# Hybrid Optical Radio Frequency Communication Channel Model Progress Report (Part B)

Aditya Venugopalan Nediyirippil
**a1899824**

May 2, 2025

Report submitted for **Data Science Research** at the School of Mathematical Sciences, University of Adelaide


THE UNIVERSITY of ADELAIDE

Project Area: **Data Science**
Project Supervisor: **Siu Wai Ho**

**Abstract**

Hybrid Free Space Optical (FSO) and Radio Frequency (RF) systems offer unique advantages in high-speed wireless communication; however, their performance is highly affected by weather-induced signal attenuation. Accurate modeling of signal loss under various atmospheric conditions—such as fog, dust, rain, and snow—is critical for reliable link adaptation.

This project investigates three machine learning-based strategies using Random Forest models to predict RF and FSO attenuation:

Method 1 (Weather-Specific Modeling): RF and FSO attenuation are predicted separately for each SYNOP weather class using stratified models.

Method 2 (Sequential Hybrid Modeling): FSO attenuation is predicted using RF signal and weather features.

Method 3 (Reverse Hybrid Modeling): RF attenuation is predicted using FSO signal and weather features.

Each method was evaluated using Root Mean Squared Error (RMSE), Coefficient of Determination ($R^2$), Out-of-Bag (OOB) score, Mutual Information (MI), and Pearson correlation to assess both accuracy and structural preservation.

Method 1 achieved the highest Pearson correlation (r 0.9883 for RF, 0.9900 for FSO) and strong MI (2.45 for RF, 2.83 for FSO), confirming that per-weather stratification preserves signal structure well.

Method 2, despite achieving the lowest RMSE (0.8021) for FSO prediction, showed poor correlation (Pearson r = -0.0139), indicating signal inversion or collapse. MI was moderate (2.5445), suggesting partial structural retention.

Method 3 offered the best trade-off: RMSE = 0.8022, $R^2$ = 0.9357, Pearson r = 0.9675, and MI = 2.7427—outperforming the general model while preserving correlation well.

Compared to the Generic RF/FSO models (e.g., General RF RMSE = 0.8223, $R^2$ = 0.9325, OOB = 0.9266), Method 3 demonstrated enhanced predictive accuracy and better retention of information structure, particularly under challenging weather conditions.

These findings validate the effectiveness of hybrid-sequential learning, especially Method 3, for modeling signal attenuation in real-world optical–RF systems and highlight the importance of structural evaluation (MI, Pearson) in addition to accuracy metrics.

# 1 Introduction

In recent times it has been observed that Hybrid communication systems that combine Radio Frequency (RF) and Free Space Optics (FSO) technologies have become very important to meet the ever-growing demand for high-capacity and reliable wireless communication links [1]. The RF links offer reliable connections under diverse weather conditions but are restricted by their bandwidth limitations . On the other hand the FSO links provide comparatively high data rates but suffer significantly from atmospheric attenuation during adverse weather events such as fog, rain, and snow[2].As communication networks demand for higher reliability increases, hybrid RF–FSO systems are being evaluated as a solution to overcome the limitations of individual technologies[3].An important challenge in designing of hybrid communication design is to accurately predict the pattern of signal loss under various environmental scenarios. Modeling the complex interactions between atmospheric phenomena and signal propagation is important to enable the adaptive link management and system optimization. Machine learning approaches, mainly tree-based methods such as Random Forests display understandable success in capturing nonlinear relationships between weather conditions and communication performance metrics.

This project aims for developing and evaluating machine learning models to predict RF and FSO signal attenuation. Three modeling strategies were used. The first approach (Method 1) involves training specific Random Forest regression models for each weather condition, giving highly specific predictions. The second approach (Method 2) used RF attenuation and weather features as inputs to predict FSO attenuation, offering a hybrid prediction strategy. The third approach (Method 3) uses opposite approach where using predicted FSO attenuation we predict and try to estimate RF attenuation. Each method aims to achieve low prediction error and tries to observe important statistical values such as Pearson correlation and mutual information between the measured and predicted attenuation values.

Feature selection was performed using backward elimination accompanied by out-of-bag (OOB) error, ensuring that models were trained on informative yet compact feature sets. Model performance was analyzed across various metrics, including Root Mean Squared Error (RMSE), $R^2$ score, Pearson correlation coefficient, and normalized mutual information. After that entropy-based analyses were conducted to evaluate the information preservation across prediction pipelines under various weather conditions.

The idea for exploring hybrid and sequential prediction strategies

lies in the possibility to improve communication links under difficult atmospheric conditions. In particular, leveraging cross-modal predictions between RF and FSO could enable more adaptive and reliable network systems. While specific per-weather models may achieve high accuracy however hybrid strategies may offer better adaptability and flexibility in real-world scenarios.

This report presents the methodology, results, and important evaluation of three modeling approaches. It aims for a better understanding of how machine learning can help in dynamic hybrid communication system designing and proposes potential future solutions based on the results observed .

# 2 Background

## 2.1 Hybrid RF/FSO Communication Systems

Radio Frequency (RF) and Free Space Optics (FSO) are two important technologies in wireless communications, each having its own advantages and disadvantages. RF systems depend on electromagnetic waves like in the microwave and millimeter-wave bands and are frequently for their robustness in non-line-of-sight scenarios and consistent performance under different atmospheric environments (Fayed & Aly, 2019). But RF links are limited by their spectrum , interference, and other constraints, particularly since their demand keeps on increasing for higher-bandwidth.

On the other hand FSO uses LED's for high-speed, line-of-sight transmission through free space. FSO links offer fiber-like bandwidth, low latency, and are unaffected by electromagnetic interference, making them optimal choice for point-to-point communication in dense areas or between satellites and ground stations. However, they suffer on a huge scale under atmospheric disturbances such as fog, haze, and heavy precipitation, where the light beam is scattered or absorbed.

In order to deal with the limitations of each individual system, A proposal for building hybrid RF–FSO architectures have been established. These systems effectively switch or extend both channels depending on weather conditions, ensuring high availability and reliability. The practical use of such systems, especially in urban or disaster prone scenarios, is increasingly supported by revolution in weather-aware signal management.

## 2.2 Atmospheric Effects on Signal Attenuation

Environmental factors play a vital role in spoiling the performance of RF and FSO links. FSO signals are vulnerable to fog, which causes Mie scattering due to water droplets with diameters similar to optical wavelengths, resulting in intense attenuation (JOCM, 2013). Rain and snow, although less severe than fog in terms of FSO impact, can also contribute to signal degradation depending on particle density and rate.

