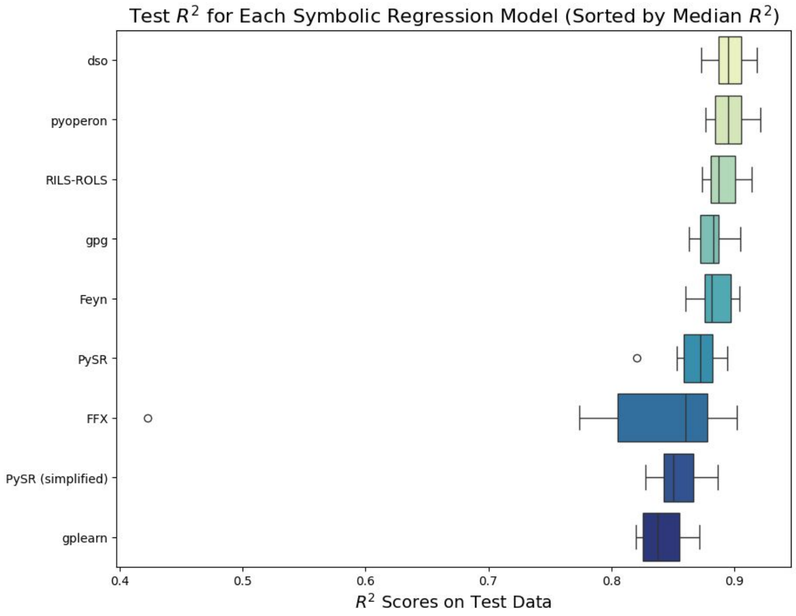
- Knowledge-based loss terms for gpg loss function

4.1 Benchmarking Symbolic Regression Methods

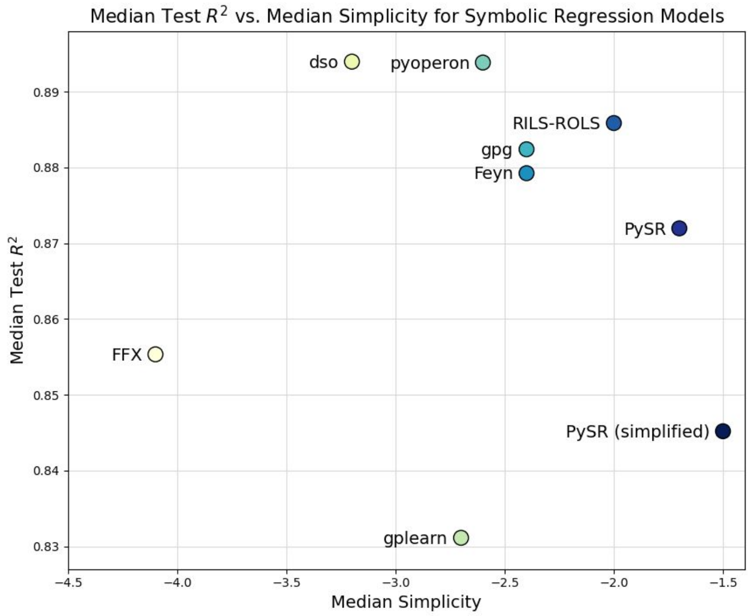
- Implemented eight symbolic regression methods based on criteria by La Cava et al. [1].
  - Genetic programming (gplearn, gpg, PySR, Feyn, pyoperon)
  - Deep learning (dso)
  - Other (FFX, RILS-ROLS)
- Run 10 trials with different training (75%) and testing (25%) sets.
- Analyzed accuracy using the coefficient of determination (R^2) and the normalized root square mean error (NRMSE).
- Analyzed equation complexity with a simplicity score indicating the number of components within the equation.

$$R^2 = 1 - \frac{\sum_{i=1}^{k}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{k}(y_i - \bar{y}_i)^2} \qquad NRMSE = \frac{\sqrt{\frac{1}{k}\sum_{i=1}^{k}(y_i - \hat{y}_i)^2}}{\bar{y}} \qquad simplicity = -log_5(s)$$

