Data Science Research Project in the School of Mathematical Sciences

Ning Ni
a1869549

July 12, 2024

Report submitted for **MATHS 7097B** at the School of Mathematical Sciences, University of Adelaide

THE UNIVERSITY of ADELAIDE

Project Area: **Hybrid Optical/Radio Frequency Communication Channel Model**
Project Supervisor: **Siu Wai Ho**

In submitting this work I am indicating that I have read the University's Academic Integrity Policy. I declare that all material in this assessment is my own work except where there is clear acknowledgement and reference to the work of others.

I give permission for this work to be reproduced and submitted to other academic staff for educational purposes.

OPTIONAL: I give permission this work to be reproduced and provided to future students as an exemplar report.

Abstract

This paper investigates the impact of weather on channel attenuation in hybrid Radio Frequency Line (RFL) and Free Space Optical (FSO) communication systems, focusing on establishing robust predictive models for attenuation under varying weather conditions. By employing ensemble algorithms, particularly Random Forest, the study explores the relationship between weather parameters and channel attenuation, identifying key predictors for different weather scenarios. Through extensive data preprocessing, exploratory data analysis (EDA), and model development, specific, generic, and hybrid models are constructed for attenuation prediction.

The research underscores the importance of feature pruning and hyperparameter tuning in optimizing model complexity and performance. Notably, the streamlined generic models outperform specific and hybrid models, reducing computational resource demands and enhancing interpretability. The study reveals the critical influence of weather factors on attenuation, identifying distinct sets of predictors for RFL and FSO channels across various weather conditions. The generic models demonstrate superior predictive capabilities, particularly excelling in foggy and dusty environments with smaller sample sizes. However, specific models outperform generic models in clear and snowy conditions for the RF channel.

On the other hand, hybrid models, which utilize the predicted attenuation of one channel as training data for predicting the attenuation of the other channel, effectively capture the correlation between RFL and FSO channels. Nevertheless, this approach does not improve prediction performance compared to generic models.

Comparisons with actual values show that generic models provide effective and accurate predictions of channel attenuation in both RFL and FSO channels. Overall, these findings offer valuable insights for mitigating weather-induced challenges in hybrid RFL/FSO communication systems, fostering the development of more resilient and efficient wireless connectivity solutions in diverse environmental conditions.

1 Introduction

In the realm of wireless communication systems, the integration of RFL and FSO technologies has emerged as a promising frontier. This hybrid approach leverages the unique advantages of both RFL and FSO channels to enhance data rates and link availability. However, the effectiveness of hybrid RFL/FSO systems can be compromised by environmental factors such as rain, fog, or dust storms, which significantly impact channel attenuation and reduce received signal power. Understanding and mitigating these weather-induced effects are crucial for maximizing system efficacy and reliability.

This study aims to comprehensively explore weather-induced channel attenuation in hybrid RFL/FSO communication systems and establish reliable channel models to predict RFL and FSO channel attenuation. To achieve this, ensemble algorithms such as Random Forest [1], Adaptive Boosting (AdaBoost) [2], and Extreme Gradient Boosting (XGBoost) [3] may be employed. These algorithms will help recognize the pivotal role of weather in system performance and determine which channel is more effective in specific weather conditions. Additionally, this study develops hybrid models aimed at exploring the correlation between RFL and FSO channels, including both linear and nonlinear relationships [4], under different weather conditions.

The outcome of this research will be a suite of weather-aware attenuation models tailored to hybrid RFL/FSO communication systems, which contain generic, specific and hybrid models. Through rigorous evaluation and comparison, these models will illuminate the intricate interplay between weather conditions and channel attenuation, providing reliable predictions for channel attenuation in real-world scenarios. Ultimately, these findings will inform the future development of hybrid RFL/FSO satellite communication systems, ushering in a new era of high-throughput wireless connectivity.

2 Background

Hybrid RFL/FSO systems, just like Fig. 1, merge RFL and FSO technologies to enhance performance, reliability, and flexibility in wireless communication networks [5]. Recognizing the unique strengths and limitations of each, RFL offers resilience to atmospheric conditions but struggles with high data rates, while FSO provides high data rates but is susceptible to turbulence and obstructions [6] [7].

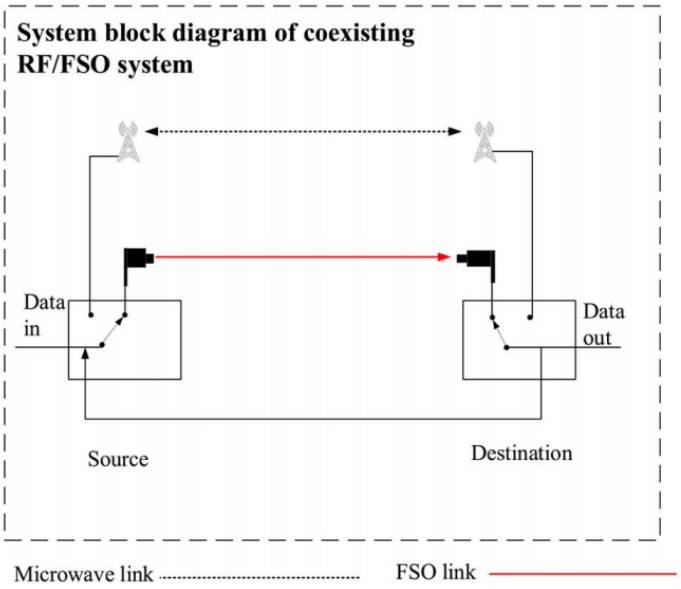


Figure 1: Example Diagram of A Hybrid RFL/FSO System

By combining RFL and FSO, hybrid systems leverage the advantages of both. RFL acts as a reliable backup when FSO is disrupted by weather or obstacles like fog and snow. Conversely, FSO boosts data rates compared to RFL during periods of low turbulence, maximizing network capacity [8].

Implementation involves seamless integration of RFL and FSO transceivers with intelligent switching mechanisms. Machine learning algorithms enable autonomous adaptation to environmental changes, optimizing resource allocation and mitigating turbulence impact [9] [10]. These models facilitate collaboration between RFL and FSO components, enabling efficient data routing and fault tolerance.

Overall, hybrid RFL/FSO systems offer improved reliability, flexibility, and performance in wireless networks. Ongoing research aims to further optimize their design for applications including telecommunications, disaster recovery, and remote sensing.

3 Methods

The data source for this research comprises real empirical data collected from a hybrid system operating in six cities globally.The methodology of this study will be divided into four main parts, including data cleaning, exploratory data analysis, model selection and evaluation metrics, generic, specific and hybrid model design. The Integrated Development Environment (IDE) utilized for this research is Visual Studio Code (version 1.86.1), and the programming language employed is Python 3.9.

3.1 Data Cleaning

The dataset studied in this research comprises 91,379 samples and 27 variables. "FSO_Att" and "RFL_Att" are the target variables. "SYNOPCode","Time",and "Frequency" are categorical variables. Others are numerical variables.

The dataset is free from any duplicated values and missing entries. This study utilize the Interquartile Range (IQR) method to identify potential outliers in some variables, as depicted in Fig. 2, yet they fall within reasonable ranges. The "RainIntensity" reaches a maximum of 90 mm/h, and the "Particulate" concentration, indicating the presence of particulate matter in the air, can peak at 1600 ug/m^3. Despite these values appearing as outliers, it's possible to occur within extreme natural environments.

3.2 Exploratory Data Analysis (EDA)

3.2.1 Target Variables Analysis

Through the analysis depicted in Fig. 3, it is observed that "FSO_Att" exhibits a bimodal right-skewed distribution within the range of 0.788 dB to 32.455 dB. On the other hand, "RFL_Att" displays a unimodal right-skewed distribution spanning from 0.027 dB to 46.893 dB.

3.2.2 Predictor Variables Analysis

The "AbsoluteHumidity" ranges from 1 g/m^3 to 25 g/m^3, with a median around 7 g/m^3. The "Particulate" is mostly distributed around 0 ug/m^3, indicating low particle content in the air and relatively clear visibility. In rare cases, it can exceed 200 ug/m^3, especially during dust storms, significantly impacting visibility. "RainIntensity" reflects mostly clear days, with occasional instances of heavy rainfall reaching up to 90 mm/h, as Fig. 4.

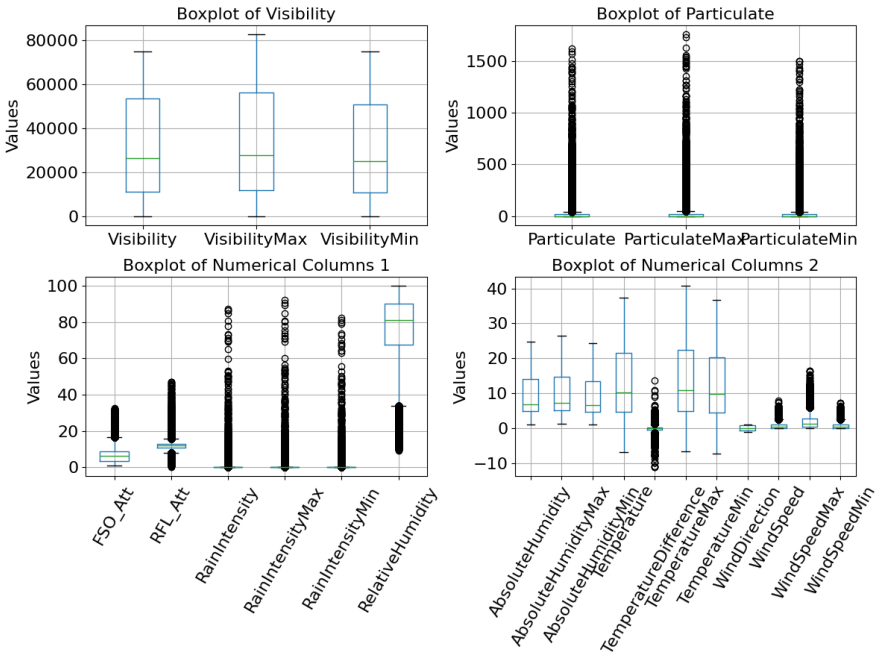


Figure 2: Outliers Analysis

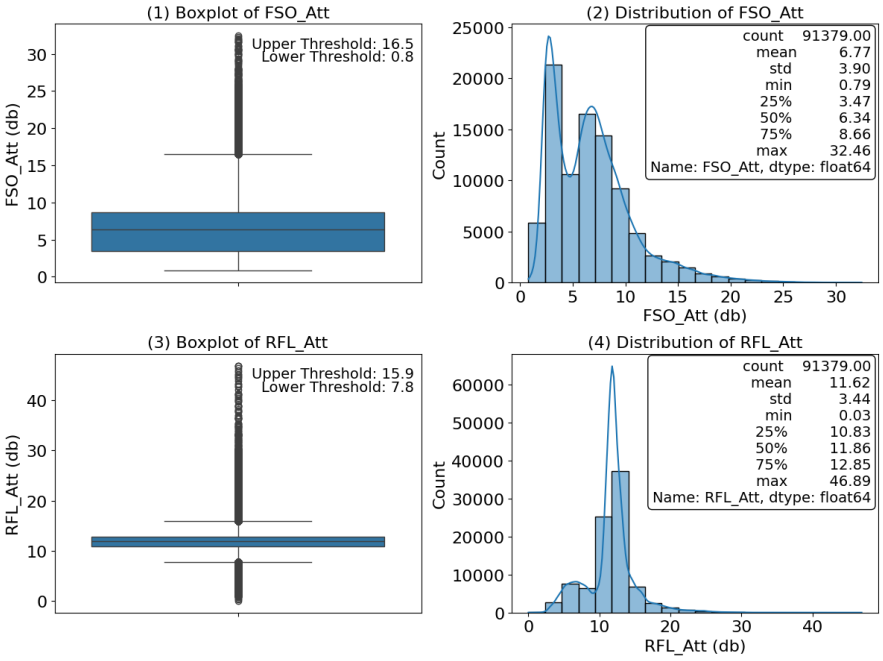


Figure 3: Target Variables Analysis

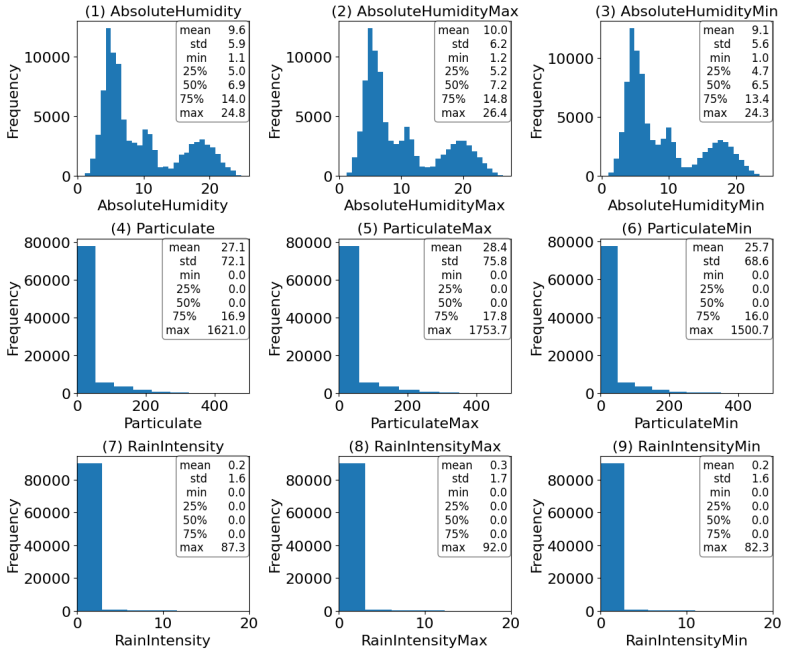


Figure 4: The Histogram of "Particulate","AbsoluteHumidity" and "RainIntensity"

Figure 5 shows that "Temperature" ranges from 1 degree Celsius to 40 degrees Celsius, with a median around 10.3 degrees Celsius, exhibiting a distribution pattern similar to "AbsoluteHumidity". More than three-quarters of "Visibility" exceed 10,000 meters. "WindSpeed" is concentrated between 0 and 5 m/s, with occasional maximum speeds reaching 16 m/s.

Figure 6 indicates that "Distance" is mainly distributed at 2100 meters, 2950 meters, 3950 meters, and 4800 meters. "WindDirection" indicates predominantly northerly and northwesterly winds, while other wind directions are distributed fairly evenly. "Frequency" shows that only microwave frequencies of 73.5 GHz and 83.5 GHz are present in the RF channel. "SYNOPCode" represents the overall weather conditions of the day, where 0 indicates clear skies, accounting for the majority of cases. Following this are 5 representing drizzle and 6 for rain. Other conditions include 3 for dust, 4 for fog, 7 for snow, and 8 for showers, but these are rare occurrences. The data spans from 0 to 24 hours in time intervals of 1 hour, evenly distributed.