RF signals, particularly those in the Ka-band and above, are affected by attenuation mainly from rain droplets that absorb and scatter microwave signals, causing link loss. The effect is more found at higher frequencies because of greater water absorption (Wikipedia, "Rain fade"). Atmospheric gases and clouds can cause signal propagation.

Although various empirical and semi-analytical models have been developed to analyze these effects such as the ITU-R models for RF and the Beer-Lambert law for FSO—their accuracy is limited in complex real

world scenarios where data on droplet size, density, or temperature may not be precisely known (NASA, 2019). Additionally these models often fail to generalize wide range of changing weather types and geographic regions.

This ignites the passion for the use of data-driven methods that can directly learn from ancient data and observed signal loss patterns, which are compatible and flexibly to noise, missing data, or previously unknown conditions.

## 2.3 Machine Learning for Attenuation Prediction

Machine learning (ML) is an effective alternative to traditional physical models . By learning difficult, nonlinear relationships between environmental features and signal degradation patterns. While talking about hybrid communication systems, ML models can be trained to predict RF or FSO attenuation using weather conditions like temperature, humidity, visibility, rainfall, and atmospheric pressure.

Among several ML methods, Random Forests (RFs) are well-suited for this purpose because of their power to maintain high-dimensional feature spaces, resistance to overfitting, and built-in estimation of feature importance. Random Forests also provide Out-of-Bag (OOB) error estimates that allow model evaluation without requiring cross-validation, which is crucial when working with large datasets.

While previous literature has focused on single-channel modeling, this study expands its boundaries by investigating whether predictions across RF and FSO can be hybridized to maintain both performance and signal structure with statistical dependencies like Pearson correlation and Mutual Information.

## 2.4 Research Gap and Motivation for This Study

In spite of the growth for ML models for channel attenuation, several gaps are present in the literature. First, most prior studies have treated RF and FSO channels independently, ignoring their dependence during co-deployment. Second, common and basic evaluation metrics like RMSE or $R^2$ alone may not detect the desired results especially when sequential modeling is deployed. Correlation metrics, like Pearson r and Mutual Information, are often determined , despite their importance in signal recovery and link adaptation.

This motivates a deeper building of hybrid modeling pipelines where: Weather-specific models are compared to general models (Method 1), RF predictions are used to improve FSO prediction (Method 2), and

FSO predictions are used in reverse to inform RF attenuation estimation (Method 3).

These methods are analyzed not only in terms of prediction accuracy but also in how well they maintain statistical relationships between RF and FSO under several weather types. This method evaluation is very valuable for real-world systems that need accurate cross-modal redundancy, dynamic constant switching.

## 2.5 Previous Results from Part A

In Part A of this project, the main focus was carry a thorough exploratory data analysis (EDA) and building baseline general Random Forest models to predict RF and FSO signal attenuation. The dataset included 27 atmospheric features, such as absolute humidity, visibility, wind speed, and particulate matter.

An extensive EDA was performed, including:

Distribution plots (KDE) and boxplots to identify skewness and outliers across the dataset

Correlation heatmaps to analyze inter-feature relationships,

Identification of high skewness in features such as RainIntensity and Particulate, which were dealt with using logarithmic and square-root transformations.

Data cleaning involved:

removing outliers using 0.5th and 99.5th percentile clipping,

Skewness reduction on asymmetric features to improve model generalization,

Resampling to balance the SYNOPCode distribution by downsampling and bootstrapping.

For predictive modeling, Random Forest regressors were trained for both FSO and RFL attenuation using entire dataset. Key results included:

General RF Model: RMSE = 0.4901, $R^2$ = 0.9797, OOB Score = 0.978

General FSO Model: RMSE = 0.7845, $R^2$ = 0.9590, OOB Score = 0.956

These results confirmed great baseline predictive performance, especially for RF attenuation.

SYNOP-Specific Modeling (Method 1 Preparation) To understand weather-specific signal behavior, individual Random Forest models were trained for each different SYNOP code These per synop weather models often showed lower RMSE and higher $R^2$ scores compared to the general model, especially for fog-heavy and rain-heavy SYNOP classes.

Backward Feature Elimination (BFE) In order to improve the models backward feature elimination based on Out-of-Bag (OOB) error was done for both RF and FSO models. This elimination process prioritized features that contributed most to prediction accuracy, ultimately reducing features complexity and better generalization.

### 2.5.1 Limitations of Part A

Part A provided a important base for understanding the hybrid RF–FSO attenuation dataset, but had several limitations that restricted the model's generalization and performance

- **No Feature Selection:** Models were trained using all features without analyzing their individual importance. This increased the risk of overfitting and model
- **Skewed Data Distribution:** The training data had highly imbalanced SYNOPCode distributions. clear weather dominated the results, which led to bias in general models
- **Lack of Weather-Aware Modeling:** The initial models did not use weather stratification. Weather conditions hugely affect signal behavior, but this factor was not modeled in Part A.
- **No Cascaded or Sequential Models:** Part A used predictions without exploring cascaded approaches
- **No Mutual Information or Entropy Analysis:** Part A was evaluated only on traditional basic evaluation metrics (e.g., RMSE, $R^2$), ignoring on informative measures like correlation and mutual information that are crucial for signal analysis.
- **Unoptimized Hyperparameter:** The random forest models used default parameters, which may not have best for the given data, potentially under utilizing and undetermined the model capacity.

These limitations motivated the extended work in Part B, where we introduced feature selection, weather based stratified modeling, sequential learning, mutual information analysis, and performance comparisons across methods.

# 3 Methods

## 3.1 Libraries and Tools Used

This project was developed in **Python 3.11** using a combination of scientific computing and machine learning libraries:

- **pandas, numpy:** For data preprocessing, and transformation of datasets.
- **scikit-learn (sklearn):** The main learning library used for implementing Random Forest regressors, train-test splitting, model evaluation (RMSE, $R^2$), and feature selection.
- **matplotlib, seaborn:** Used for visualization of feature importance, correlation heatmaps, RMSE trends, and method comparison plots.
- **scipy:** Specifically for statistical computations such as Pearson correlation.

All experiments were run locally using Jupyter Notebook on an Anaconda distribution environment.

## 3.2 Random Forest Regression Algorithm

All predictive models in this project were developed using the Random Forest regression algorithm, Random Forests work by aggregating the outputs of multiple decision trees, each trained on a random bootstrap sample of the dataset. At each split, a random subset of features is considered, which increases variance reduction and reduces overfitting.