4.1 Benchmarking Symbolic Regression Methods



Each model's test R^2 scores over ten trials sorted by median test R^2

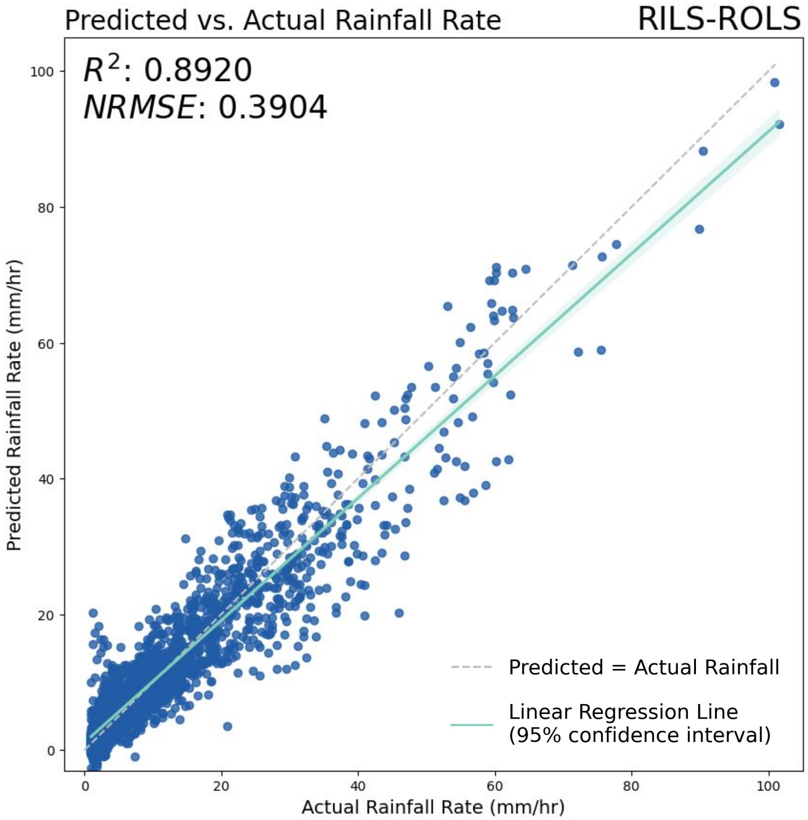


Each model's median test R^2 vs. median simplicity (simplicity closer to zero indicates simpler equations)

4.1 Benchmarking Symbolic Regression Methods

Equation with best combination of **accuracy and simplicity** using symbolic regression with RILS-ROLS

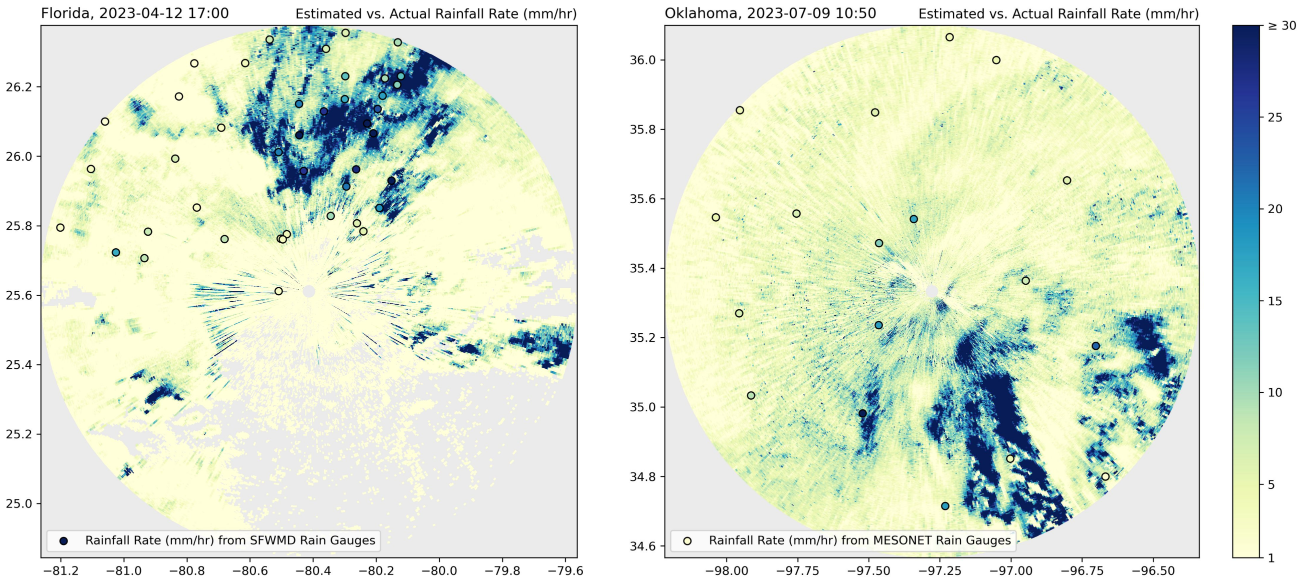
$$R = 1.208Z(K_{DP})\rho_{hv}^3 - 20.088K_{DP} + 2(10^{-6})\rho_{hv}^4 Z^4 e^{\cos(Z_{DR})} - 0.643$$