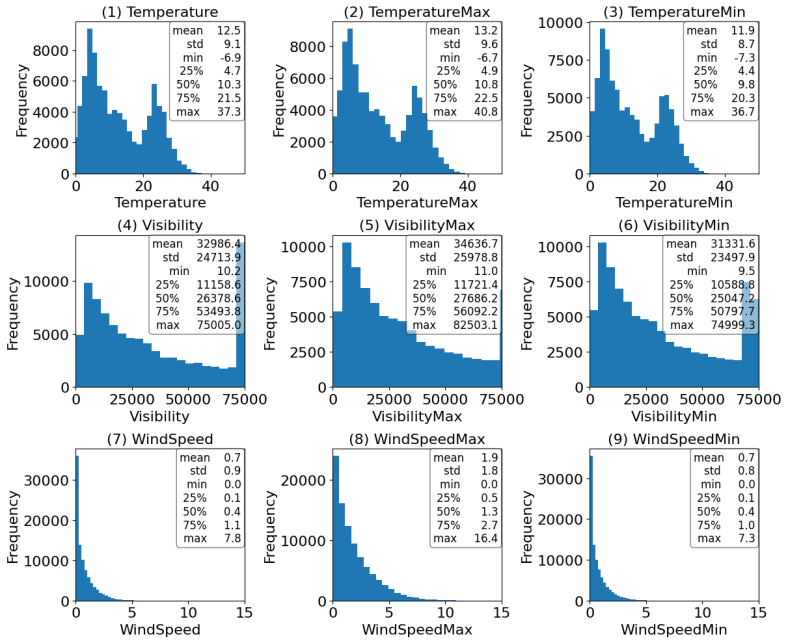


Figure 5: The Histogram of "Temperature","Visibility" and "WindSpeed"

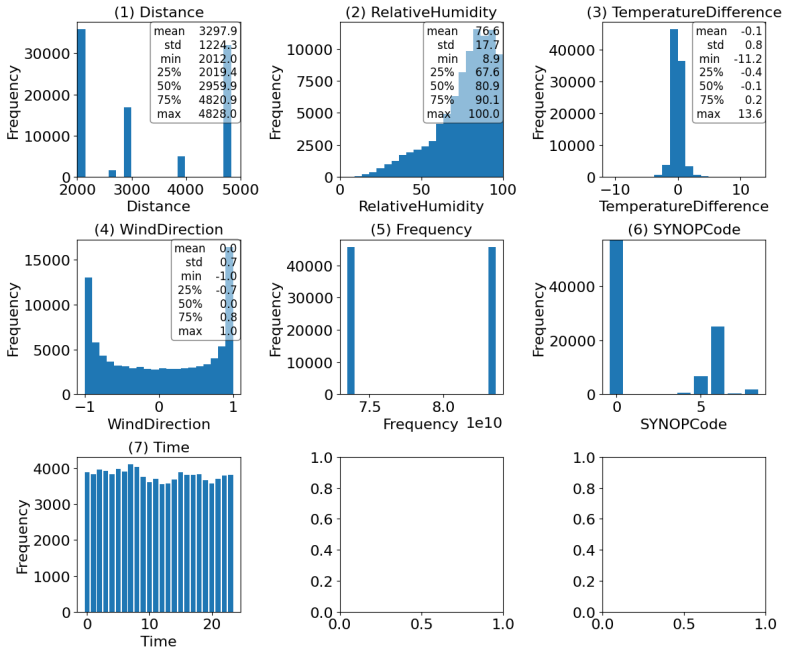


Figure 6: The Histogram of Other Features

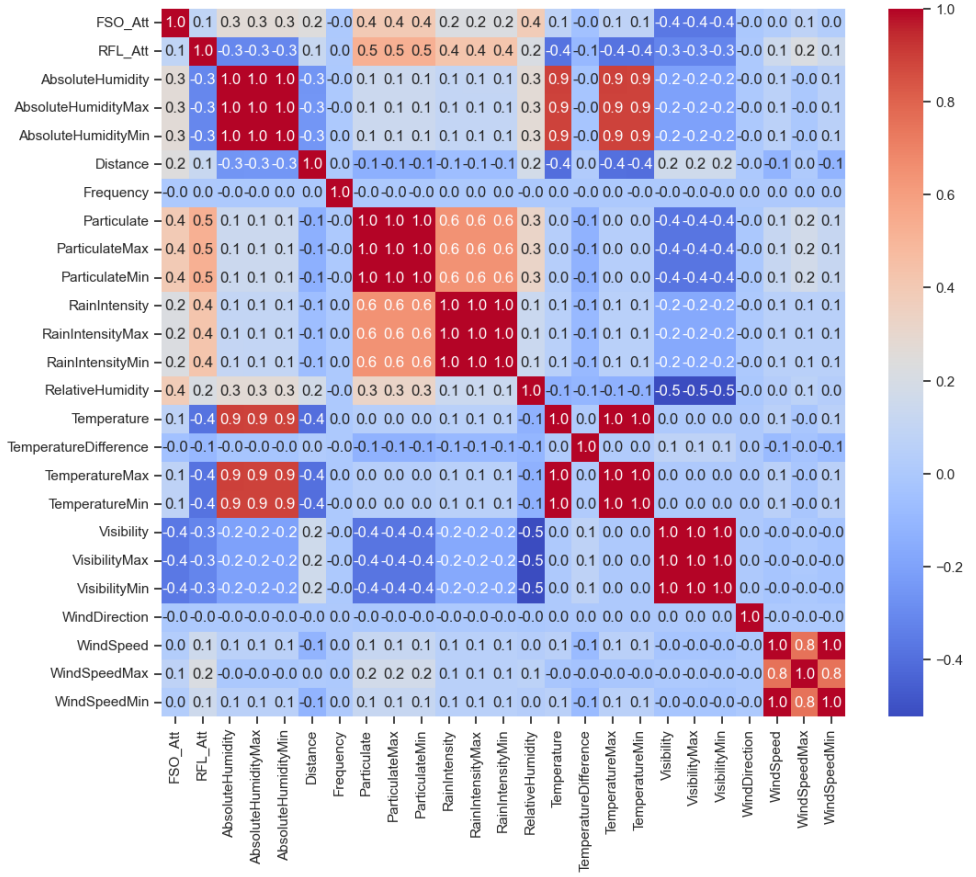


Figure 7: Correlation Heatmap of Target vs. Numerical Variables

3.2.3 The correlation analysis between Target and Predictor Variables

Combining Fig. 7 and Fig. 8, there is a weak linear correlation between FSO_Att and "Particulate" and "RelativeHumidity", while RFL_Att shows weak linear correlations with "Particulate", "RainIntensity", and "Temperature". The attenuation of both channels does not exhibit linear correlations with other numerical variables. "Temperature" has a strong linear correlation with "AbsoluteHumidity", and "Particulate" shows a clear linear correlation with "RainIntensity". Some features also exhibit strong linear correlations with their corresponding maximum and minimum values.

Through Fig. 9, it reveals the characteristics of the FSO channel and RF channel. In foggy (SYNOPCode 4), dusty (SYNOPCode 3), and snowy (SYNOPCode 7) weather conditions, the optical channel attenuation is greater than in other weather environments, while the RF channel remains relatively stable in these environments, less susceptible to inter-

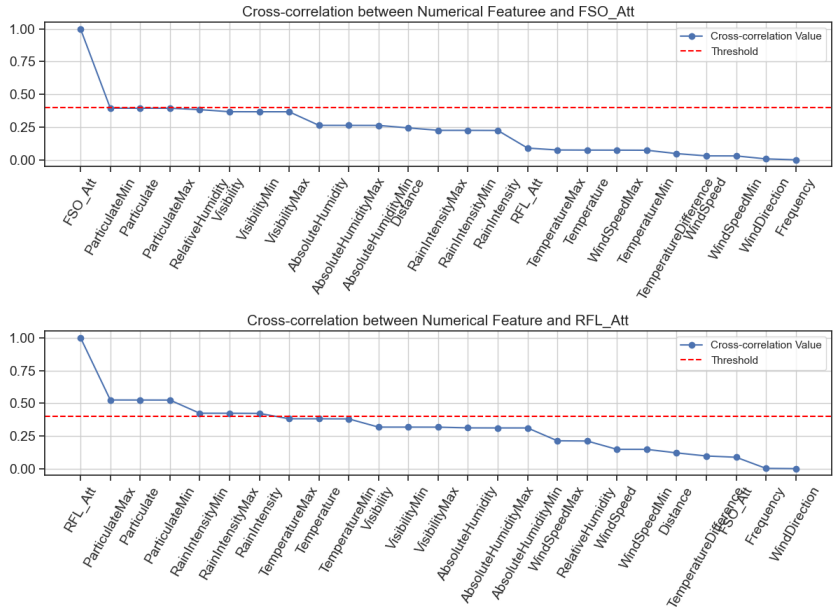


Figure 8: Cross-correlation between Numerical Feature and Attenuation

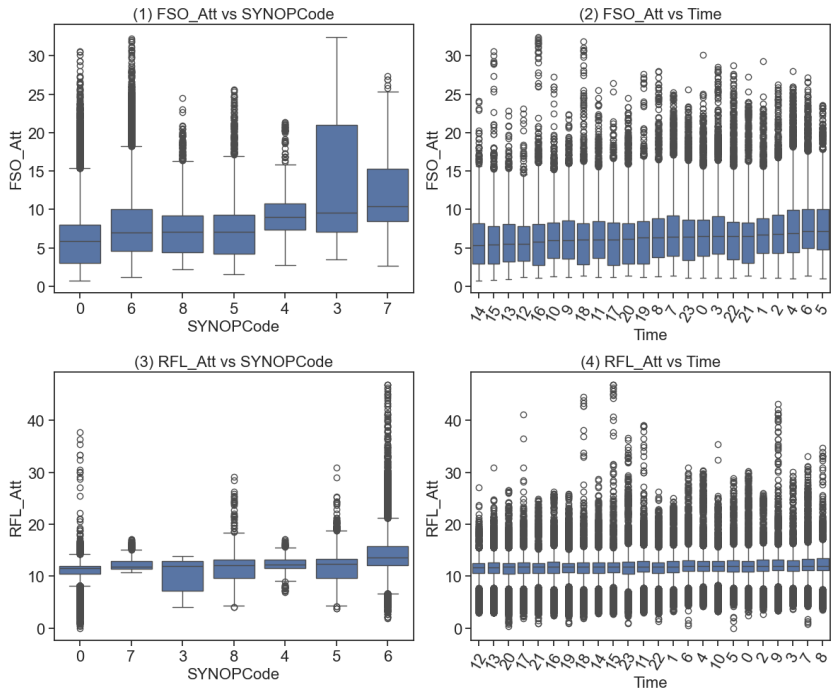


Figure 9: Target vs. Categorical Variables

ference. The FSO channel appears to have lower attenuation during the day compared to night, whereas the RF channel exhibits similar attenuation levels across different time periods without significant differences.

3.3 Model Selection and Evaluation Metrics

3.3.1 The Basic Architecture of Random Forest

Random Forest is an ensemble learning method that constructs a multitude of decision trees during training and then combines their outputs to make predictions. Each decision tree in the Random Forest is trained independently, typically using a random subset of the training data and a random subset of the features. This randomness helps to reduce overfitting and enhance robustness. During prediction, the results of all individual trees are aggregated through voting (for classification tasks) or averaging (for regression tasks) to produce the final prediction. This ensemble approach often results in a more accurate and robust model compared to a single decision tree, as showing Fig. 10.

In Fig. 10, The training dataset, consisting of 250 rows and 100 columns, is randomly sampled with replacement n times. Subsequently, a decision tree is trained on each sample. Finally, during the prediction phase, the outcomes of all n trees are aggregated to yield a final decision.

3.3.2 Bagging

Bagging, or Bootstrap Aggregating, is a fundamental sampling technique in Random Forests. It operates by training each decision tree on a randomly sampled subset of the training data, with replacement, resulting in multiple bootstrap samples, as Fig. 11. In an ideal case, about 36.8 % of the total training data forms the "out-of-bag (OOB)" sample and 63.2% of that contributes each bootstrap sample. This can be shown as Eq.(1),where N is the total number of samples.

Each tree is then independently trained on one of these samples, imparting diversity as they learn different aspects of the data. During prediction, the results from all trees are aggregated, often through majority voting for classification or averaging for regression tasks. This aggregation reduces variance and mitigates overfitting, making Random Forests effective at generalizing to unseen data.

It's worth nothing that the concept of OOB is specific to each individual tree. Although a sample may be considered as "out-of-bag data" for one tree, it could also be part of the training set for another tree. As the number of trees increases, there is no concept of OOB data for

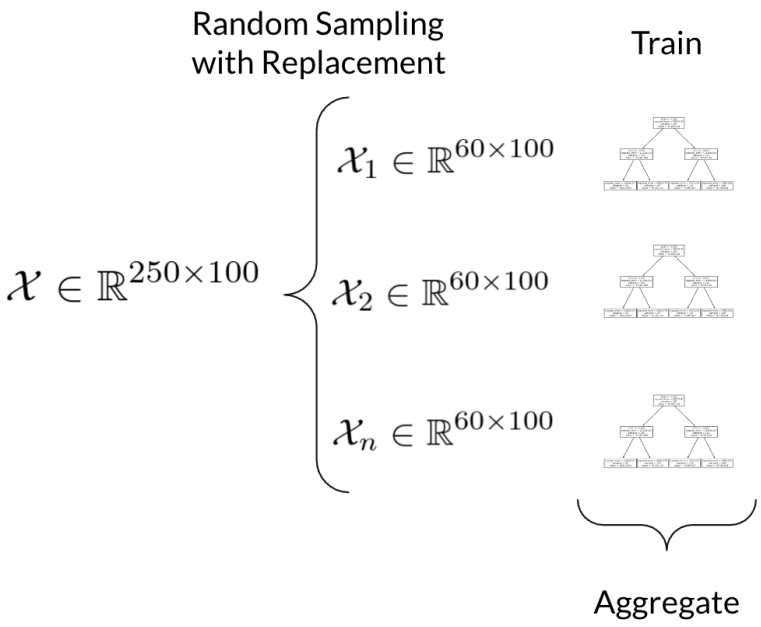


Figure 10: Basic Construction of Random Forest [11]

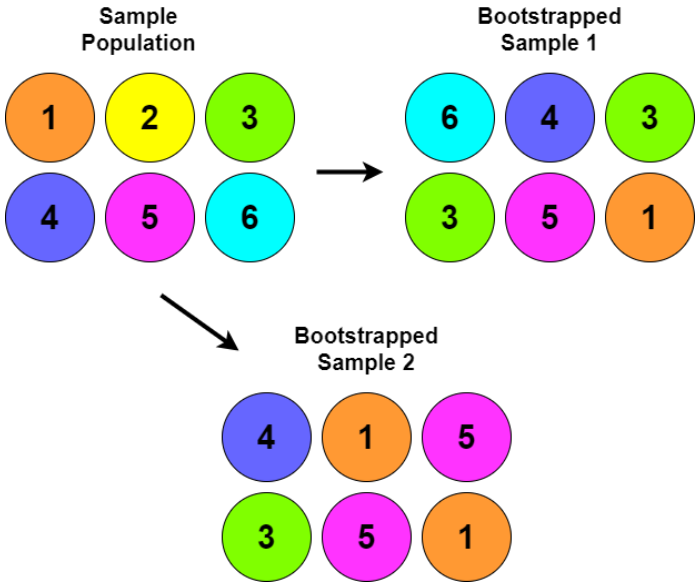


Figure 11: Bagging

the entire random forest because the entire forest is trained on the entire dataset.

$$\lim_{N\to\infty}\left(1-\frac{1}{N}\right)^{N}=e^{-1}=0.368 \tag{1}$$

3.3.3 Branching Principles

Decision trees measure impurity using either the Gini coefficient or information entropy for classification tasks, and Mean Squared Error (MSE) for regression tasks. At each split, the tree evaluates impurity for all features, selecting the feature that minimizes impurity for branching. Subsequently, at each child node, impurity is recalculated for each feature, and the process iterates, choosing the feature that minimizes impurity. With each branching layer, the overall impurity decreases, as the tree seeks to minimize impurity. The decision tree continues branching until no more features are available or the impurity metric is optimized.

However, decision trees are prone to overfitting. To mitigate this, random forests offer an effective solution. Firstly, they utilize bagging to ensure that each decision tree in the forest operates on a different subset of samples. Additionally, at each node, random forests randomly select a subset of features. It's important to note that earlier random decision forests employed the "random subspace method" [12], where each tree received a random subset of features. However, the current approach involves selecting different subsets of features for each node, while providing each tree with the complete set of features [13]. In summary, while each tree in a random forest receives the full set of features, only a random subset of features is considered at each node.

3.3.4 Evaluation Metrics

In this study, Root Mean Squared Error (RMSE) and R-square (R^2) are employed to evaluate models' performances, as Eq.(2) and Eq.(3), where $\bar{y}$ is the numerical average of y_n over all N. RMSE provides a measure of the model's error in the same units as the target variable, while R^2 evaluates the proportion of variance in the target variable explained by the model [14].

$$RMSE=\sqrt{\frac{1}{N}\sum_{n=1}^{N}\left(y_n-\hat{y}_n\right)^2} \tag{2}$$

$$R^2=1-\frac{\sum_{n=1}^{N}\left(y_n-\hat{y}_n\right)^2}{\sum_{n=1}^{N}\left(y_n-\bar{y}\right)^2} \tag{3}$$

In addition to RMSE and R2 for assessing the prediction accuracy of individual channels, this project employs the Pearson Correlation Coefficient (PCC) and Mutual Information (MI) methods to analyze and evaluate the linear and information correlations between the two channels under specific weather conditions. This evaluation helps assess the performance of the hybrid prediction model.