The algorithm offers several benefits in the context of hybrid RF–FSO modeling:

- Handles high-dimensional data , which in our case was necessary
- Comes with built-in feature importance metrics, which are used during backward elimination.
- Supports Out-of-Bag (OOB) scoring, which enables internal cross-validation.
- Is robust to noise and nonlinear relationships, which are common in issues in weather-influenced data.

Random Forest regressors were used for all modeling stages, including baseline and optimized baseline, per-SYNOP, and sequential (hybrid) models. Unless otherwise specified, all forests used 100 estimators, with ‘oob_score=True‘ and a fixed random state to ensure reproducibility.

## 3.3 Data Preparation And Dataset Analysis

This section explains the dataset, exploratory steps, and the data preprocessing pipeline adopted for model building and robustness. The dataset comprised of 91,379 rows and 27 columns, including two target variables: FSO_Att (Free Space Optics Attenuation) and RFL_Att (Radio Frequency Attenuation), accompanied with 25 environmental features such as temperature, humidity, particulate concentration, visibility, frequency, and wind metrics.

### 3.3.1 Dataset Overview and Initial Inspection

Analysis began by importing the dataset (RFLFSODataFull.csv) and undergoing data type check from figure 1 . A null value check as seen in figure 2 eliminating the need for imputation since no missing values (figure 3). A type check was also done to make sure all columns were numerical as for further statistical analysis and modeling (figure 2)

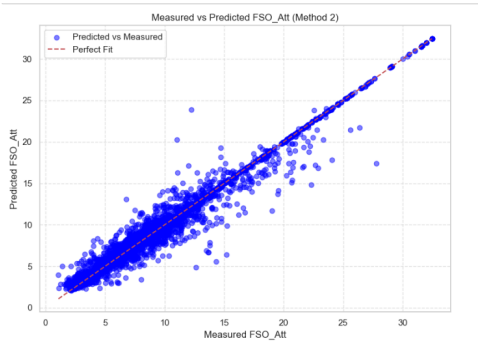
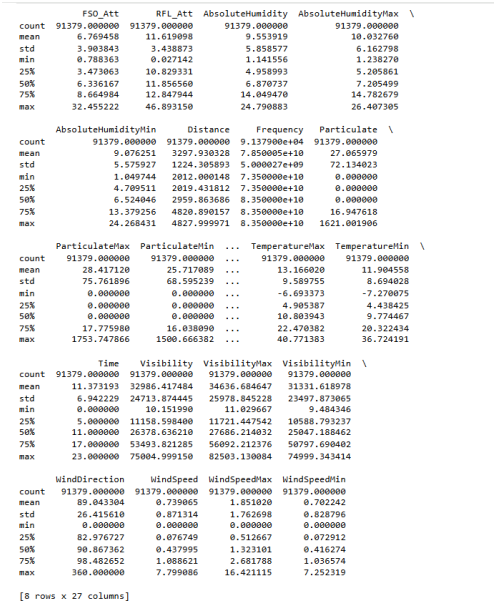



Figure 1: Dataset preview showing all 27 features and data types



| | FSO_Att | RFL_Att | AbsoluteHumidity | AbsoluteHumidityMax |
|---|---|---|---|---|
| count | 91379.000000 | 91379.000000 | 91379.000000 | 91379.000000 |
| mean | 6.769458 | 11.619098 | 9.553919 | 10.032760 |
| std | 3.903843 | 3.438873 | 5.858577 | 6.162798 |
| min | 0.788363 | 0.027142 | 1.141556 | 1.238270 |
| 25% | 3.473063 | 10.829331 | 4.958993 | 5.205861 |
| 50% | 6.336167 | 11.856560 | 6.870737 | 7.205499 |
| 75% | 8.664984 | 12.847944 | 14.049470 | 14.782679 |
| max | 32.455222 | 46.893150 | 24.790883 | 26.407305 |

| | AbsoluteHumidityMin | Distance | Frequency | Particulate |
|---|---|---|---|---|
| count | 91379.000000 | 91379.000000 | 9.137900e+04 | 91379.000000 |
| mean | 9.076251 | 3297.930328 | 7.850005e+10 | 27.065979 |
| std | 5.575927 | 1224.305893 | 5.000027e+09 | 72.134023 |
| min | 1.049744 | 2012.000148 | 7.350000e+10 | 0.000000 |
| 25% | 4.709511 | 2019.431812 | 7.350000e+10 | 0.000000 |
| 50% | 6.524046 | 2959.863686 | 8.350000e+10 | 0.000000 |
| 75% | 13.379256 | 4820.890157 | 8.350000e+10 | 16.947618 |
| max | 24.268431 | 4827.999971 | 8.350000e+10 | 1621.001906 |

| | ParticulateMax | ParticulateMin | ... | TemperatureMax | TemperatureMin |
|---|---|---|---|---|---|
| count | 91379.000000 | 91379.000000 | ... | 91379.000000 | 91379.000000 |
| mean | 28.417120 | 25.717089 | ... | 13.166020 | 11.904558 |
| std | 75.761896 | 68.595239 | ... | 9.589755 | 8.694028 |
| min | 0.000000 | 0.000000 | ... | -6.693373 | -7.270075 |
| 25% | 0.000000 | 0.000000 | ... | 4.905387 | 4.438425 |
| 50% | 0.000000 | 0.000000 | ... | 10.803943 | 9.774467 |
| 75% | 17.775980 | 16.038090 | ... | 22.470382 | 20.322434 |
| max | 1753.747866 | 1500.666382 | ... | 40.771383 | 36.724191 |

| | Time | Visibility | VisibilityMax | VisibilityMin |
|---|---|---|---|---|
| count | 91379.000000 | 91379.000000 | 91379.000000 | 91379.000000 |
| mean | 11.373193 | 32986.417484 | 34636.684647 | 31331.618978 |
| std | 6.942229 | 24713.874445 | 25978.845228 | 23497.873065 |
| min | 0.000000 | 10.151990 | 11.029667 | 9.484346 |
| 25% | 5.000000 | 11158.598400 | 11721.447542 | 10588.793237 |
| 50% | 11.000000 | 26378.636210 | 27686.214032 | 25047.188462 |
| 75% | 17.000000 | 53493.821285 | 56092.212376 | 50797.690402 |
| max | 23.000000 | 75004.999150 | 82503.130084 | 74999.343414 |

| | WindDirection | WindSpeed | WindSpeedMax | WindSpeedMin |
|---|---|---|---|---|
| count | 91379.000000 | 91379.000000 | 91379.000000 | 91379.000000 |
| mean | 89.043304 | 0.739065 | 1.851020 | 0.702242 |
| std | 26.415610 | 0.871314 | 1.762698 | 0.828796 |
| min | 0.000000 | 0.000000 | 0.000000 | 0.000000 |
| 25% | 82.976727 | 0.076749 | 0.512667 | 0.072912 |
| 50% | 90.867362 | 0.437995 | 1.323101 | 0.416274 |
| 75% | 98.482652 | 1.088621 | 2.681788 | 1.036574 |
| max | 360.000000 | 7.799086 | 16.421115 | 7.252319 |