4.1 Benchmarking Symbolic Regression Methods

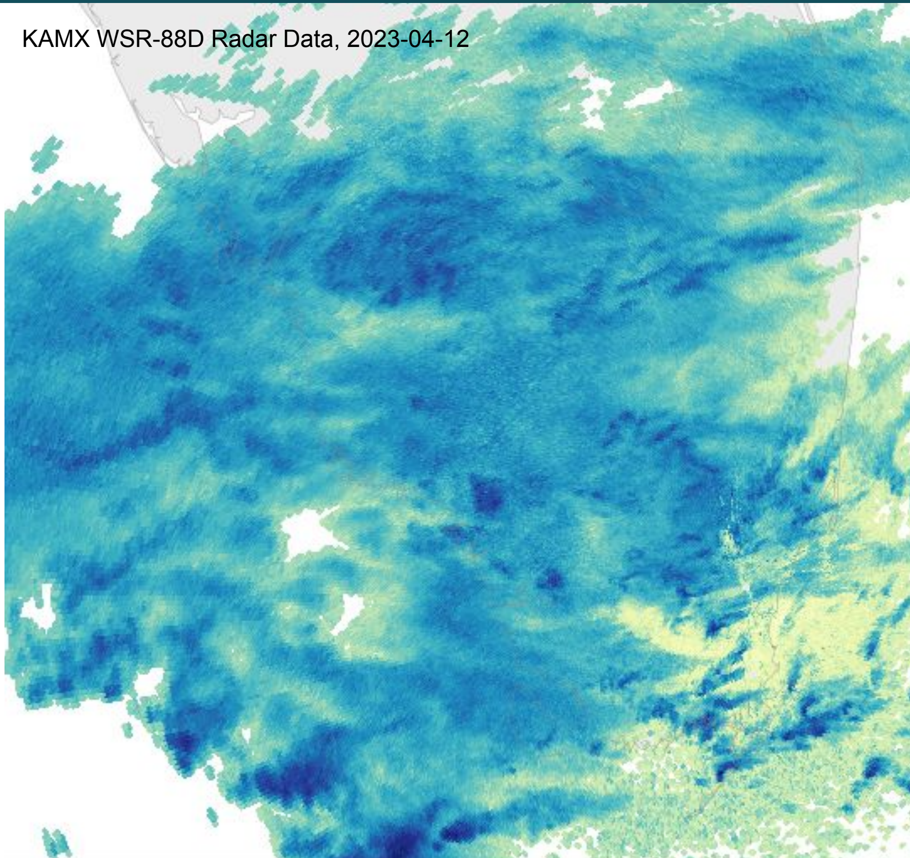
Predicted rainfall rate based on equation from RILS-ROLS

$$R = 1.208Z(K_{DP})\rho_{hv}^{3} - 20.088K_{DP} + 2(10^{-6})\rho_{hv}^{4}Z^{4}e^{\cos(Z_{DR})} - 0.643$$



4.2 Symbolic Regression on Subsets of Data

- One significant challenge to precipitation estimation is capturing **different precipitation types, distributions, and intensities**.
- Previous research has found that pre-processing the data to distinguish rainfall intensities has improved QPE accuracy [3].
- We test clustering algorithms, decision trees, and setting a threshold using Z_{DR} and ρhv to subset the data prior to running symbolic regression.