PCC is a statistical measure that quantifies the linear correlation between two sets of data. It is defined as the ratio of the covariance of the two variables to the product of their standard deviations, thereby normalizing the measure of covariance, as Eq.(4). This normalization ensures that the PCC value always falls between -1 and 1.The PCC can only capture linear relationships between variables, shown as Fig. 12. In the study, (-0.2, 0.2) is defined as weak linear correlation, [0.2, 0.5] and [-0.5, -0.2] represent moderate linear correlation, (0.5, 1] and [-1, 0.5) represent strong linear correlation.

In the project, X_i and Y_i represent the values of RFL_Att. and FSO_Att. in the sample, respectively. $\bar{X}$ and $\bar{Y}$ represent their average values. The PCC can reflect the linear relationship between the attenuation of the two channels, which can be used to evaluate whether the established models accurately represent the real linear relationship of the two channels.

$$PCC = \frac{\sum_{i=1}^{n} \left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sqrt{\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2} \sqrt{\sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2}} \tag{4}$$

In information theory, entropy measures the uncertainty or the amount of information contained in a message. Introduced by Claude Shannon, it quantifies the expected value of the information contained in a message. Shannon entropy H for a discrete random variable X with possible values $\{x_1, x_2, \ldots, x_n\}$ and probability mass function $P(X)$ is shown as Eq.(5).

$$H(X) = -\sum_{i=1}^{n} P\left(x_i\right) \log P\left(x_i\right) \tag{5}$$

The joint entropy $H(X, Y)$ of two random variables X and Y is a measure of the total uncertainty or information content associated with the pair (X, Y). It quantifies the amount of information needed to describe the outcomes of both X and Y together. The joint entropy $H(X, Y)$ is defined as Eq.(6), where $P(x, y)$ is the joint probability mass function of X and Y.

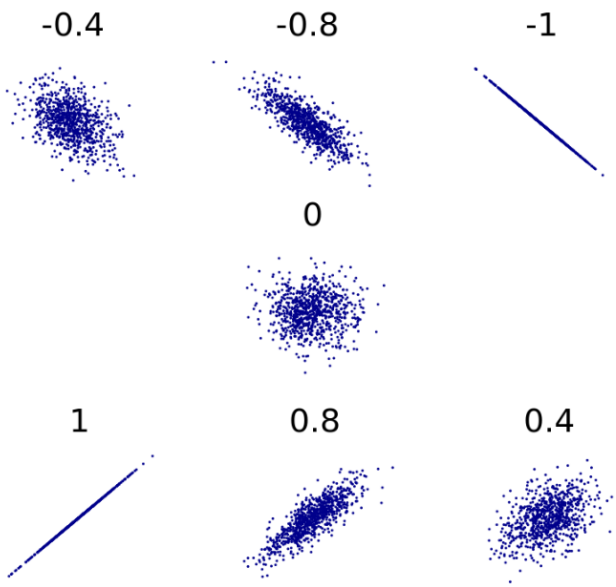


Figure 12: Example of Correlation Coefficient [15]

$$H(X,Y) = -\sum_{x\in X}\sum_{y\in Y} P(x,y)\log P(x,y) \tag{6}$$

MI quantifies the amount of information one random variable contains about another, capturing their mutual dependence. MI is defined based on the concept of entropy,as Eq.(7). It measures the reduction in uncertainty of one variable due to knowledge of the other, thus quantifying how much knowing the value of one variable reduces uncertainty about the other. Unlike linear correlation, which assesses the strength and direction of a linear relationship between two variables, MI evaluates the total dependence, encompassing both linear and non-linear correlations. This capability allows MI to detect more complex dependencies that linear correlation might miss, making it a versatile and powerful tool for understanding interactions between variables.

$$I(X;Y) = H(X) + H(Y) - H(X,Y) \tag{7}$$

However, MI is not upper bounded, and its values can vary widely, which does not provide an intuitive sense of the degree of dependency between variables. In contrast, Normalized Mutual Information (NMI) normalizes the MI value, providing a standardized measure within a fixed range of 0 to 1, shown as Eq.(8). This normalization makes it easier to interpret the degree of similarity or dependence between variables. A higher NMI directly translates to a stronger relationship.

Table 1: Data Splitting

Data set	sample number	ratio
training	63965	70%
validation	13707	15%
test	13707	15%

$$\mathrm{NMI}(X, Y) = \frac{I(X;Y)}{H(X,Y)} \tag{8}$$

3.3.5 Data Splitting

In this study, the dataset is divided into training, validation and testing set, as Tab. 1.

The training set is used to build the random forest model, with each decision tree randomly selecting samples from it for training, ensuring that each tree is constructed based on a different subset of the data.

The validation set is employed to tune the model's hyperparameters, evaluating the performance of different parameter settings and selecting the best configuration to enhance the model's generalization ability.

The test set is utilized to assess the model's performance, validating its predictive capability on unseen data and providing performance metrics for real-world applications in predicting attenuation, thus evaluating the model's reliability and utility.

Note that in the process of developing and evaluating hybrid models, the use of test datasets involves merging the validation dataset to constitute 30% of the total dataset.

3.3.6 Decision Tree Model

Random Forest is a type of ensemble learning method, which combines multiple weak learners to form a strong learner. Decision tree serves as the fundamental weak learner in Random Forest. Each decision tree is trained on a random subset of the training data, and then integrated by methods such as voting or averaging to produce the final prediction.

The significance of exploring weak learners lies in two aspects:

For Random Forest to yield reliable results, individual decision trees must demonstrate certain predictive performance, achieving high accuracy on both the training and testing sets (for classification problems) or low error rates (for regression problems). Otherwise, even with multiple weak learners integrated, Random Forest cannot provide reliable predictions.

Table 2: Decision Tree Model Hyperparameters

Hyperparameter	Range
splitter	"best" or "random"
criterion	"gini" or "entropy"
max_depth	*range(10,41,2)
min_samples_leaf	[1,10,50,100]
min_samples_split	[2,11,51,101]

Decision trees in Random Forest should have moderate complexity, neither too simple (high bias) nor too complex (high variance). Decision trees that are too simple may lead to underfitting, while overly complex decision trees may result in overfitting. By investigating the performance of individual decision trees, suitable parameter ranges can be obtained, such as the maximum depth of the decision tree (max_depth), the minimum number of samples required to be at a leaf node (min_samples_leaf), and the minimum number of samples required to split an internal node (min_samples_split). This provides an approximate parameter range for Random Forest hyperparameter tuning, reducing the learning time for hyperparameter adjustment.

Based on the value ranges provided in Tab. 2 for training the model and tuning hyperparameters, Figure 13 and Figure 14 demonstrate that as the max_depth increases, the model's RMSE decreases and R^2 increases. Once max_depth reaches approximately 20, the model's performance stabilizes. Additionally, the model achieves its best performance when min_samples_leaf is set to 10, and further increasing this value does not yield improvements in model performance. In the experiment, parameters such as splitter, criterion, and min_samples_split have minimal influence on model performance. Consequently, the optimized parameters are max_depth set to 22 and min_samples_leaf set to 10. This parameter combination can serve as a reference for training subsequent Random Forest models.

The specific results are recorded in Tab. 3, indicating that the Decision Tree with the optimized parameters performs well in predicting both "FSO_Att" and "RFL_Att", which is validated by tests on the validation set, showing the model's robust generalization capabilities.

3.4 Generic Random Forest Models

3.4.1 The Framework of Generic Model

The Generic Random Forest Model, referred to as the generic model, is designed to predict channel attenuation across all weather conditions.

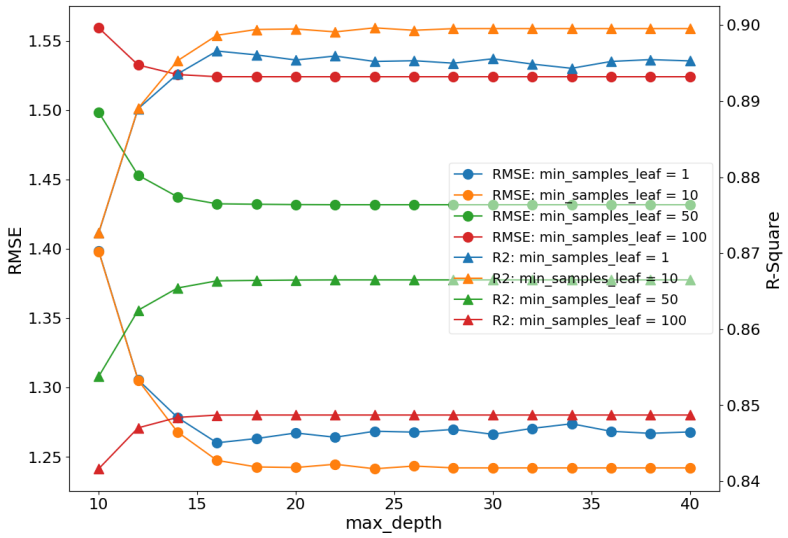


Figure 13: Learning Curve of Decision Tree Regression Model on FSO Training Set

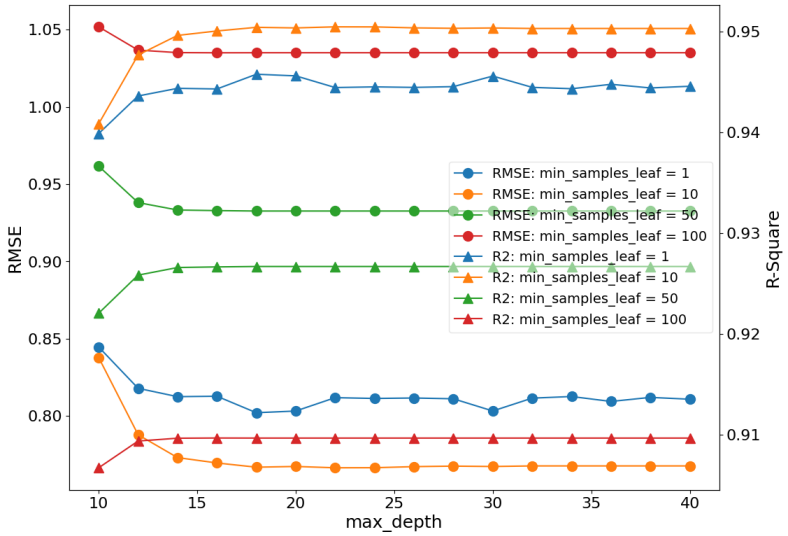


Figure 14: Learning Curve of Decision Tree Regression Model on RFL Training Set

Table 3: Decision Tree Model Metrics

Data set	RMSE	R^2
training_FSO	1.24	90%
validation_FSO	1.19	90.5%
training_RFL	0.77	95%
validation_RFL	0.73	95.5%

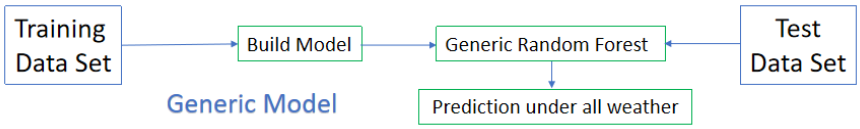


Figure 15: The Framework of Generic Model

Table 4: Hyperparameters Tuning of RF Model for FSO and RFL

Hyperparameter	Coarse Tune	Fine Tune(FSO)	Fine Tune(RFL)
n_estimators	range(10,301,20)	[120,130,140]	range(100,141,10)
max_depth	range(5,31,5)	range(27,33,1)	range(23,28,1)
min_samples_leaf	[4,5,6,7]	[1,2,3,4]	[1,2,3,4]
min_samples_split	[5,6,7,8]	[2,3,4,5]	[2,3,4,5]

This model is trained using the entire training dataset, ensuring its versatility and robustness. Figure 15 illustrates the framework of generic model.

3.4.2 Generic Model Establishment

During the coarse tuning of hyperparameters for FSO prediction, Figure 16 and Figure 17 demonstrate a clear trend that as the value of n_estimators increases, the RMSE decreases and R^2 increases slightly across both the training and validation sets. However, beyond a certain point, specifically when n_estimators reaches 130, further increments fail to enhance the model's performance in either dataset.

Regarding max_depth, optimal performance is achieved when it is set to 30. However, the improvement in model performance when max_depth is 30 compared to when it's 20 is marginal. Based on these observations, the fine tuning ranges will be outlined in Tab. 4.

In the fine-tuning process, as illustrated in Fig. 18, it is evident that the model's performance improves in the training set as the number of n_estimators and max_depth increase. However, in the validation set, the optimal configuration occurs when n_estimators is set to 130 and max_depth is 30. Beyond this point, further increases lead to a decline in model performance, indicating overfitting. Additionally, it is observed in Fig. 19 that reducing min_samples_leaf continues to enhance model performance until this parameter reaches a value of 1.

In the coarse tuning phase of the RFL model, as Fig. 20 and Fig. 21 it was determined that the optimal hyperparameter configuration occurred when n_estimators was set to 100, max_depth to 25, and min_samples_leaf to 4, resulting in the best model performance.

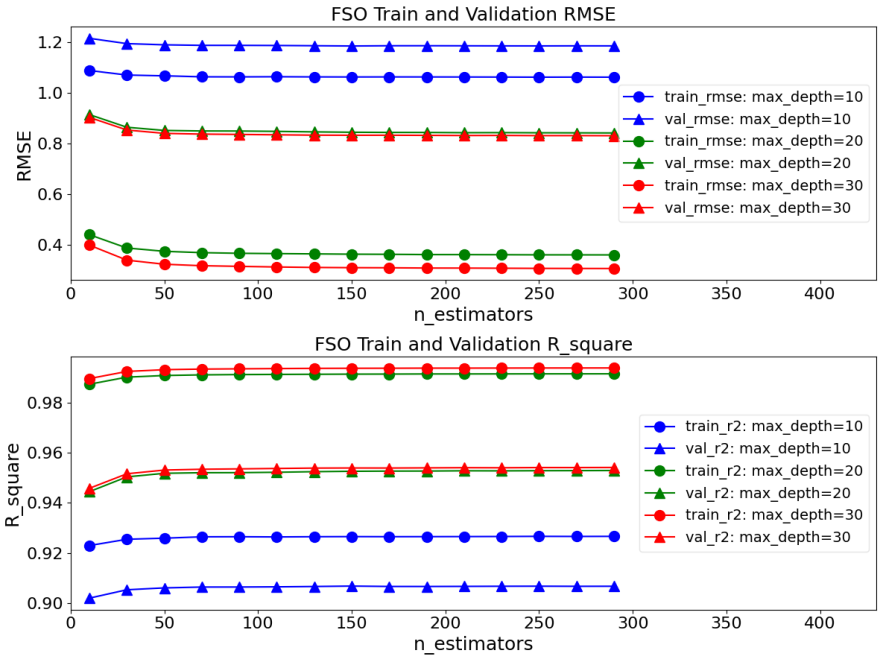


Figure 16: FSO Coarse Learning Curve of Random Forest Regression Model with Respect to n_estimators and max_depth

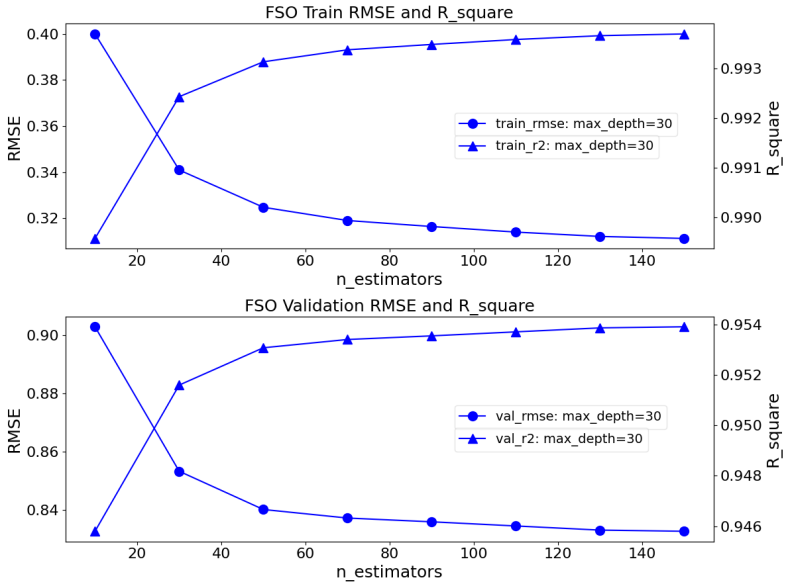


Figure 17: FSO Coarse Learning Curve of Random Forest Regression Model with Respect to n_estimators and max_depth = 30

