Supplementary Material: Understanding cirrus clouds using explainable machine learning

A. ML Model Hyperparameters

Table A.1. *Best performing hyperparameters for both ML models found by Bayesian optimization.*

(a) XGBoost

Maximum tree depth	15
Alpha	38
Lambda	7
Subsample ratio of the training data	0.4
Column subsample ratio for each tree	0.8
Number of trees	250
Learning rate	0.02

(b) LSTM + Attention

LSTM	hidden layer size	250
Final layers	Layer sizes	100, 50
	Dropout	0.5
	Activation function	ReLU
	Regularizer	Batch normalization
General	Maximum epochs	50
	Batch size	1000
	Learning rate	1e-5

B. XAI evaluation metrics

In this section, we provide further details on the metrics used to evaluate the post-hoc explanations. Since the metrics were originally proposed for classification tasks, they were adapted to fit the regression setting of this study.

Stability Relative Input Stability (RIS) and Relative Output Stability (ROS) are applied to evaluate the robustness of post-hoc feature attributions towards small changes in the input data. First, a modified input data set is created by adding random noise to the original samples. Then, explanations for the slightly modified samples are calculated. Finally, the relative distance between original and modified explanations with respect to the distance between the original and modified sample (RIS) and original and modified prediction (ROS) are calculated with smaller relative distances representing more stable explanations. The code to calculate the stability metrics was adapted from (Agarwal et al., 2022).

Estimated Faithfulness The metric is calculated by incrementally removing each of the attributes deemed important by the post-hoc feature attribution method and evaluating the effect on the performance. Features are removed by replacing them with their mean value. The correlation between the importance score of a feature and the feature's effect on model performance yields the faithfulness measure with higher correlations representing more faithful explanations. Our implementation is an adaption of (Arya et al., 2019).

Supplementary Material References

Agarwal, C., Krishna, S., Saxena, E., Pawelczyk, M., Johnson, N., Puri, I., Zitnik, M., and Lakkaraju, H. (2022). OpenXAI: Towards a Transparent Evaluation of Model Explanations.

Arya, V., Bellamy, R. K. E., Chen, P.-Y., Dhurandhar, A., Hind, M., Hoffman, S. C., Houde, S., Liao, Q. V., Luss, R., Mojsilović, A., Mourad, S., Pedemonte, P., Raghavendra, R., Richards, J., Sattigeri, P., Shanmugam, K., Singh, M., Varshney, K. R., Wei, D., and Zhang, Y. (2019). One Explanation Does Not Fit All: A Toolkit and Taxonomy of AI Explainability Techniques.