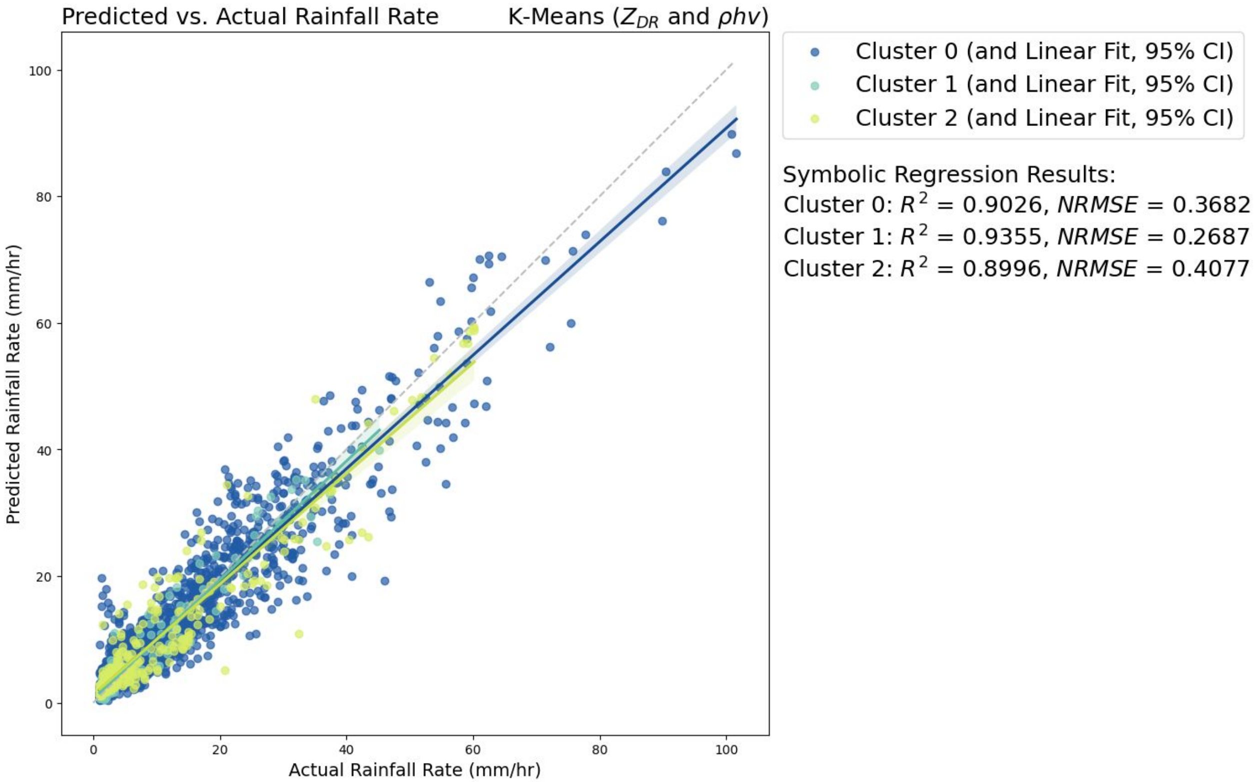


4.2 Symbolic Regression on Clusters

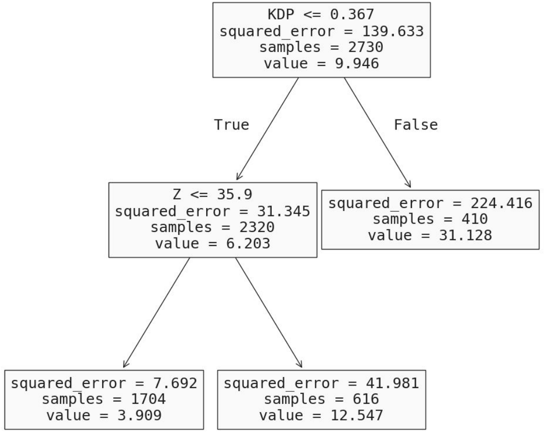
Mean metrics of three clusters from the trial with the highest mean test R^2 score for each clustering method using Feyn