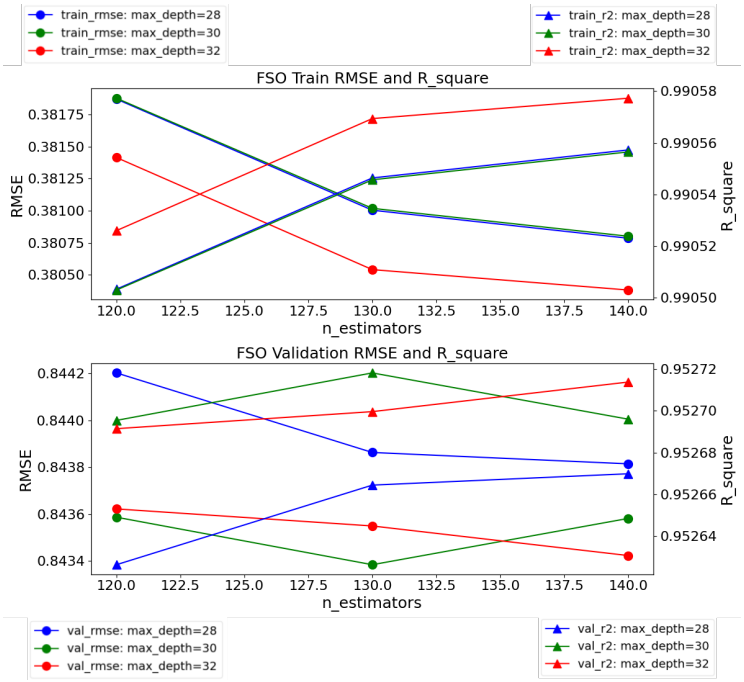


Figure 18: FSO Fine Learning Curve of Random Forest Regression Model with Respect to n_estimators and max_depth

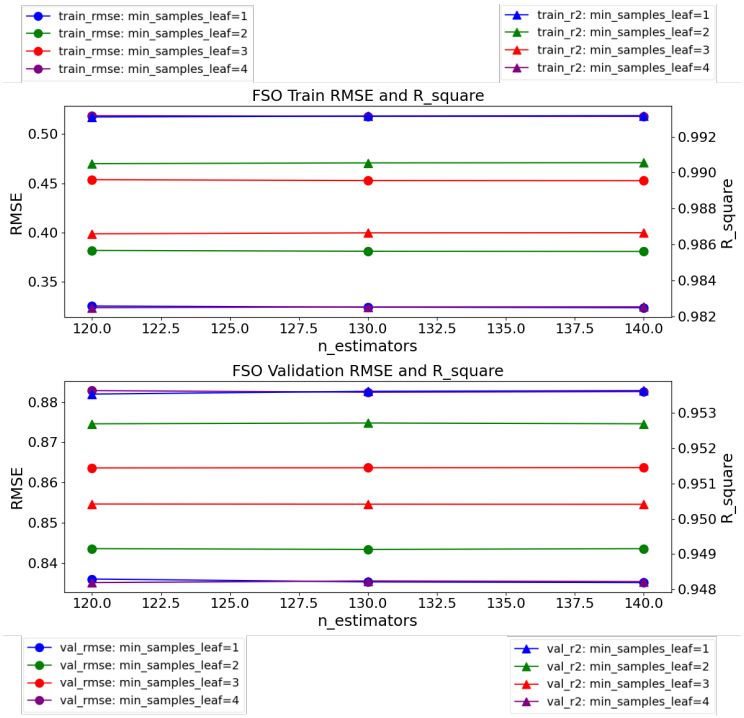


Figure 19: FSO Fine Learning Curve of Random Forest Regression Model with Respect to n_estimators and min_sample_leaf

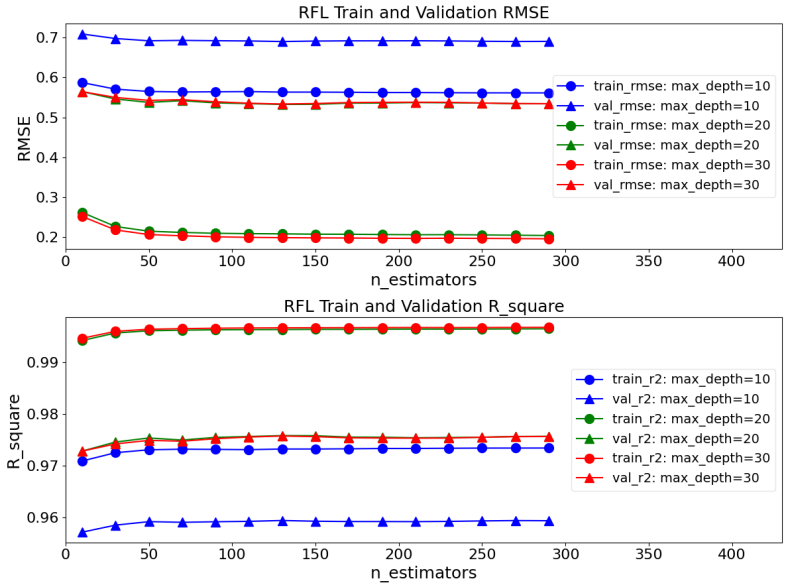


Figure 20: RFL Coarse Learning Curve of Random Forest Regression Model with Respect to n_estimators and max_depth

Table 5: Optimal Hyperparameters for Models

Hyperparameter	FSO Model	RFL Model
n_estimators	130	110
max_depth	30	28
min_samples_leaf	1	1
min_samples_split	2	2

During the fine-tuning process of the RFL model, depicted in Fig. 22, it became apparent that the model's performance in the training set improved with increasing values of n_estimators and max_depth. Yet, in the validation set, the optimal parameters were identified as n_estimators being 110 and max_depth being 28. Further escalation of these parameters led to a decline in performance, signaling overfitting. Additionally, Figure 23 illustrated that reducing min_samples_leaf improved model performance, with the optimal value transitioning from 4 to 1.

Finally, Table 5 presents the optimal hyperparameters for both models, while Table 6 displays the performance of both models on the training, validation, and test datasets.

3.4.3 Features Importance and Pruning of Generic Model

As discussed in Section 3.3.3, during Random Forest training, values are collected to measure how, on average, the node split decreases the impu-

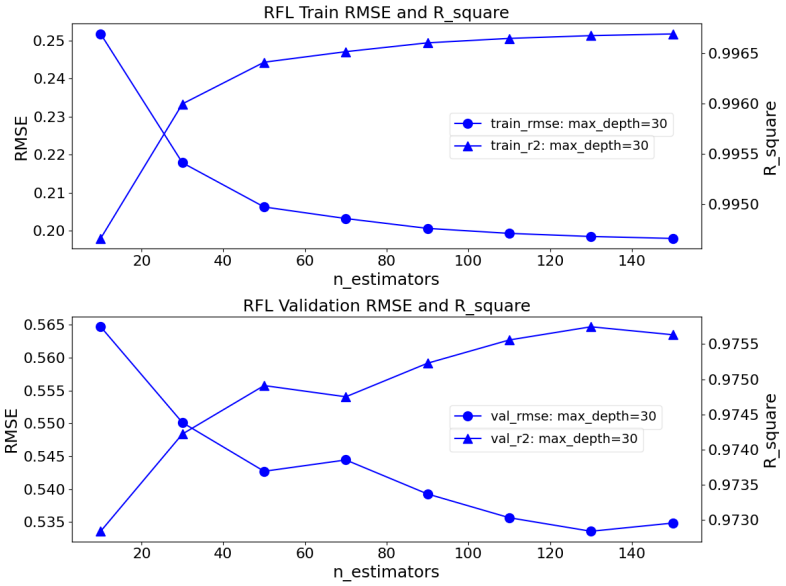


Figure 21: RFL Coarse Learning Curve of Random Forest Regression Model with Respect to n_estimators and max_depth = 30

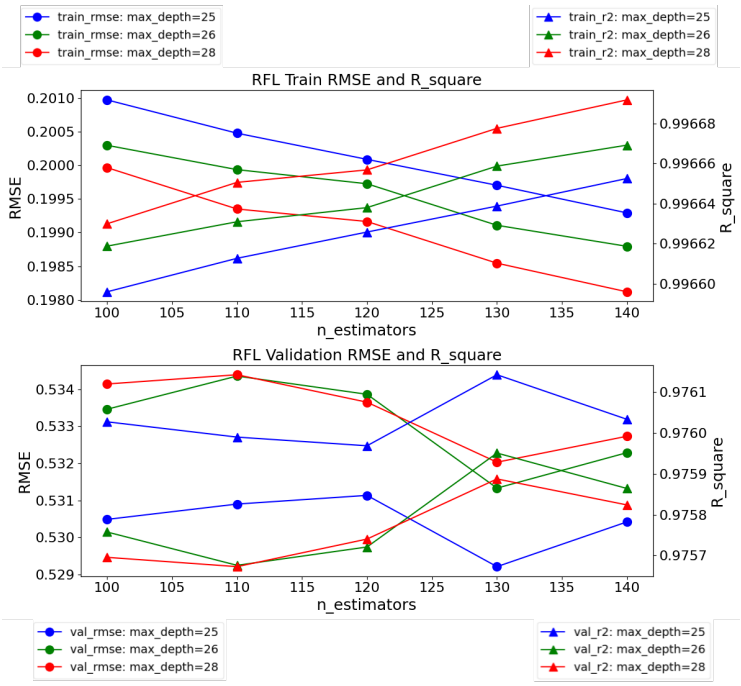


Figure 22: RFL Fine Learning Curve of Random Forest Regression Model with Respect to n_estimators and max_depth

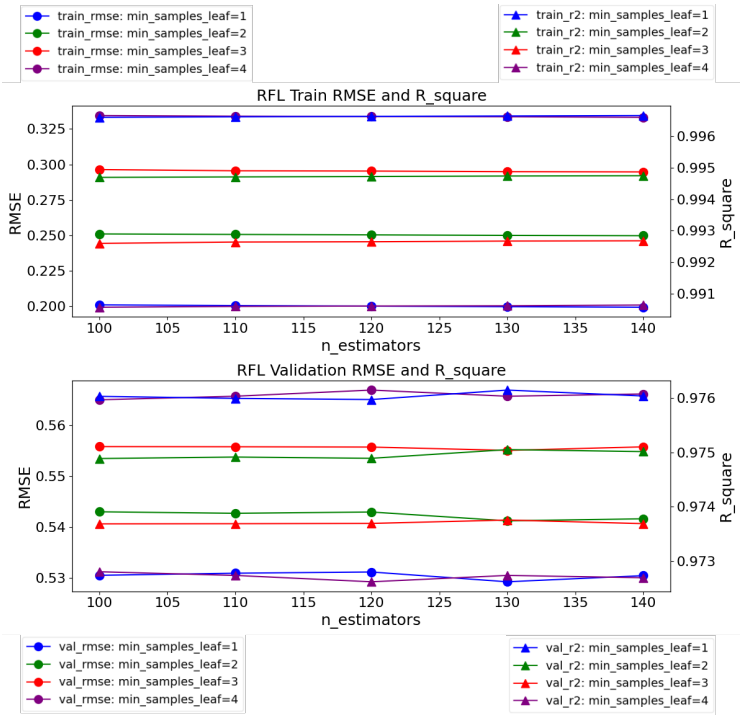


Figure 23: RFL Fine Learning Curve of Random Forest Regression Model with Respect to n_estimators and min_samples_leaf

Table 6: Metrics of Optimal Models

Metrics	FSO Model	RFL Model
training RMSE	0.31	0.20
validation RMSE	0.83	0.54
testing RMSE	0.78	0.50
training R^2	99.4%	99.7%
validation R^2	95.4%	97.6%
testing R^2	95.9%	97.9%

rity or MSE. The average over all trees in the forest yields the measure of feature importance [16]. This method is implemented in scikit-learn's Random Forest [17].The computed importances is the relative values since they are normalized. One significant advantage of this method is its computational speed, all necessary values are computed during Random Forest training. However, in the case of correlated features, it may select one feature over another, potentially leading to incorrect conclusions.

Another method is Permutation Importance [18] that is to conduct a "destructive test" on each feature in the model by randomly altering the order of its values, and then observing the change in model performance to evaluate the contribution of the feature to the model predictions. First, the trained model is used to predict the test data, and the prediction results are recorded as a baseline. Next, for each feature to be evaluated, the order of its values is randomly shuffled, and the model's predictions on the shuffled data are recalculated. The change in model performance on the shuffled data is then computed. The importance of a feature is defined as the degree of change in model performance. If the model performance decreases significantly after shuffling the feature values, it indicates that the feature has a significant impact on the model's performance and thus has high importance; conversely, if the performance change is small, it suggests that the feature has low importance. Permutation Importance provides intuitive interpretability and can offer useful information even in the presence of feature correlations or nonlinear relationships.

The above mentioned methods provide a sorted order of features based on their importance scores. Wrapper methods [19], like Recursive Feature Elimination (RFE), go beyond this by selecting features based on actual model performance. RFE evaluates subsets of features using Random Forests, by iteratively removing features and assessing their impact on model performance. This process involves training the model on the entire feature set and iteratively eliminating the least important features based on their derived feature importances. This iterative procedure continues until the desired number of features is achieved. By directly evaluating their influence on model performance, wrapper methods offer a more comprehensive understanding of feature importance.

Embedded methods [20], on the other hand, incorporate feature selection as an integral part of the model training process. In Random Forests, feature importance is inherently embedded within the model construction, as the splitting criteria at each node are based on feature importance measures.

Embedded methods can select or delete groups of features based on certain thresholds of feature importance scores. Using embedded methods enables the rapid identification of features with significant predictive

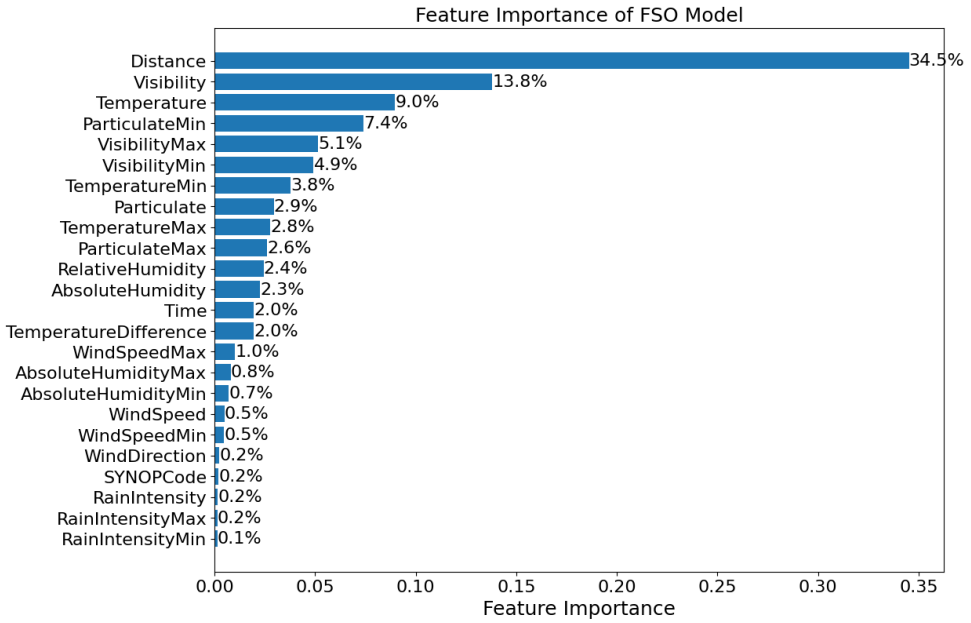


Figure 24: Feature Importance of FSO Model

power. Subsequently, the selected features can be further refined using wrapper methods like RFE. This combined approach maximizes the efficiency of embedded methods and the precision of wrapper methods, allowing for more effective feature selection and modeling, particularly in high-dimensional datasets.

Given that this research dataset comprises 25 features, employing wrapper methods alone should suffice for eliminating redundant features.

Figure 24 and Figure 25 respectively illustrate the features importance of FSO and RFL channels. This study employs the wrapper method to prune features.

The process of feature pruning is iteratively removing features based on their feature importance scores, starting from the least important to the most important, as depicted in Fig. 24 and Fig. 25. Subsequently, the performance of the models is observed. If the models' performance does not exhibit significant changes, pruning occurs. During this process, consideration is given to the interactions between features and their impact on model performance.