[8 rows x 27 columns]

Figure 2: Statistical Analysis Of The Data Set

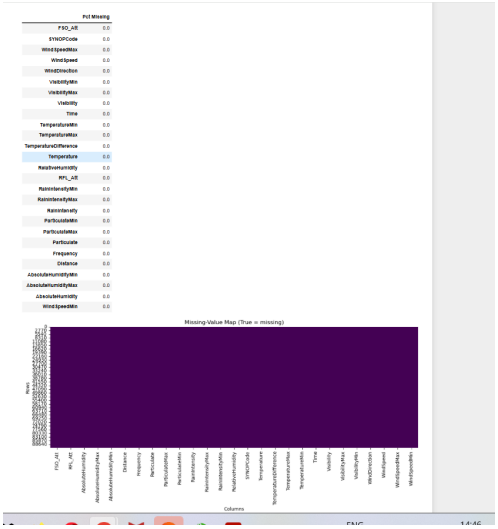



Figure 3: Missing percentage and missing heatmap

### 3.3.2 Exploratory Data Analysis (EDA)

Several analyses were performed to understand the data distribution and relationships

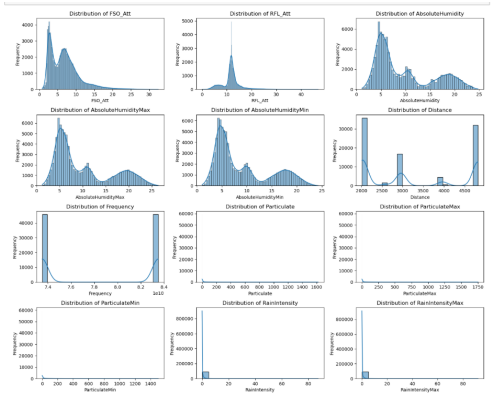



Figure 4: KDE plots to inspect skewness

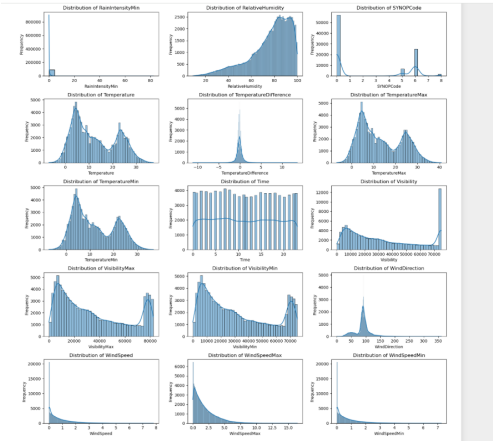



Figure 5: KDE plots to inspect skewness

KDE and Histogram Plots were used to visualize the distribution and skewness in the dataset( figure 4 and figure 5)

From figure 6 Correlation Heatmap matrix was plotted to analyze multicollinearity and relationships(Hastie, Tibshirani and Friedman, 2009). FSO_Att and RFL_Att showed high correlation with multiple features like AbsoluteHumidity, Temperature, and Visibility.

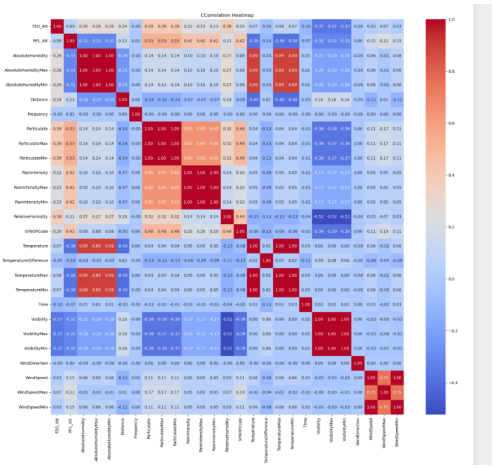



Figure 6: Heatmap to understand correlation between features

Weather Distribution Plot were plotted to match SYNOPCode to human-readable weather categories (e.g., clear, rain, snow) and visualized their frequency using a count-plot to inspect dataset balance across weather types. (Figure 7)

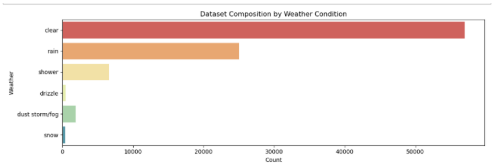



Figure 7: Weather Condition Distribution (Countplot)

### 3.3.3 Outlier Treatment

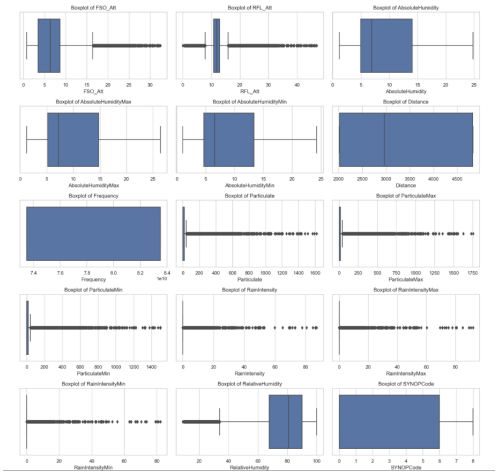


Figure 8: Before removal of outliers

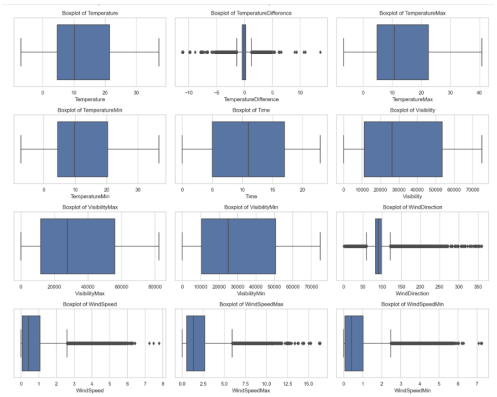


Figure 9: Before removal of outliers

Outliers were detected using boxplots(figure 8 and figure 9). To reduce their influence without deleting data, we capped extreme values at the 0.5th and 99.5th percentiles as we can see from figure (10,11,12,13,14,15,16,17,18)(Aggarwal, 2017)