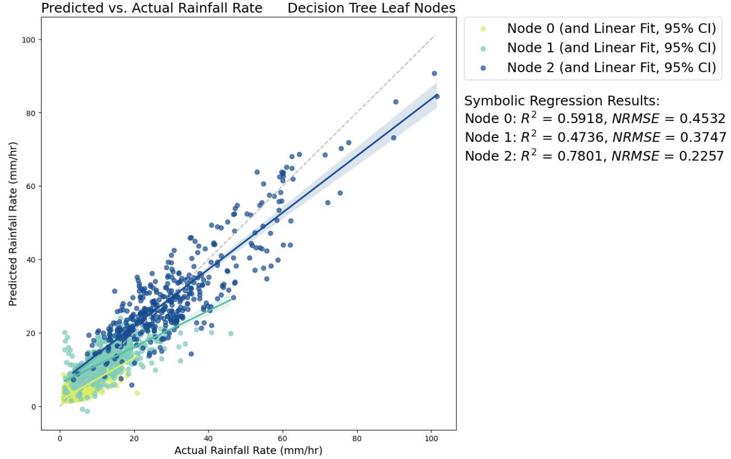
Cluster	Variable	Train R^2	Test R^2	Train NRMSE	Test NRMSE	Simplicity
All Data (Without Clustering)		0.8757	0.9046	0.4116	0.3817	-2.4
K-Means	All Radar	0.7382	0.7826	0.4142	0.3784	-2.1
	ρ_{hv} and Z_{DR}	0.9048	**0.9200**	0.3605	0.3250	-2.2
	Rain	0.6318	0.6764	0.2456	0.2434	-2.1
Bisecting K-Means	All Radar	0.7129	0.7527	0.4300	0.3936	**-2.0**
	ρ_{hv} and Z_{DR}	**0.9064**	0.8980	0.3554	0.3722	-2.1
	Rain	0.6106	0.6069	0.2214	0.2265	**-2.0**
Agglomerative	All Radar	0.7556	0.7887	0.4006	0.3727	-2.1
	ρ_{hv} and Z_{DR}	0.8973	0.8746	0.3758	0.3980	-2.3
	Rain	0.6466	0.6623	**0.2201**	**0.2248**	**-2.0**