The objective is to identify the most relevant subset of features capable of effectively predicting FSO and RFL attenuation while simultaneously minimizing overfitting and computational overhead.

Figure 26 and Figure 27 demonstrate the impact of feature pruning on both generic models. Without compromising predictive performance, the complexity of features is significantly reduced, from 27 features to

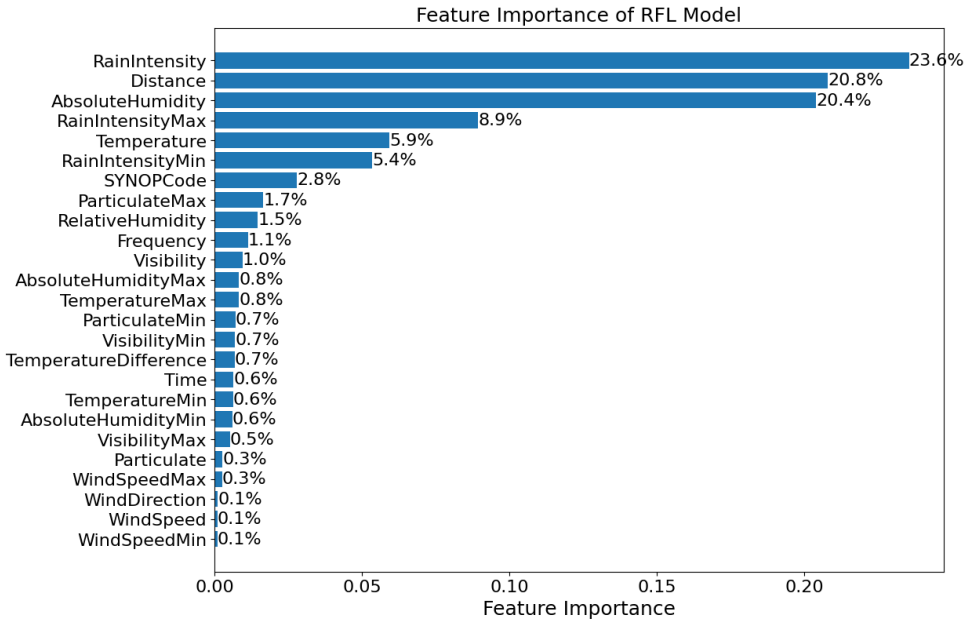


Figure 25: Feature Importance of RFL Model

approximately 14 for the FSO Model or 11 features for the RFL Model, as indicated by the green line in figures. Additionally, Distance and Visibility are the most significant factors in the FSO generic model, while in the RFL generic model, Absolute Humidity and Rain Intensity takes precedence.

Utilizing the pruned features, further hyperparameter tuning is conducted for both models. The coarse tuning and fine tuning processes remain consistent with those outlined in Section 3.4.

Following the descriptions provided in Fig. 28 and Fig. 29, the hyperparameters are determined considering the trade-off between model complexity and performance. The results are presented in Tab. 7. Both of n_estimators and max_depth in optimal hyperparameter with feature pruning are a little larger than that of optimal hyperparameter without feature pruning in Tab. 5.

Table 8 displays the evaluation metrics for the two generic models without and with feature pruning, indicating that the performances of two generic models with feature pruning are almost some as models without feature pruning.

Through feature pruning and hyperparameter tuning, the overall complexity of the two generic random forest models are significantly reduced. This helps improve the interpretability and generalization ability of the models while reducing the risk of overfitting.

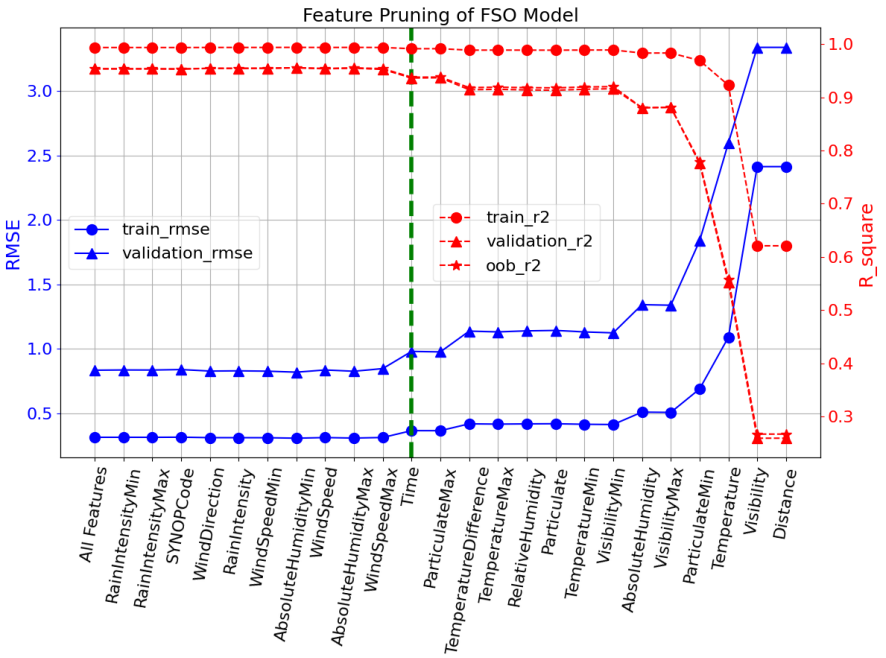


Figure 26: Feature Importance of FSO Model

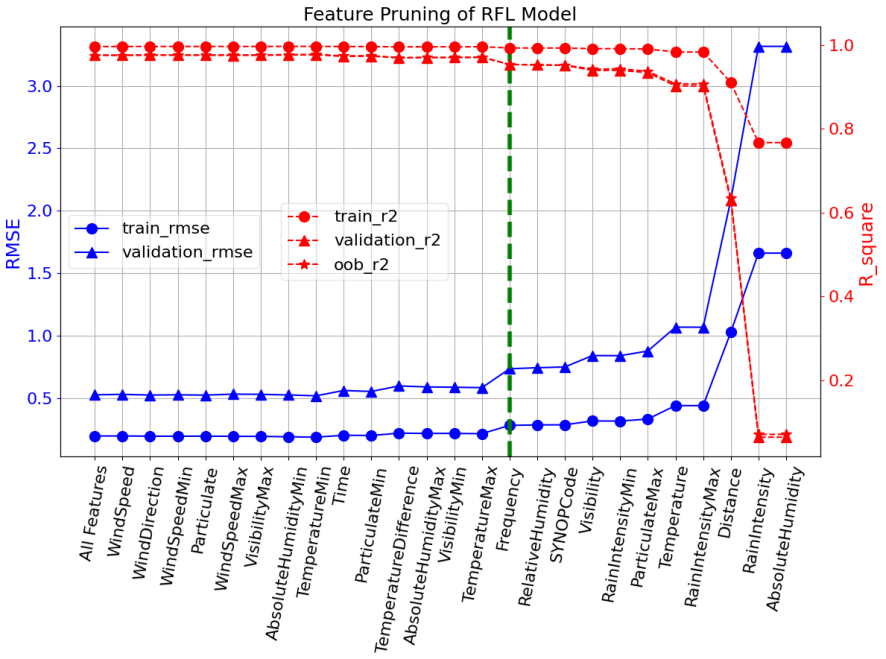


Figure 27: Feature Importance of RFL Model

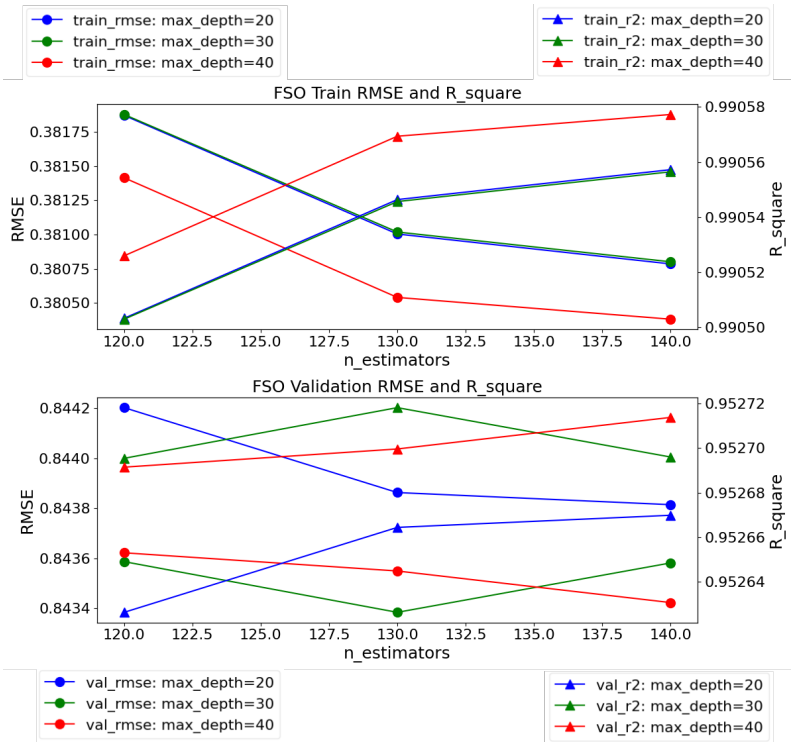


Figure 28: Hyperparameter Tuning for FSO Model with Feature Pruning

Table 7: Optimal Hyperparameters for Models with Feature Pruning

Hyperparameter	FSO Model	RFL Model
n_estimators	150	120
max_depth	34	32
min_samples_leaf	1	1
min_samples_split	2	2

Table 8: Generic Model Without(WO) and With(W) Feature Pruning

Models	RMSE	$\hat{R2}$
FSO (WO))	0.78	95.9%
FSO (W)	0.78	95.9%
RFL (WO))	0.50	97.9%
RFL (W)	0.54	97.5%

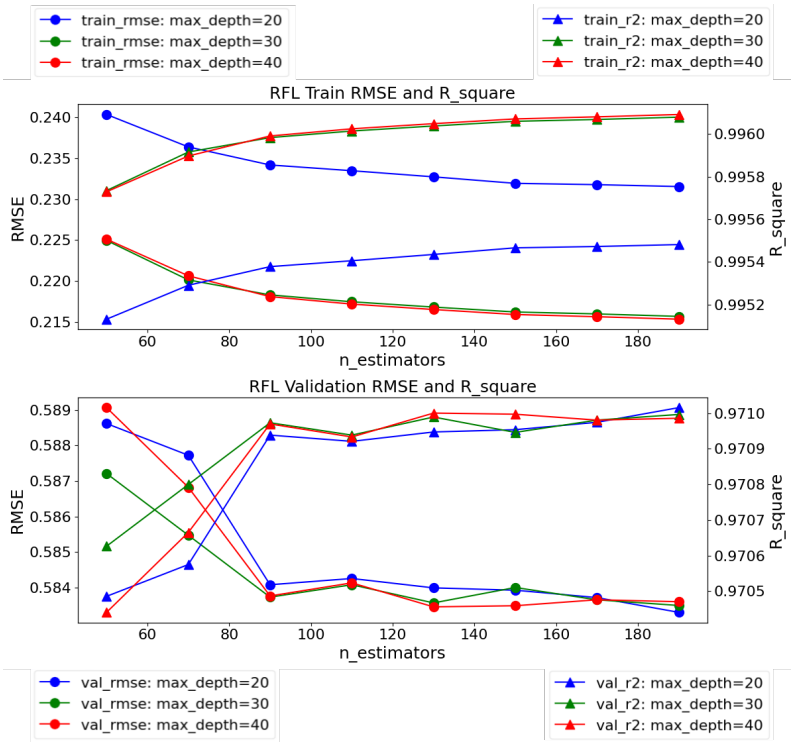


Figure 29: Hyperparameter Tuning for RFL Model with Feature Pruning

3.5 Specific Random Forest Models

3.5.1 The Framework of Specific Model

In contrast to the generic model, specific models are tailored to predict attenuation under particular weather conditions. Each of these models is trained using data specific to one weather scenario, enabling predictions within that context. This study will develop seven specific models, with each model corresponding to a distinct weather condition. The framework is shown as Fig. 30. The hyperparameter tuning process for the specific models is similar with that of the generic model. Therefore, this process will not be repeated in this section.

3.5.2 Feature Importance and Pruning of Specific models

In the original dataset, the data will be divided into subsets based on the values of the feature "SYNOPCode", representing different weather conditions, shown as Tab. 10. Each subset of data corresponding to specific weather conditions will be used to build a random forest model separately. This process will establish seven distinct models to predict the impact of weather on signal attenuation for specific weather scenar-

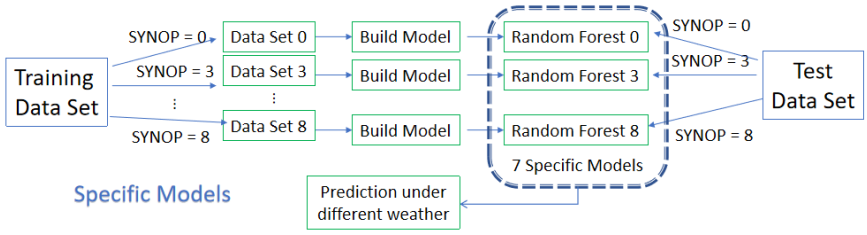


Figure 30: The Framework of Specific Model

Table 9: Top Feature Importance Under Different Weather Conditions

Weather	RFL Channel	FSO Channel
Clear	*DIS(50.4%), AH(24.5%)	DIS(47.3%), Temp(10.1%)
Dust	Time(22.2%), AH(18%), AHmax(16.4%)	DIS(22.3%), VMmin(16.3%), RH(14.6%)
Fog	RH(40.4%), AH(14.3%), RI(11.1%)	PM(26.8%), PMmin(25.8%), PMmax(20.2%)
Drizzle	AH(47.5%), AHmax(23.1%)	DIS(34%), PMmax(12.4%)
Rain	RImax(25%), AH(16.3%), RI(14.4%)	DIS(26.3%), PMmax(12.4%)
Snow	WSmax(30.9%), Tdiff(19.2%), Time(15.5%)	PM(25.4%), PMmax(25.3%), PMmin(15.4%)
Showers	AH(17.0%), AHmin(13.5%), RI(12.2%)	VMmin(28.3%), DIS(27.8%)

* DIS: Distance, AH: AbsoluteHumidity, RH: RelativeHumidity, Temp: Temperture, VM: Visibility, PM: Particulate, RI: RainIntensity, WS: WindSpeed, Tdiff: TemperatureDifference.

ios. Table 9 illustrates the top important features under specific weather conditions.

Figure 31 and Figure 32 depict the performance of each specific model based on different weather conditions. Each specific model has its own set of important predictors, indicating that certain features are more effective in predicting attenuation under specific weather conditions.

Specifically, when the weather is rainy or showers, the RFL model requires additional features to maintain prediction performance at a stable level. The RMSE in rainy or shower conditions is higher compared to other weather conditions, indicating a significant impact of rainy weather on the RFL channel. In rainy weather, the RMSE of FSO channel attenuation resembles that of RFL channels. However, the attenuation in the FSO channel is typically higher than in other weather conditions, particularly in dusty weather, where the FSO model performs poorly.

Table 10: Feature Pruning of Specific Models under Different Weather

SYNOP	Weather	*Amount of Important Feature (RFL/FSO)
0	Clear	10/9
3	Dust	13/5
4	Fog	9/9
5	Drizzle	12/5
6	Rain	17/10
7	Snow	6/6
8	Showers	15/8

* The Amount column indicates the number of important features required by the specific models under different weather conditions. The first number represents the amount in the RFL channel, while the second number represents the amount in the FSO channel.