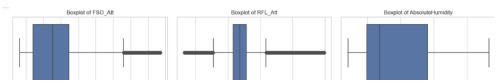


Figure 10: After removal of outliers

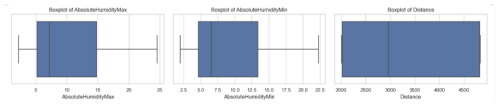



Figure 11: After removal of outliers

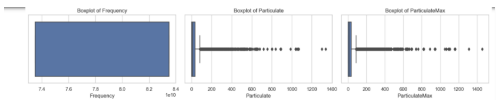



Figure 12: After removal of outliers

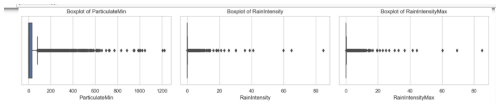



Figure 13: After removal of outliers

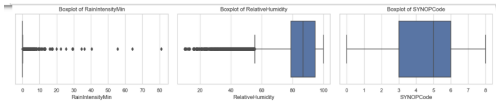



Figure 14: After removal of outliers

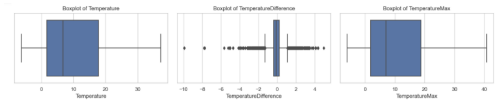



Figure 15: After removal of outliers

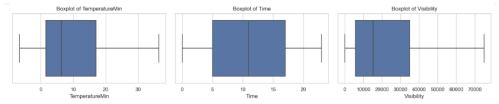



Figure 16: After removal of outliers


Figure 17: After removal of outliers

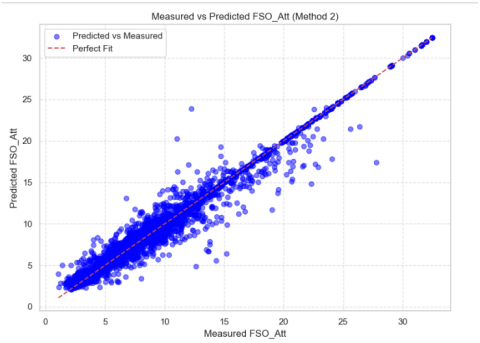



Figure 18: After removal of outliers

### 3.3.4 Skewness Correction

```
FSO_Att                 1.180187
RFL_Att                 0.333365
AbsoluteHumidity        0.836056
AbsoluteHumidityMax     0.841151
AbsoluteHumidityMin     0.841513
Distance                0.244208
Frequency              -0.000022
Particulate             3.614893
ParticulateMax          3.610046
ParticulateMin          3.615347
RainIntensity           5.972863
RainIntensityMax        5.959528
RainIntensityMin        5.947654
RelativeHumidity       -1.045672
SYNOPCode               0.579932
Temperature             0.395913
TemperatureDifference   0.270284
TemperatureMax          0.402804
TemperatureMin          0.398743
Time                    0.030113
Visibility              0.524502
VisibilityMax           0.530113
VisibilityMin           0.530280
WindDirection           0.333365
WindSpeed               1.622634
WindSpeedMax            1.380512
WindSpeedMin            1.627424
Weather                     NaN
dtype: float64
```

Figure 19: Before skewness transformation

Various features displayed high skewness (e.g., RainIntensity, WindSpeedMin, ParticulateMin), which can affect model predictions as seen in figure 19 These were reduced by log1p transformations for positively skewed variables (e.g., Particulate, WindSpeedMax). Sqrt for extreme outliers in rainfall features.Power transform for negatively skewed features such

as RelativeHumidity (figure 20) (ResearchGate, n.d.).These transformations reduce skewness, and normalize distributions (Weisberg, 2005)

| | |
|---|---|
| FSO_Att | 1.180187 |
| RFL_Att | 0.333365 |
| AbsoluteHumidity | 0.836056 |
| AbsoluteHumidityMax | 0.841151 |
| AbsoluteHumidityMin | 0.841513 |
| Distance | 0.244208 |
| Frequency | -0.000022 |
| Particulate | 0.995452 |
| ParticulateMax | 0.988955 |
| ParticulateMin | 1.002831 |
| RainIntensity | 2.302150 |
| RainIntensityMax | 2.288917 |
| RainIntensityMin | 2.315559 |
| RelativeHumidity | -0.513451 |
| SYNOPCode | 0.579932 |
| Temperature | 0.395913 |
| TemperatureDifference | 0.270284 |
| TemperatureMax | 0.402804 |
| TemperatureMin | 0.398743 |
| Time | 0.030113 |
| Visibility | 0.524502 |
| VisibilityMax | 0.530113 |
| VisibilityMin | 0.530280 |
| WindDirection | 0.333365 |
| WindSpeed | 1.622634 |
| WindSpeedMax | 0.298834 |
| WindSpeedMin | 0.795607 |
| Weather | NaN |

Figure 20: After skewness transformation

### 3.3.5 Summary

After EDA and preprocessing , The dataset was clean, complete, and statistically stable. No imputation was necessary. All transformations were applied before feature selection and model training.

## 3.4 Feature Selection and General Model Training

The following section presents the development of model full-feature baselines to optimized models based on feature elimination. The aim was to build Random Forest regressors for both RF and FSO attenuation prediction using accurate and easy feature subsets.

### 3.4.1 Initial General Models (All Features)

The first trained two general Random Forest models in this project were developed using the complete data set excluding any selection techniques.

This step was done to have baseline models

Although these models delivered strong predictive power, the high dimensionality posed risks of redundancy, overfitting, and possible bias of classes with more data.

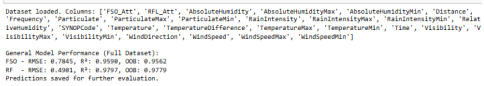



Figure 21: General Model Performance (Full Dataset)

### 3.4.2 Backward Feature Elimination using OOB Score

In order to avoid overfitting and bias issues we implemented a backward feature elimination (BFE) routine tagged by the Out-of-Bag (OOB) score. For Method 1, feature elimination was performed independently on each attenuation target (RFL and FSO) using the balanced dataset. At each iteration, the least contributing feature was removed, and models were retrained until the performance worsen (early stopping). The optimal feature sets achieved lower RMSE and higher $R^2$ and OOB scores compared to using all features.