4.2 Symbolic Regression on Clusters



4.2 Symbolic Regression on Decision Tree Leaf Nodes

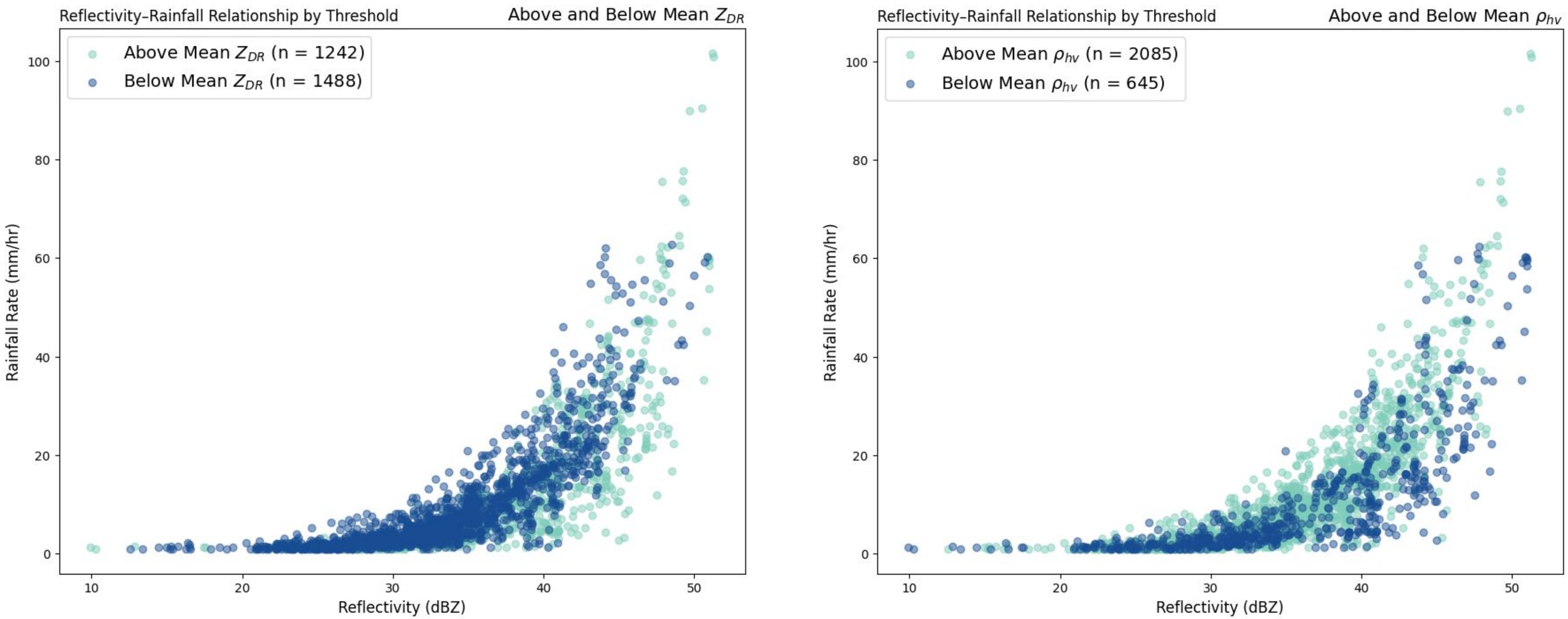




Metrics from the trial with the highest test R^2 score for each node using Feyn

Subset	Size	Train R^2	Test R^2	Train NRMSE	Test NRMSE	Simplicity
All Data	2730	0.8757	0.9046	0.4116	0.3817	-2.4
Node 1	1704	0.5775	0.6330	0.4620	0.4271	-2.1
Node 2	616	0.4510	0.5624	0.3923	0.3123	-2.1
Node 3	410	0.7773	0.7826	0.2277	0.2199	-2.0