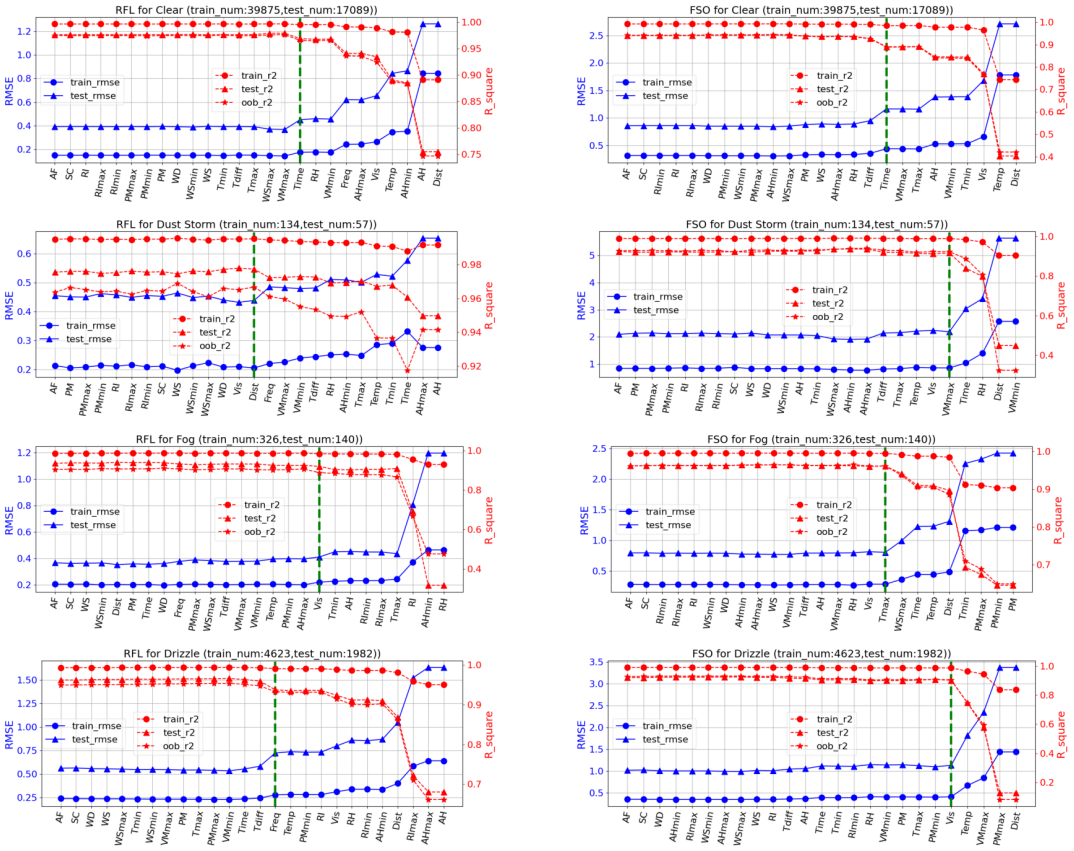


Figure 31: Predictor Importance for Clear, Dust, Fog, Drizzle

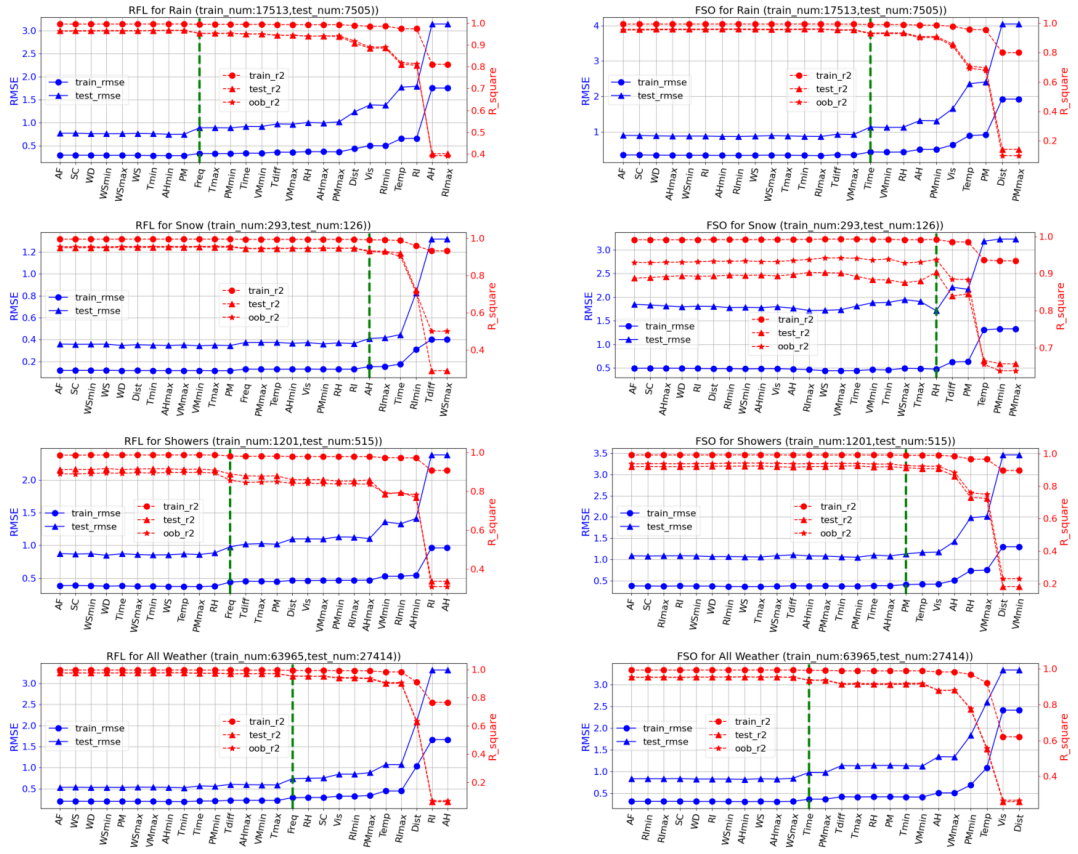


Figure 32: Predictor Importance for Rain, Snow, Showers and All Weather

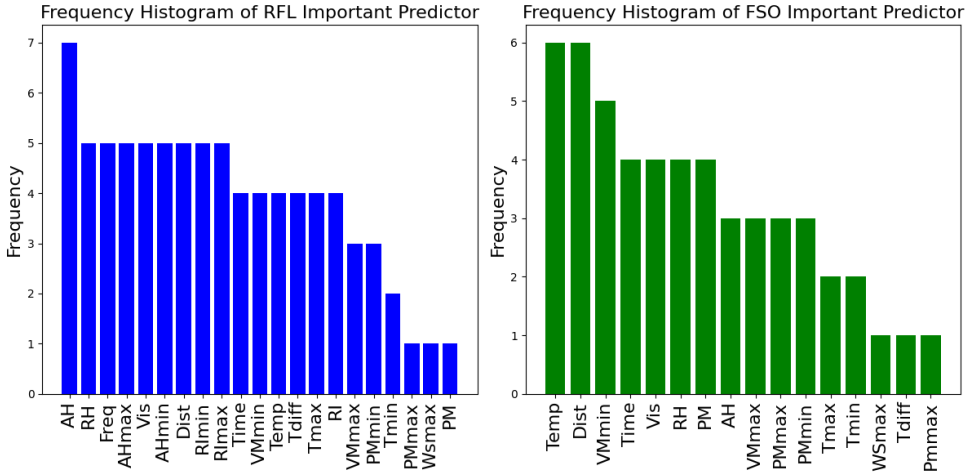


Figure 33: Important Predictor Bar Chart of All Specific Models

Additionally, Table 10 presents the number of important features required by the specific models under different weather conditions, which indicates that T specific models have different important features.

Figure 33 illustrates the frequency of important predictors across all specific models. In the RFL channel, Absolute Humidity is utilized in all the special models and features related to Rain Intensity, Relative Humidity, Frequency, Distance, and Visibility are commonly utilized for prediction in various weather conditions, while in the FSO channel, features related to Distance, Temperature, and Visibility are commonly applied for prediction in different weather environments. The consistency between the frequency of important predictors in specific models and the important predictors identified in the generic model demonstrates the reliability of building prediction models based on these important predictors.

3.6 Hybrid Random Forest Models

3.6.1 The Framework of Hybrid Model

In the first stage of the hybrid model, random forest 1 is trained using all weather features. The first stage have the same processing as the generic model establishment, but the difference is that the first predicted channel attenuation is then used as part of the training set for the second stage, where random forest 2 is trained to predict the attenuation of the other channel, shown as Fig. 34. This study assumes the implementation of two hybrid methods: Hybrid Method 1 (M1) involves using predicted RFL_Att. to forecast FSO_Att., while Hybrid Method 2 (M2) entails using predicted FSO_Att. to predict RFL_Att.

This approach maximizes the utilization of the predicted attenuation from one channel. However, errors from the first stage predictions may propagate to the second stage, potentially affecting the accuracy of the attenuation prediction in the second stage.

3.6.2 PCC Metrics in Hybrid Model

Figure 35 describes the distribution of True and Predicted pair of RFL_Att. and FSO_Att. in different weather conditions.

From Fig. 36, there is essentially no significant linear correlation between RFL_Att. and FSO_Att. in the overall sample. However, under specific weather conditions, they exhibit a certain degree of linear correlation. In dusty environments, the attenuation of the two channels shows a moderate negative linear correlation, whereas in rainy conditions, the attenuation exhibits a moderate positive linear correlation. In foggy and

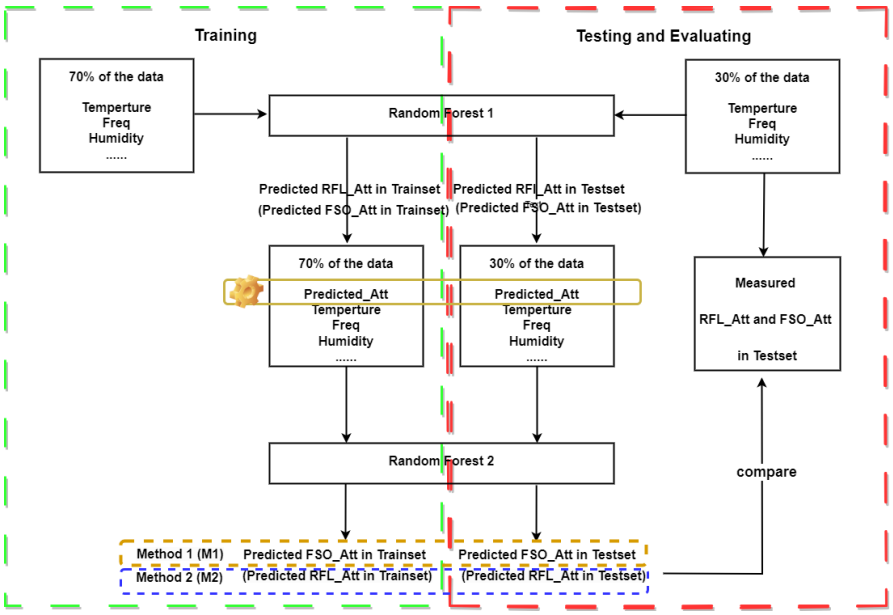


Figure 34: The Framework of Hybrid Model

snowy conditions, the attenuation of both channels shows a strong positive linear correlation, while the weak linear correlation in clear, drizzle, and shower conditions does not provide useful information.

Despite the strong positive correlation observed in foggy, snowy, and rainy conditions, it implies that the attenuation of both channels tend to increase or decrease together under the same weather conditions, only in dusty conditions do the channels exhibit a moderate negative linear correlation, which suggests that when the communication signal of one channel deteriorates significantly, the other channel might maintain better communication quality. However, this moderate negative linear correlation is not sufficient to definitively establish a compensatory relationship. In conclusion, while the hybrid model can capture the real attenuation correlation between the two channels, this correlation does not provide a definitive conclusion on which channel, RFL or FSO, is more suitable under most specific weather conditions.

Figure 37 illustrates that different frequencies of RFL_Att. exhibit varying degrees of linear correlation under dusty conditions. Specifically, the 83.5GHz RFL_Att. demonstrates a stronger negative linear correlation with FSO_Att. compared to the 78.5GHz RFL_Att. This suggests that the 83.5GHz RFL channel is a better choice in dusty weather conditions.

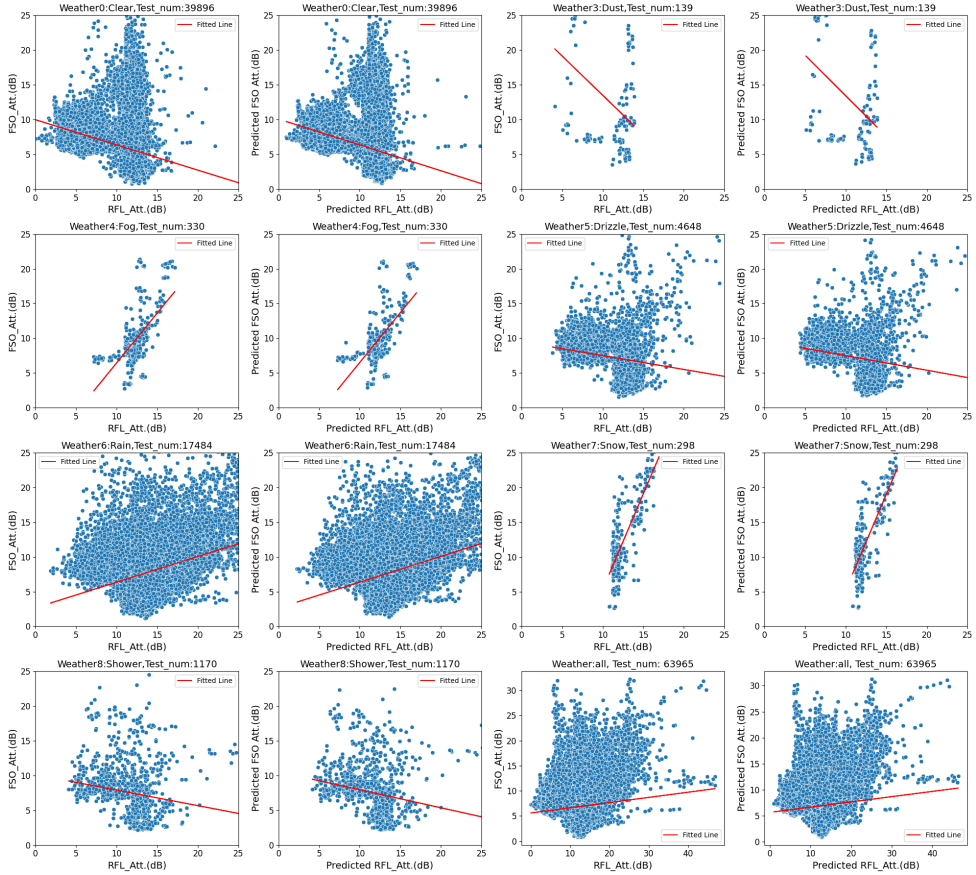


Figure 35: The Distribution of True and Predicted pair of RFL_Att. and FSO_Att. by SYNOPcode

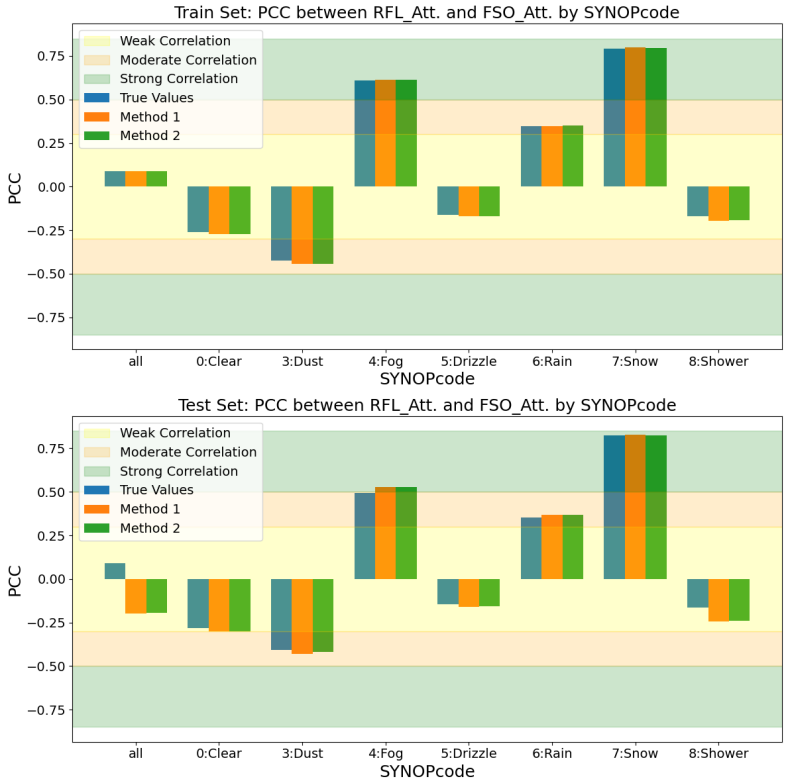


Figure 36: The PCC of RFL_Att. and FSO_Att. by SYNOPcode

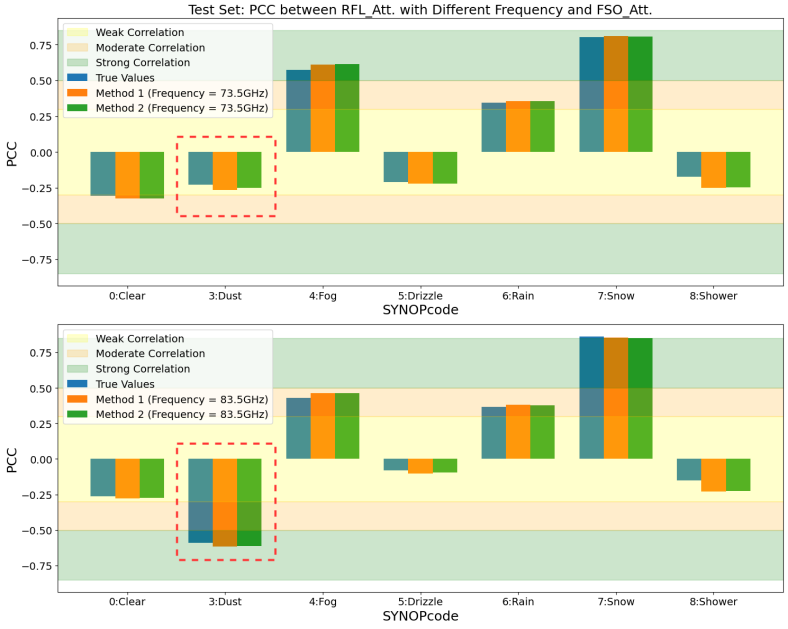


Figure 37: PCC between RFL_Att. with Different Frequency and FSO_Att.