For Method 2 a hybrid method was used to predict FSO attenuation. We used the same BFE function, but the input now included measured RFL attenuation along with weather features. After eliminating less important features, the final subset obtained improved generalization performance, confirmed by reduced test RMSE and increased OOB score.

For Method 3 we first trained a Random Forest model to predict FSO Attenuation from weather features, then added the predicted FSO Attenuation as a new input to a second model predicting RFL Attenuation. We again applied BFE . BFE was executed after the FSO prediction stage to ensure independence of input . Which resulted in compact feature set that preserved the predictive relationship between RF and FSO .

Visualizations of feature elimination curves and CSVs of selected features were saved for all methods. These selected features were then used to retrain the final models.

# 4 Model Building And Evaluation

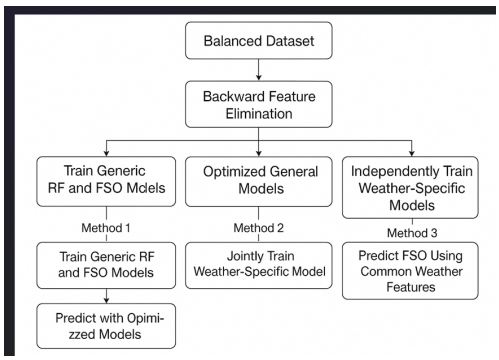



Figure 22: Tree Diagram Explaining the workflow

### 4.0.1 Optimized General Models (Post-BFE)

Following feature selection using backward elimination, we retrained the general Random Forest models for both RFL and FSO using only the retained optimal features. The backward elimination process was motivated by Out-of-Bag (OOB) score and Root Mean Squared Error (RMSE) and R squared, with early stopping used when further removal of features resulted in performance degradation.

These retrained models are the optimized baseline (general) predictors for RF and FSO attenuation. The training and evaluation were conducted on the same balanced dataset used in early baseline models. By reducing the features, these optimized models aimed to improve generalization and reduce overfitting.The selected feature lists for both targets were saved and reused for final evaluation steps and visual comparisons.

### 4.0.2 Hyperparameter Tuning for Optimized RF and FSO modles

Despite hyperparameter tuning using GridSearchCV, the optimized models didn't improve in terms of RMSE or $R^2$.

In fact, RMSE increased slightly, suggesting that the feature selection was optimal

$R^2$ and OOB scores also slightly degraded for both RF and FSO after tuning, confirming no benefit from hyperparameter changes.

This result highlights that feature selection by backward elimination had better effect than tuning.

#### 4.0.3 Evaluation of General vs Optimized Models

To avail the benefits of feature selection, a direct comparison was conducted between the general models trained using all features and the optimized models trained using only features selected via backward elimination.

Separate Random Forest regressors were trained for both RFL_Att and FSO_Att targets using the same stratified train-test split to ensure fair comparison. Performance was evaluated using Root Mean Squared Error (RMSE), Coefficient of Determination ($R^2$), and Out-of-Bag (OOB) score.

This comparison allowed us to focus the effect of feature selection by observing whether the optimized models generalize better to unseen data or not . The results of this experiment are discussed in the Results section.

#### 4.0.4 Method 1: Per-SYNOP Code Specific Models

In this method, separate Random Forest models were trained for each SYNOP weather condition to analyze attenuation behavior under specific environmental condition. This approach gives freedom to models to explore individual patterns that may be seen in a general model.

For each unique SYNOP code, the a unique subset of the training data was extracted. Backward feature elimination was applied independently on these subsets using the same logic with OOB, resulting an optimal set of features dedicated to that weather condition. Two sets of models were trained per SYNOP:

- One for predicting RF attenuation (RFL_Att)
- One for predicting FSO attenuation (FSO_Att)

These specialized models were then analyzed on test samples belonging to the same SYNOP class. The selected features and performance metrics (RMSE, $R^2$, and OOB score) for each model were saved and later compared to those of the generic and hybrid models(later)

This method allows analysis of weather-specific attenuation patterns which will later be used as a benchmark to evaluate the effectiveness of hybrid modeling approaches.

#### 4.0.5 Method 2: Hybrid FSO Prediction (FSO ← RF + Weather)

Method 2 aimed to improve the prediction of FSO attenuation by adding RF signal attenuation (RFL_Att) as a feature along with other features .

This approach displays a sequential modeling strategy, where the correlation between RF and FSO is used together to improve generalization, especially under challenging weather conditions.

The model consisted of all available weather features plus the true measured RFL predicted value. A Random Forest regressor was trained on this hybrid data to predict FSO Attenuation. To prevent overfitting backward feature elimination was re-applied on this combined input using the same OOB strategy. The model performance (RMSE, $R^2$, OOB) was tracked to determine the optimal subset.

Once the optimal feature set was obtained, the final hybrid model was retrained using only the retained features. Its predictions were saved including mutual information and correlation analysis by SYNOP weather code.

#### 4.0.6 Method 3: Sequential RF Prediction (RF ← FSO + Weather)

Method 3 explored a sequential modeling setup in which RF attenuation was predicted using weather conditions with predicted FSO attenuation as a feature. This approach simulates a real-world scenario where FSO readings (e.g., from satellite) are available, and RF signal behavior must be inferred under different environmental conditions.

The method was implemented in two stages. First, a Random Forest model was trained to predict FSO Attenuation using weather features alone. The predicted FSO values were then added to the original weather features to form a new data set. This hybrid input was used to train a second Random Forest model that predicts RFL Attenuation.

Just like previous methods, backward feature elimination was applied to the hybrid input (FSO prediction + weather), using OOB score and RMSE for selection . This ensured that only the most important features, including the predicted FSO, were retained. The final RFL model was trained on the reduced feature set and evaluated using a test set.

### 4.1 Summary of Modeling Workflow

The process of model building begin with general Random Forest models for RF and FSO attenuation using a balanced dataset and applied backward feature elimination to reduce dimensionality. Followed by the development of per-SYNOP models (Method 1) to analyze weather-specific behavior using stratified subsets feature selection.(figure 23)

From figure 23 , Next two hybrid modeling strategies were implemented. Method 2 added measured RF signals and weather features to predict FSO attenuation, while Method 3 used predicted FSO and

weather conditions to predicted RF attenuation. Both approaches included backward feature elimination to select hybrid feature sets.