4.2 Grouping by Radar Variable Mean

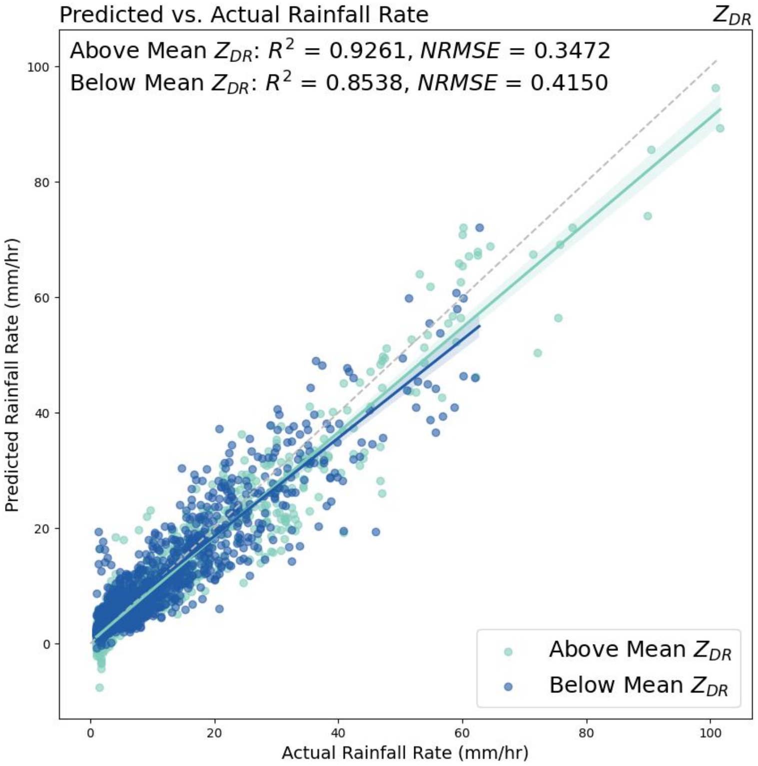


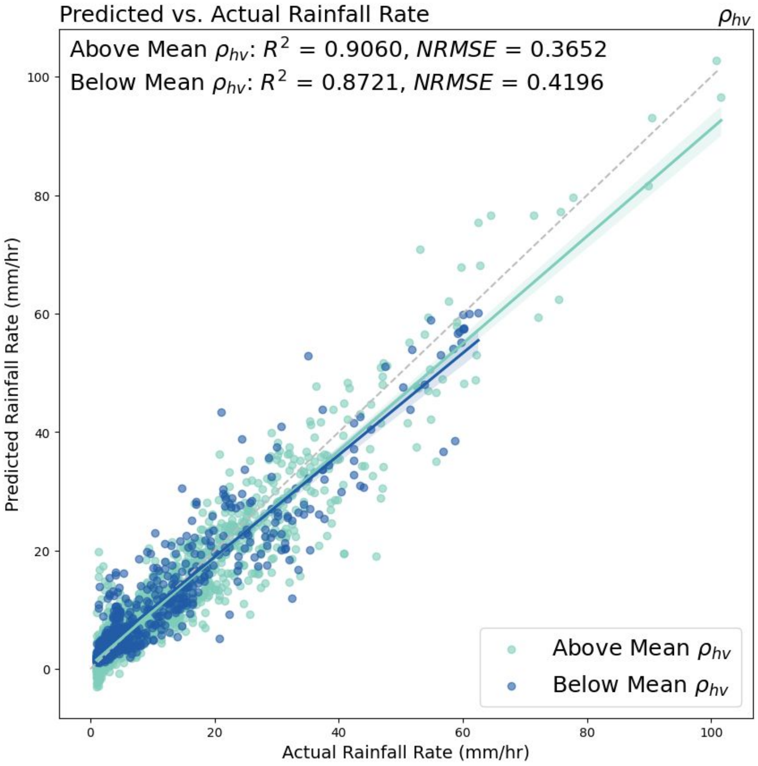
4.2 Grouping by Radar Variable Mean

Metrics from the trial with the highest test R^2 score for each group using Feyn

Variable	Group	Size	Train R^2	Test R^2	Train NRMSE	Test NRMSE	Simplicity
All Data		2730	0.8757	0.9046	0.4116	0.3817	-2.4
Z_{DR}	Above Mean	1242	**0.9161**	**0.9519**	0.3620	**0.2976**	-2.2
	Below Mean	1488	0.8538	0.8538	0.4096	0.4314	-2.3
ρ_{hv}	Above Mean	2085	0.9023	0.9132	**0.3554**	0.3898	-2.1
	Below Mean	645	0.8586	0.9025	0.4268	0.3985	-2.4

4.2 Grouping by Radar Variable Mean





4.3 Exploring New Symbolic Regression Models with gpg

- Incorporated knowledge-based loss terms [6] into the loss function of gpg
    - Z−R relation ($Z = aR^b$) loss term: $loss_f = loss(Y, \hat{Y}) + \lambda * loss(\hat{Y}, (\frac{Z}{a})^{\frac{1}{b}})$
    - Cluster-based loss term: $loss_f = loss(Y, \hat{Y}) - \lambda * silhouette\_score(\hat{Y}, \mathrm{L})$
        - Silhouette score: measures how well the rainfall rates are assigned to their predetermined clusters
    - Binned rainfall loss term: $loss_f = loss(Y, \hat{Y}) + \lambda * [loss(\widehat{Y_j}, L_j) + loss(\widehat{Y_j}, U_j)]$
        - Prior to training: split the data into three groups of low, medium, and high rainfall rate
        - Adds to the loss if the model predicts a rainfall rate not aligning with the groups `

Y = ground-truth rainfall rate
$\hat{Y}$ = predicted rainfall rate
Z = Reflectivity (mm^6/mm^{-3})
L = cluster labels
$\widehat{Y_j}$ = predicted rainfall rate for group $j = 1, 2, 3$
L_j = lower bound for group $j = 1, 2, 3$
U_j = upper bound for group $j = 1, 2, 3$
λ = weight parameter