3.6.3 The MI and NMI Metrics in the Hybrid Models

Figure 38 displays the joint probability distribution of RFL_Att. and FSO_Att. across different weather conditions, with highlighted points indicating areas of high pair value density. This figure uses a bin interval of 1 dB for better illustration. However, in the subsequent calculation of MI and NMI, this study employs a bin interval of 0.1 dB for better exploring the correlation between two channels.

Figure 39 quantifies the MI between RFL_Att. and FSO_Att..The MI is notably higher in dusty, foggy, snowy, and shower conditions compared to other weather conditions, suggesting that the correlation between RFL_Att. and FSO_Att. may provide more useful information for mutual prediction under these four conditions.

Figure 40 further illustrates that the correlation between RFL_Att. and FSO_Att was strong under dusty, snowy and foggy conditions, demonstrating that the hybrid models can effectively capture the correlation between the two channels.

3.6.4 The Feature Importance in Hybrid Model

Figure 41 presents that the predicted RFL_Att. only contributes 4.8% to the feature importance for predicting FSO_Att., and the predicted FSO_Att. contributes 4.8% to the feature importance for predicting RFL_Att in hybrid model. This is a sign that one channel's attenuation is not an effective predictor for predicting another channel's attenuation.

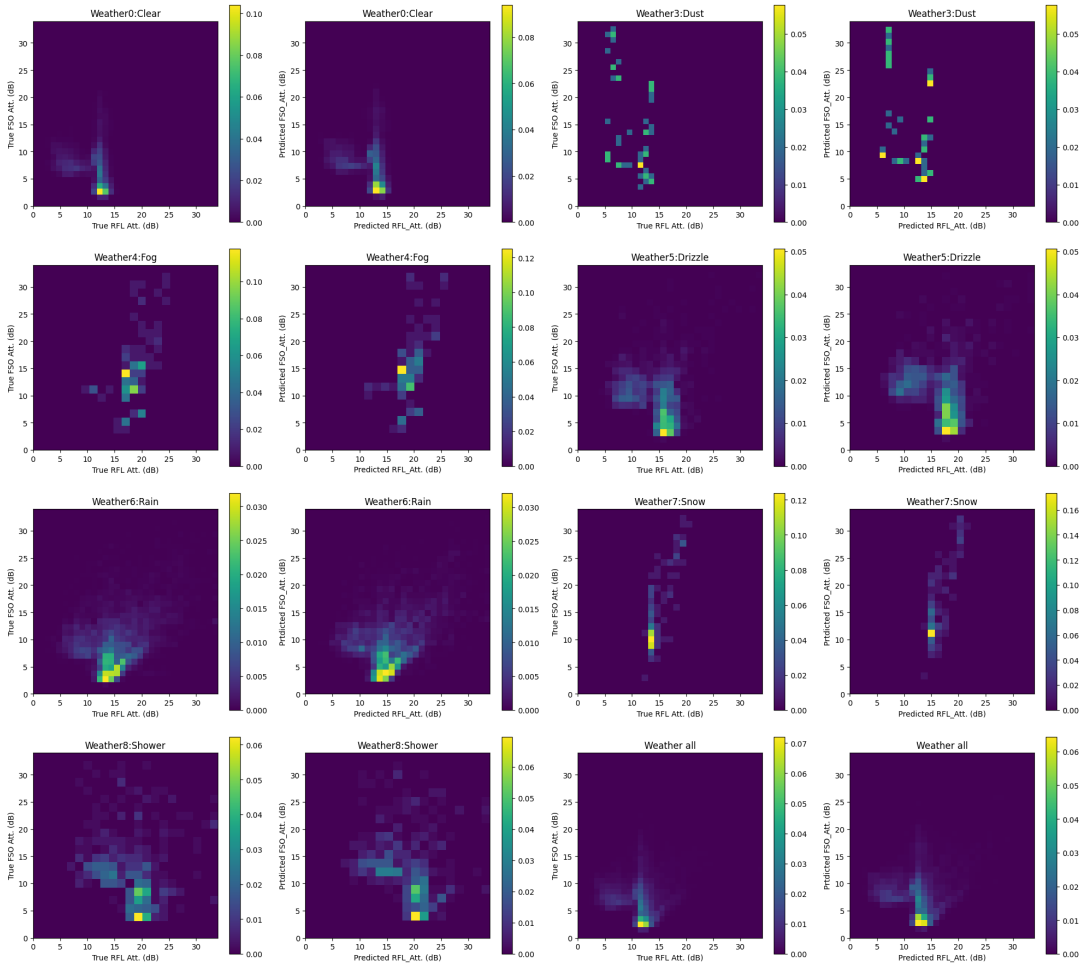


Figure 38: The Joint Probability Distribution of RFL_Att. and FSO_Att.

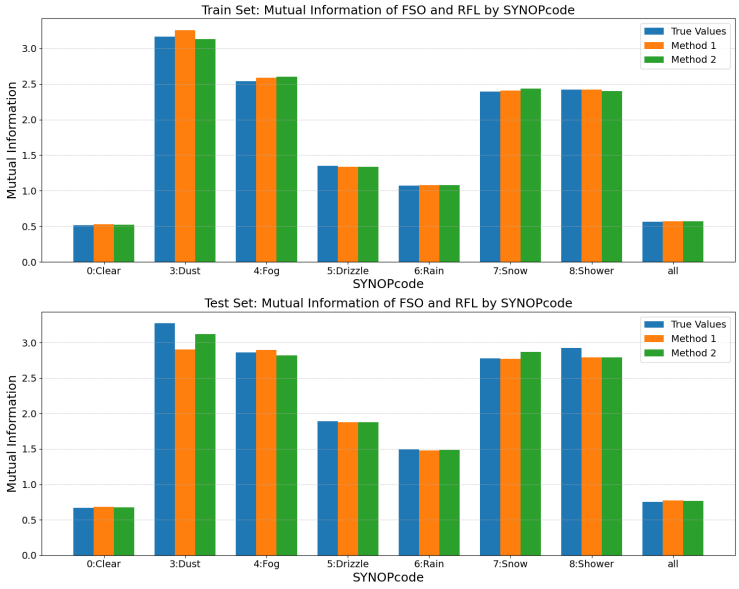


Figure 39: MI of FSO_Att. and RFL_Att. (bin_interval = 0.1)

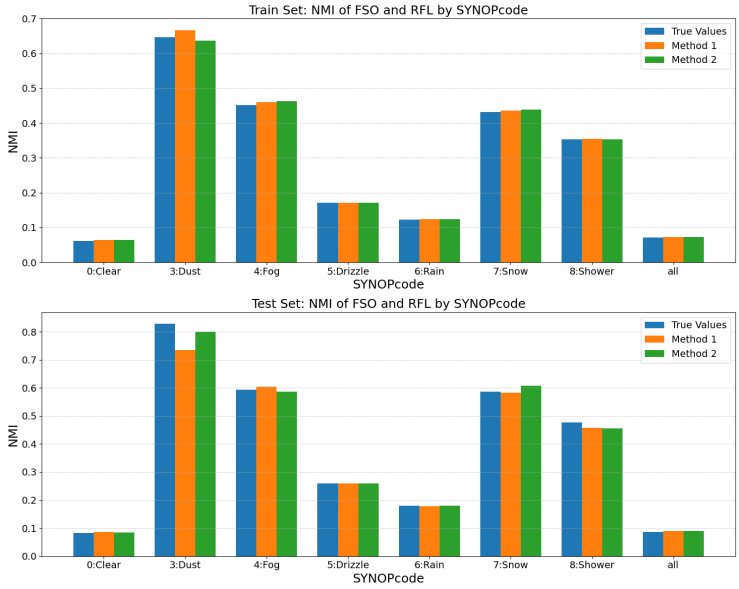


Figure 40: NMI of FSO_Att. and RFL_Att. (bin_interval = 0.1)

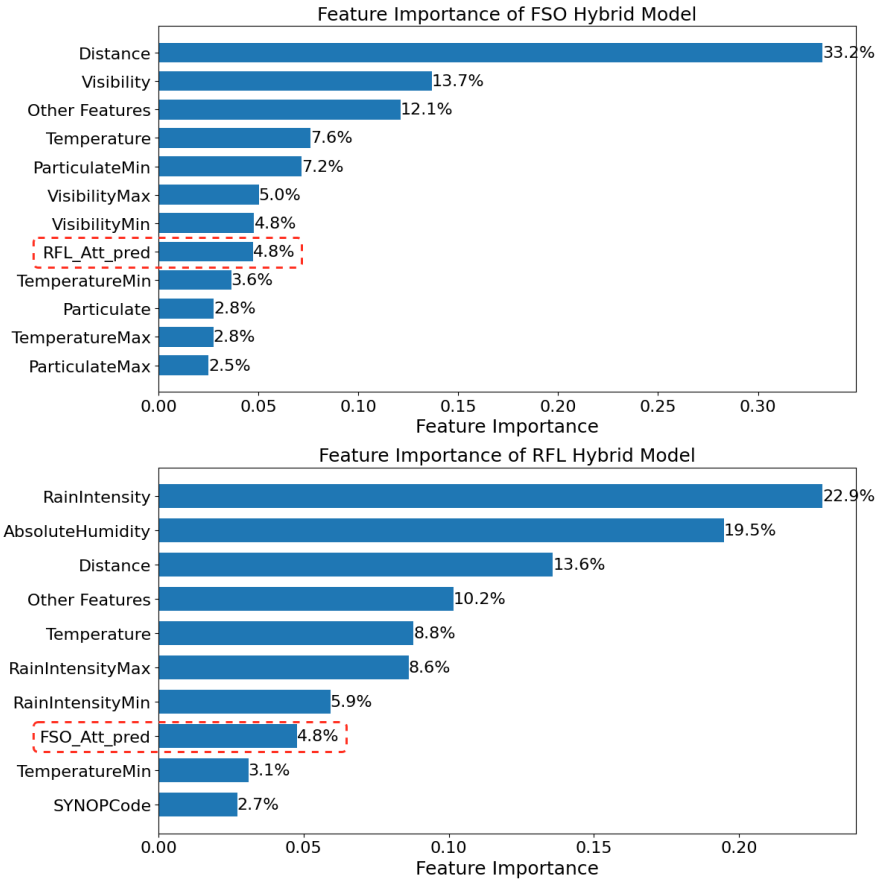


Figure 41: The Feature Importance in Hybrid Model

4 Results

4.1 The Comparison of Results between Specific Models and Generic Models

As shown Fig. 42, the performance of the specific RFL model is slightly better under clear and snowy conditions. This is because half of the training data is collected under clear conditions, resulting in the special RFL model being able to better explore the relationship between features and predictive attenuation in the RFL channel. On the other hand, the most important predictor of Maximum Wind Speed under snowy conditions is only relevant for predicting under snow conditions and do not contribute to predictions in other weather conditions. Under other weather conditions, the performance of generic RFL model is better than special RFL model.

In the FSO channel, the generic model outperforms the specific model under dusty and snowy conditions. This is due to the small size of the training set in such weather conditions, making the specific model unreliable. However, the predicted results are good in the generic model under dust and fog conditions, possibly due to data from other environments contributing to predicting attenuation under these conditions. In other weather conditions, the performance of generic FSO model is better than special FSO models.

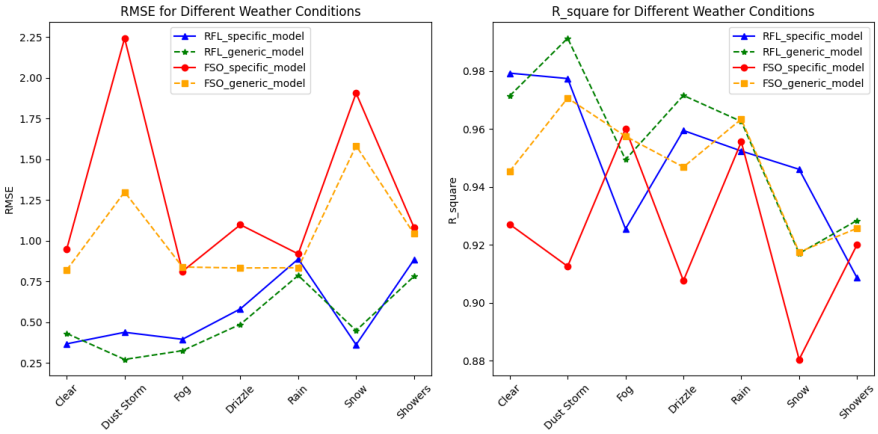


Figure 42: Comparison of Specific and Generic Model

4.2 The Result in the Hybrid Models

As mentioned in Section 3.3.4, mentioned, PCC can only capture the linear correlation, while NMI can explore the correlation between RFL_Att.

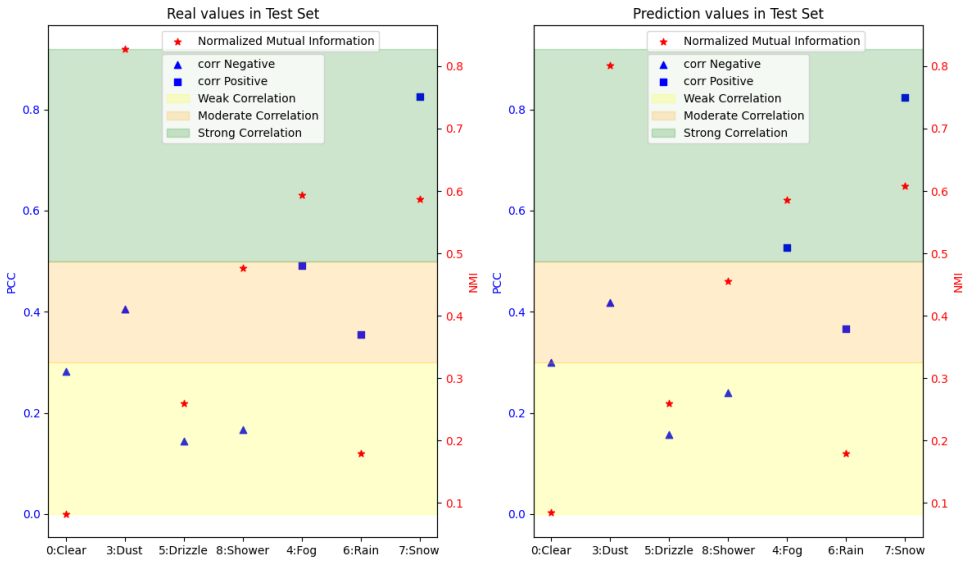


Figure 43: The linear and nonlinear correlation of FSO_Att. and RFL_Att.

and FSO_Att. regardless of whether it is linear or nonlinear. Figure 43 highlights the true and predicted PCC and NMI in the testing dataset. It demonstrates that Hybrid Models can effectively capture both PCC and NMI. Specifically, NMI is higher than PCC under conditions of dust, fog, and showers, indicating that RFL_Att. and FSO_Att. may exhibit more nonlinear or complex relationships in these weather conditions, as NMI can capture more shared information, resulting in higher values. However, under snow conditions, NMI is notably lower than PCC, which suggests that RFL_Att. and FSO_Att. likely have a primarily linear relationship, and in this case, PCC more effectively captures this linear correlation. Additionally, both PCC and NMI are low under clear, drizzle, and rain conditions, which means that RFL_Att. and FSO_Att. do not have significant associations in these weather conditions.