All models were trained and tested using stratified splits to adopt fairness in comparison. Performance was evaluated using RMSE, $R^2$, OOB score, Pearson correlation, and mutual information. The resulting predictions and selected feature sets were to evaluate and assess both accuracy and the preservation of correlation between channels under different environmental conditions as we can see in figure 23

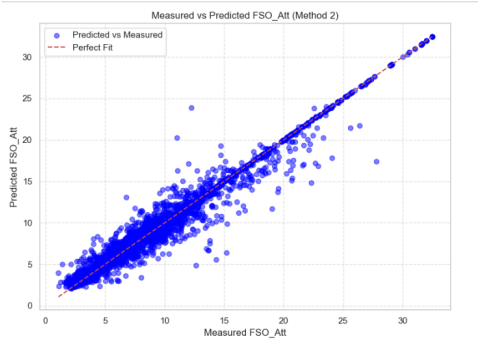



Figure 23: Table showing the summary of the models built

## 4.2 Evaluation Metrics

To evaluate model performances, the use of traditional regression metrics and information based metric were done. These metrics were computed globally and per SYNOP weather condition to evaluate both prediction accuracy and preservation of signal information.

### 4.2.1 1. Root Mean Squared Error (RMSE)

RMSE measures the mean magnitude of prediction errors and penalizes larger deviations more strongly:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Where:

- $y_i$: Ground truth (measured) value
- $\hat{y}_i$: Predicted value
- $n$: Total number of samples

In our project, RMSE was computed for all model (generic, optimized, per-SYNOP, Method 2, and Method 3) using `mean_squared_error(..., squared=False)` from `sklearn.metrics`. It was also calculated separately per SYNOPCode to identify weather related weaknesses.

#### 4.2.2 2. Coefficient of Determination ($R^2$)

The $R^2$ score signifies the proportion of variance in the true values that is captured by the model predictions:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Where $\bar{y}$ is the mean of observed values.

In our code, this was computed using `r2_score` from `sklearn.metrics` for each model and plotted in both summary tables and per-weather evaluations.

#### 4.2.3 3. Out-of-Bag (OOB) Score

The OOB score is an internal validation metric present in Random Forests. During training, each tree uses a bootstrap sample of the data, excluding a portion as "out-of-bag." The OOB score evaluates predictions on test samples, pretending as a form of internal cross-validation.

In our models, the OOB score was enabled using `oob_score=True` and used using `model.oob_score_`. It was tracked during feature elimination and model optimization for models.

#### 4.2.4 4. Pearson Correlation Coefficient

Pearson correlation coefficient measures the linear relationship between predicted and true attenuation values:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Where:

- $x_i$: Predicted values
- $y_i$: True values
- $\bar{x}, \bar{y}$: Mean values

Pearson correlation was computed using `scipy.stats.pearsonr` in the code, particularly to assess correlation per SYNOPCode. This evaluated how each model preserved the statistical relationship between RF and FSO attenuation under various weather conditions.

#### 4.2.5 5. Mutual Information (MI)

Mutual Information shows the amount of information shared between predicted and true values. Unlike correlation, it captures both linear and nonlinear dependencies:

$$I(X;Y) = H(X) + H(Y) - H(X,Y)$$

Where:

- $H(X)$: Entropy of predicted values
- $H(Y)$: Entropy of true values
- $H(X,Y)$: Joint entropy

**Binning Strategy.** Since MI operates on discrete distributions, we approximate continuous values using fixed-width binning:

$$\text{Bins} = [\min(x), \min(x) + \delta, \min(x) + 2\delta, \ldots, \max(x)]$$

In our implementation, binning was done using `np.histogram` and `np.histogram2d` with manually selected bin widths. We tested $\delta = 0.1$, 0.23, and 1.0 dB to evaluate the effect of bin granularity.

The final normalized mutual information was computed as:

$$\text{Normalized MI} = \frac{I(X;Y)}{H(X,Y)}$$

This metric was computed per SYNOPCode for measured, Method 1, Method 2, and Method 3 predictions, enabling accurate comparison of correlation preservation across models.

All entropy and MI processes were implemented in NumPy, consistent with the project goal to measure information preservation under hybrid channel modeling.

**Notes:** This template suggests that your next section should describe methods used or developed in this report. Methods that are simple background material should go in the previous section. This section focuses on those that are novel, or in some cases just more difficult and more important for the work.

Sometimes the section will be called "methods," but I find a more specific, and descriptive heading is usually preferable.

Your focus in describing these should be reproducibility. A reader should ideally be able to recreate your work from your description.

Describe data, experiments, simulations, or solution techniques such that your reader can understand exactly what you did. It may be helpful

to keep trying to answer the 6Ws: Why, When, Where, What, Who and How.

However, the art of such writing is to balance detail and precision with brevity. Concise descriptions are to be preferred because the information is more accessible. Often we use references to allow us to abbreviate or omit some details that are common to other experiments or problems.

Mathematical notation is also very useful in composing precise, yet concise descriptions of a problem. However, do not use mathematics or jargon for its own sake. Clarity is important, and mathematics or complicated technical terms can either enhance this (when used appropriately) or detract from it (if used carelessly). Your goal is *not* to try to seem smart by using complicated words. Your goal is to communicate!

I have not sought to include a tutorial or examples of LaTeX use here as there are now many sources of such information. For instance see `http://www.maths.adelaide.edu.au/anthony.roberts/LaTeX/index.html`.

# 5 Results

## 5.1 General vs Optimized Models

**Motivation and Justification.** Baseline models were initially trained using all 27 environmental features to predict RF and FSO attenuation. This method provides a baseline, using the entire data set risks overfitting, mainly when conditions of high feature redundancy and imbalanced weather distributions.

To deal with this, we implemented **backward feature elimination (BFE)** guided by out-of-bag (OOB) score and RMSE. This allowed us to iteratively remove the least important feature until performance degradation was detected. The mission was to obtain better generalization, to reduce model complexity.

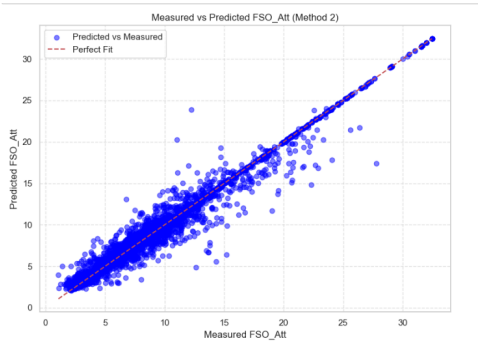



Figure 24: Hyparameter Tuning results

As seen in figure 24 hyperparameter tuning basic was done (e.g., n_estimators, max_depth), but early experiments revealed no significant improvement over default settings. As a result, our focus shifted toward feature selection as the primary optimization method.