4.3 Exploring New Symbolic Regression Models with gpg

Metrics from the model with the highest test R^2 score using custom loss functions in gpg

Loss Function	Train R^2	Test R^2	Train NRMSE	Test NRMSE	Simplicity
Original	0.8744	0.9049	0.4115	0.3842	-2.5
Z-R ($\lambda = 1$) (Equation 4.3.1)	0.8546	0.8900	0.4427	0.4132	**-2.3**
Silhouette score ($\lambda = 20$) (Equation 4.3.2)	0.8746	0.9060	0.4110	0.3819	-2.5
Binned rainfall ($\lambda = 0.01$) (Equation 4.3.3)	**0.8748**	**0.9067**	**0.4108**	**0.3804**	**-2.3**

- Including the binned rainfall term in the loss function generated a more accurate and less complex symbolic expression

5. Conclusion

Benchmarking

- Symbolic regression is effective for quantitative precipitation estimation, providing interpretable and accurate models.

Symbolic Regression on Subsets with Feyn

- There is potential for data pre-processing methods that subset the data to improve the accuracy of learned equations.
- Applying Feyn symbolic regression on three clusters resulting from k-means clustering based on Z_{DR} and ρ_{hv} achieved improved R^2 scores, lower NRMSE scores, and slightly simpler equations.

Custom Loss Functions with gpg

- Adjusting and applying custom loss functions in gpg slightly improved the R^2 scores and NRMSE scores while also improving the model simplicity.

5.1 Further Work

This study can be built upon in the following ways:

- Test symbolic regression models on a larger dataset encompassing more geographic regions and dates.
- Explore symbolic regression on time series data.
- Conduct a deeper analysis on how to incorporate domain knowledge into the loss function of symbolic regression models to further improve learned equations.

References

[1] W. La Cava, B. Burlacu, M. Virgolin, M. Kommenda, P. Orzechowski, F. O. de Franc¸a, Y. Jin, and J. H. Moore, “Contemporary symbolic regression methods and their relative performance,” Advances in neural information processing systems, vol. 2021, no. DB1, p. 1, 2021.

[2] R. Cifelli and V. Chandrasekar, Dual-Polarization Radar Rainfall Estimation. American Geophysical Union (AGU), 2010, pp. 105–125.

[3] J. Huangfu, Z. Hu, J. Zheng, L. Wang, and Y. Zhu, “Study on quantitative precipitation estimation by polarimetric radar using deep learning,” Advances in Atmospheric Sciences, pp. 1–14, 2024.

[4] M. R. Kumjian, “Principles and applications of dual-polarization weather radar. Part I: Description of the polarimetric radar variables.” Journal of Operational Meteorology, vol. 1, 2013.

[5] W. Li, H. Chen, and L. Han, “Polarimetric radar quantitative precipitation estimation using deep convolutional neural networks,” IEEE Transactions on Geoscience and Remote Sensing, vol. 61, pp. 1–11, 2023.

[6] L. Von Rueden, S. Mayer, K. Beckh, B. Georgiev, S. Giesselbach, R. Heese, B. Kirsch, J. Pfrommer, A. Pick, R. Ramamurthy et al., “Informed machine learning–a taxonomy and survey of integrating prior knowledge into learning systems,” IEEE Transactions on Knowledge and Data Engineering, vol. 35, no. 1, pp. 614–633, 2021.

[7] K. Shin, J. J. Song, W. Bang, and G. Lee, “Quantitative precipitation estimates using machine learning approaches with operational dual polarization radar data,” Remote Sensing, vol. 13, no. 4, 2021.

[8] L. Wang and H. Chen, “Machine learning for polarimetric radar quantitative precipitation estimation,” in 2023 United States National Committee of URSI National Radio Science Meeting (USNC-URSI NRSM), 2023, pp. 298–299.

[9] D. Wijayarathne, P. Coulibaly, S. Boodoo, and D. Sills, “Use of radar quantitative precipitation estimates (QPEs) for improved hydrological model calibration and flood forecasting,” Journal of Hydrometeorology, vol. 22, no. 8, pp. 2033–2053, 2021.

Thank You!