4.3 The Comparison of Results between Hybrid Models and Generic Model

Figure 44 illustrates that, in most cases, the performance of the generic model is better than the hybrid model, even though the hybrid model incorporates the information of the correlation between FSO_Att. and RFL_Att. This also demonstrates that the attenuation prediction of one channel is less important for predicting the attenuation of another channel.

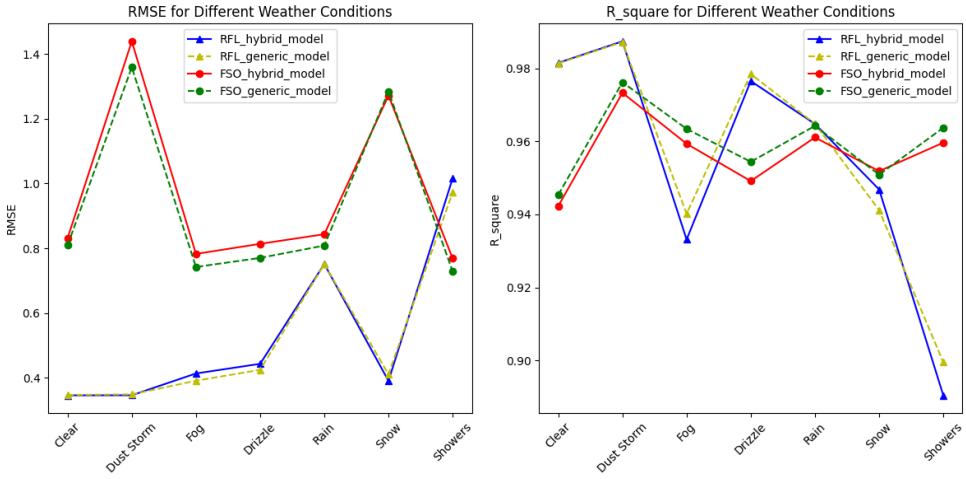


Figure 44: Comparison of Hybrid and Generic Model

Table 11: Model Metrics: Generic (S) vs. Hybrid (M1, M2)

Metrics	FSO (S)	RFL (S)	FSO (M1)	RFL (M2)
training RMSE	0.31	0.20	0.32	0.20
testing RMSE	0.78	0.50	0.84	0.52
training R^2	99.4%	99.7%	99.3%	99.7%
testing R^2	95.9%	97.9%	95.3%	97.7%

From Tab. 11, the performance of the hybrid model is comparable to that of the separate models. This indicates that the predicted attenuation of the first channel does not provide additional useful information for predicting the attenuation of the second channel.

4.4 The Comparison of Three Models

In this section, the hybrid models are trained under various weather conditions, similar to the specific models. The hybrid models generally outperform the specific models across almost all weather conditions. This superior performance is attributed to the hybrid model's use of an additional channel's attenuation as a new predictor for predicting another channel's attenuation. However, with the exception of the hybrid model under foggy conditions and the specific RFL model under snowy conditions, the generic model outperforms all other models in most scenarios.

Considering both model performance and complexity, generic models provide the best overall performance across both RFL and FSO channels, delivering more accurate predictions under most weather conditions. Hybrid models effectively capture the correlation between RFL

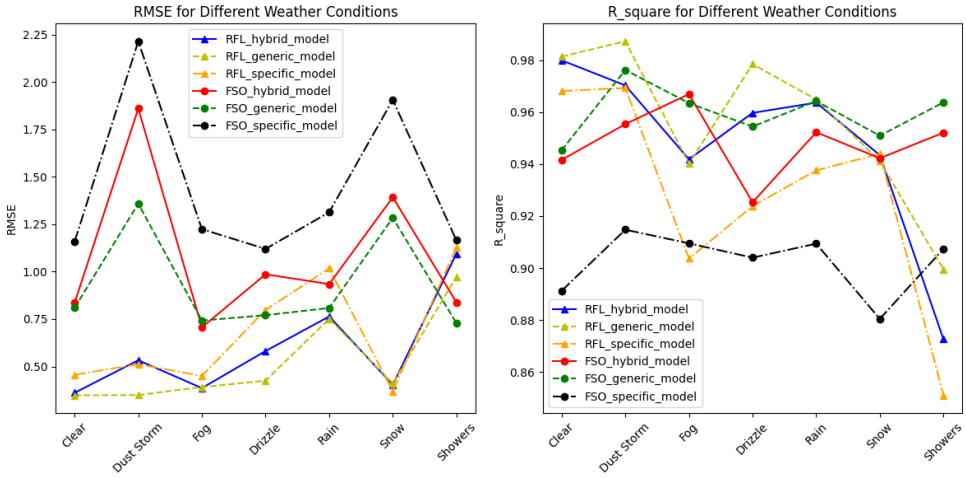


Figure 45: Comparison of Three Models

Table 12: RMSE Metrics for All Models

RMSE: dB	Clear	Dust	Fog	Drizzle	Rain	Snow	Showers
generic FSO	0.81	1.36	0.74	0.77	0.81	1.28	0.73
specific FSO	1.16	2.22	1.22	1.12	1.31	1.91	1.17
hybrid FSO(M1)	0.83	1.45	0.78	0.81	0.84	1.26	0.78
*hybrid FSO(M1)	0.84	1.86	0.71	0.99	0.93	1.39	0.84
generic RFL	0.35	0.35	0.39	0.42	0.75	0.41	0.97
specific RFL	0.46	0.51	0.45	0.80	1.02	0.37	1.13
hybrid RFL(M2)	0.34	0.34	0.30	0.44	0.75	0.39	1.00
*hybrid RFL(M2)	0.36	0.53	0.39	0.58	0.76	0.40	1.10

*hybrid FSO(M1) and *hybrid RFL(M2) were trained based on different weather conditions, indicating that each *hybrid FSO(M1) and *hybrid RFL(M2) can just predict attenuation under one specific weather condition.

and FSO channel attenuation, enabling them to surpass specific models. In conclusion, generic models are the most effective, followed by hybrid models, with specific models ranking last. Figure 45, Table 12 and Table 13 summarize the performance of all models.

Table 13: $R\hat{2}$ Metrics for All Models

$R\hat{2}$: %	Clear	Dust	Fog	Drizzle	Rain	Snow	Showers
generic FSO	94.5	97.6	96.3	95.4	96.4	95.1	96.4
specific FSO	89.1	91.5	90.9	90.4	90.9	88.0	90.7
hybrid FSO(M1)	94.2	97.3	96.0	95.0	96.1	95.2	95.9
*hybrid FSO(M1)	94.2	95.5	96.7	92.5	95.2	94.2	95.2
generic RFL	98.1	98.7	94.0	97.8	96.5	94.1	89.9
specific RFL	96.8	96.9	90.4	92.4	93.8	94.4	85.1
hybrid RFL(M2)	98.2	98.8	93.9	97.6	96.5	94.6	89.3
*hybrid RFL(M2)	98.0	97.0	94.2	96.0	96.4	94.3	87.3

5 Conclusion

This study comprehensively explored weather-induced channel attenuation in hybrid RFL/FSO communication systems, employing ensemble algorithms to establish predictive models for attenuation prediction. Through extensive data preprocessing, EDA, and model establishment, this research developed generic, specific and hybrid models capable of accurately predicting channel attenuation under various weather conditions.

The analysis emphasized the significance of feature pruning and hyperparameter tuning in optimizing model complexity and performance. Through these processes, the generic random forest model has been streamlined, reducing its reliance on 25 features to 14 for the FSO channel and 11 for the RFL channel. Despite removing some redundant features, the performance in both channel models remains basically unchanged. This reduction in complexity has significantly alleviated computational overhead while simultaneously enhancing the interpretability and generalization ability of the models.

Furthermore, the investigation revealed the pivotal role of weather factors in influencing channel attenuation. By building sepcific models for specific weather conditions and generic models encompassing diverse environments, the study demonstrated that in the RFL channel, features related to Absolute Humidity, Rain Intensity, Relative Humidity, Frequency, Distance, and Visibility are frequently employed for prediction across diverse weather conditions. Conversely, in the FSO channel, features associated with Distance, Temperature, Visibility, and Particulate are commonly utilized for prediction across various weather environments. In the RFL channel, features related to water, such as Absolute Humidity, Relative Humidity, and Rain Intensity, are often the most significant, indicating their crucial role in predicting RFL attenuation. In contrast, the FSO channel is primarily impacted by Distance, Visibility, and Particulate Matter. Notably, under foggy and snowy conditions, Particulate, PMmin, and PMmax together account for up to 70% of feature importance, suggesting that particulate matter in the air is a key predictor of FSO attenuation in these conditions.

Additionally, the generic model excels in predicting channel attenuation across diverse weather conditions. In snowy conditions, the specific model outperforms the generic model in the RFL channel, while the hybrid model performs better than the generic model in the FSO channel. Under clear and foggy conditions, the hybrid model excels in the RFL channel. However, in other conditions, the generic model outperforms the other two models. The average RMSE of the generic models is 0.335

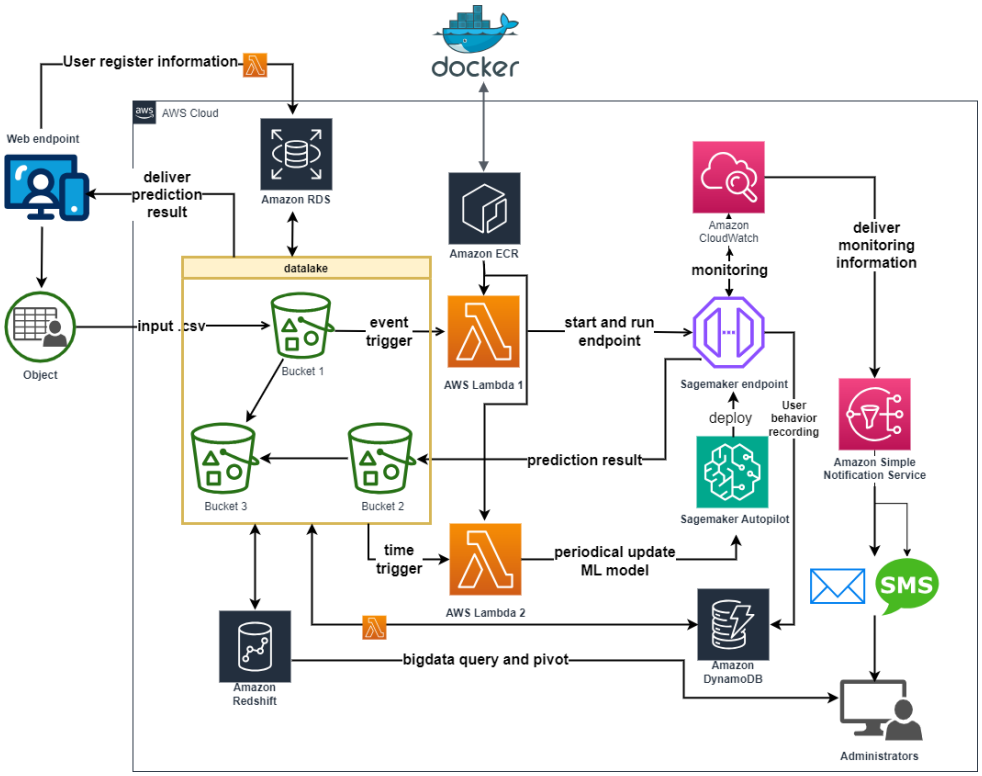
dB and 0.019 dB lower than that of the specific models and hybrid models, respectively, while the average R^2 is 4.62% and 0.17% higher than that of the specific models and hybrid models, respectively. Considering overall performance, model complexity, and versatility, the generic models are the best, followed by the hybrid models, with the specific models ranking last.

Overall, the research provides valuable insights into mitigating the impact of weather on hybrid RFL/FSO communication systems, facilitating the development of more robust and efficient wireless connectivity solutions.

6 Appendix

The code can be accessed in github:
https://github.com/hahawang1986/Hybrid-Optical-Radio-Frequency-Communication-Channel-Model.git

The model in this project has been deployed on the AWS Cloud, providing prediction services. The data pipeline is shown in section 6



References

[1] Antonios Lionis, Konstantinos Peppas, Hector E Nistazakis, Andreas Tsigopoulos, Keith Cohn, and Athanassios Zagouras. Using machine learning algorithms for accurate received optical power prediction of an fso link over a maritime environment. In *Photonics*, volume 8, page 212. MDPI, 2021.

[2] Cao Ying, Miao Qi-Guang, Liu Jia-Chen, and Gao Lin. Advance and prospects of adaboost algorithm. *Acta Automatica Sinica*, 39(6):745–758, 2013.

[3] Yi Feng, Linlan Liu, and Jian Shu. A link quality prediction method for wireless sensor networks based on xgboost. *IEEE Access*, 7:155229–155241, 2019.

[4] Mavuto M Mukaka. A guide to appropriate use of correlation coefficient in medical research. *Malawi medical journal*, 24(3):69–71, 2012.

[5] Syed Agha Hassnain Mohsan, Muhammad Asghar Khan, and Hussain Amjad. Hybrid fso/rf networks: A review of practical constraints, applications and challenges. *Optical Switching and Networking*, 47:100697, 2023.

[6] Mostafa Zaman Chowdhury, Moh Khalid Hasan, Md Shahjalal, Md Tanvir Hossan, and Yeong Min Jang. Optical wireless hybrid networks: Trends, opportunities, challenges, and research directions. *IEEE Communications Surveys & Tutorials*, 22(2):930–966, 2020.

[7] P Series. Propagation data and prediction methods required for the design of earth-space telecommunication systems. *Recommendation ITU-R*, pages 618–12, 2015.

[8] P Series. Propagation data required for the design of terrestrial free-space optical links. *Recommendation ITU-R*, 2012.

[9] Antonios Lionis, Konstantinos Peppas, Hector E Nistazakis, Andreas Tsigopoulos, Keith Cohn, and Athanassios Zagouras. Using machine learning algorithms for accurate received optical power prediction of an fso link over a maritime environment. In *Photonics*, volume 8, page 212. MDPI, 2021.

[10] Kappala Vinod Kiran, Subhanesh Perinbaraj, Jayashree Pradhan, Pradeep Kumar Mallick, Ashok Kumar Turuk, and Santos Kumar Das. Machine learning aided switching scheme for hybrid

fso/rf transmission. *Intelligent Decision Technologies*, 14(4):529–536, 2020.

[11] Wikipedia contributors Harry585. Random forest bagging illustration, 2023. [Online; accessed 1-Apr-2024].

[12] Tin Kam Ho. The random subspace method for constructing decision forests. *IEEE transactions on pattern analysis and machine intelligence*, 20(8):832–844, 1998.

[13] Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001.

[14] Davide Chicco, Matthijs J Warrens, and Giuseppe Jurman. The coefficient of determination r-squared is more informative than smape, mae, mape, mse and rmse in regression analysis evaluation. *PeerJ Computer Science*, 7:e623, 2021.

[15] DenisBoigelot and Imagecreator. Correlation examples. https://commons.wikimedia.org/wiki/File:Correlation_examples2.svg, 2011. Accessed: 2024-06-02.

[16] Leo Breiman and Adele Cutler. Manual–setting up, using, and understanding random forests v4. 0. 2003. *URL https://www. stat. berkeley. edu/˜ breiman/Using_random_forests_v4. 0. pdf*, 2011.

[17] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[18] André Altmann, Laura Toloşi, Oliver Sander, and Thomas Lengauer. Permutation importance: a corrected feature importance measure. *Bioinformatics*, 26(10):1340–1347, 2010.

[19] Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar):1157–1182, 2003.

[20] Girish Chandrashekar and Ferat Sahin. A survey on feature selection methods. *Computers & Electrical Engineering*, 40(1):16–28, 2014.