**Results Summary.** The impact of feature selection is summarized in the comparison table below, showing the performance of general vs optimized models for both RFL and FSO prediction(figure 25)

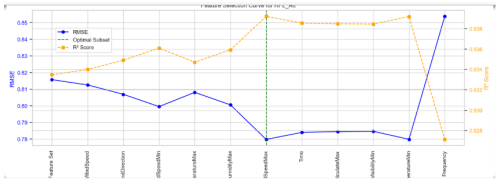


Figure 26: Feature Elimination For RFL With BFE Curve

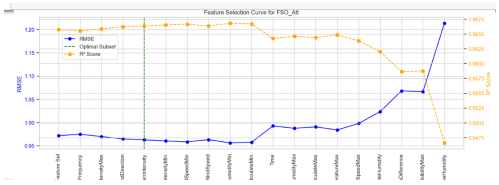


Figure 27: Feature Elimination For FSO With BFE Curve

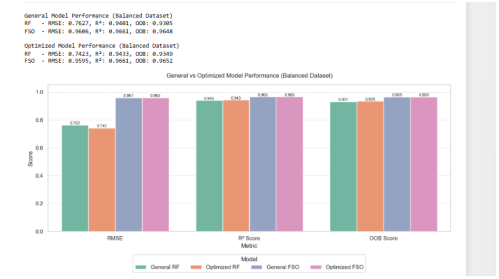


Figure 25: General vs Optimized Model Performance (Balanced Dataset

Additionally, a BFE curve was created to track both model performance at each elimination step. The figure 26 and figure 27 below shows RMSE and $R^2$ evaluation as features were removed. The optimal feature subset was selected just before OOB score or RMSE degraded.(figure 28, figure 29 and figure 30)

| | Step | Removed Feature | Remaining Features | Feature List | RMSE | R² Score | OOB Score |
|---|---|---|---|---|---|---|---|
| 0 | 0 | Full Feature Set | 24 | AbsoluteHumidity, AbsoluteHumidityMax, Absolut... | 0.815725 | 0.933481 | 0.926260 |
| 1 | 1 | WindSpeed | 23 | AbsoluteHumidity, AbsoluteHumidityMax, Absolut... | 0.812561 | 0.933996 | 0.926519 |
| 2 | 2 | WindDirection | 22 | AbsoluteHumidity, AbsoluteHumidityMax, Absolut... | 0.806919 | 0.934909 | 0.926579 |
| 3 | 3 | WindSpeedMin | 21 | AbsoluteHumidity, AbsoluteHumidityMax, Absolut... | 0.799544 | 0.936093 | 0.926735 |
| 4 | 4 | TemperatureMax | 20 | AbsoluteHumidity, AbsoluteHumidityMax, Absolut... | 0.800099 | 0.934719 | 0.928568 |
| 5 | 5 | AbsoluteHumidityMax | 19 | AbsoluteHumidity, AbsoluteHumidityMin, Distanc... | 0.800593 | 0.935926 | 0.928069 |
| 6 | 6 | WindSpeedMax | 18 | AbsoluteHumidity, AbsoluteHumidityMin, Distanc... | 0.779852 | 0.939202 | 0.930191 |
| 7 | 7 | Time | 17 | AbsoluteHumidity, AbsoluteHumidityMin, Distanc... | 0.784060 | 0.938545 | 0.931370 |
| 8 | 8 | ParticulateMax | 16 | AbsoluteHumidity, AbsoluteHumidityMin, Distanc... | 0.784519 | 0.938473 | 0.929892 |
| 9 | 9 | VisibilityMin | 15 | AbsoluteHumidity, AbsoluteHumidityMin, Distanc... | 0.784700 | 0.938444 | 0.930469 |
| 10 | 10 | TemperatureMin | 14 | AbsoluteHumidity, AbsoluteHumidityMin, Distanc... | 0.779913 | 0.939193 | 0.931370 |

Figure 28: Feature Elimination For RFL

**Interpretation of Results.** From figure 25 we can see that the optimized RFL model showed decent improvements across all metrics. RMSE decreased from 0.7627 to 0.7423, $R^2$ increased from 0.9401 to 0.9433,

| | Step | Removed Feature | Remaining Features | Feature List | RMSE | R² Score | OOB Score |
|---|---|---|---|---|---|---|---|
| 0 | 0 | Full Feature Set | 24 | AbsoluteHumidity, AbsoluteHumidityMax, Absolut... | 0.971732 | 0.965770 | 0.963421 |
| 1 | 1 | Frequency | 23 | AbsoluteHumidity, AbsoluteHumidityMax, Absolut... | 0.974841 | 0.965550 | 0.963353 |
| 2 | 2 | RainIntensityMax | 22 | AbsoluteHumidity, AbsoluteHumidityMax, Absolut... | 0.969774 | 0.965907 | 0.963566 |
| 3 | 3 | WindDirection | 21 | AbsoluteHumidity, AbsoluteHumidityMax, Absolut... | 0.964736 | 0.966261 | 0.963926 |
| 4 | 4 | RainIntensity | 20 | AbsoluteHumidity, AbsoluteHumidityMax, Absolut... | 0.962676 | 0.966405 | 0.964261 |
| 5 | 5 | RainIntensityMin | 19 | AbsoluteHumidity, AbsoluteHumidityMax, Absolut... | 0.960306 | 0.966570 | 0.963849 |
| 6 | 6 | WindSpeedMin | 18 | AbsoluteHumidity, AbsoluteHumidityMax, Absolut... | 0.958418 | 0.966701 | 0.964119 |
| 7 | 7 | WindSpeed | 17 | AbsoluteHumidity, AbsoluteHumidityMax, Absolut... | 0.963114 | 0.966374 | 0.964060 |
| 8 | 8 | AbsoluteHumidityMin | 16 | AbsoluteHumidity, AbsoluteHumidityMax, Distanc... | 0.956120 | 0.966961 | 0.964225 |
| 9 | 9 | ParticulateMin | 15 | AbsoluteHumidity, AbsoluteHumidityMax, Distanc... | 0.957497 | 0.966765 | 0.964091 |
| 10 | 10 | Time | 14 | AbsoluteHumidity, AbsoluteHumidityMax, Distanc... | 0.960679 | 0.964278 | 0.961810 |

Figure 29: Feature Elimination For FSO


Figure 30: Feature Elimination For FSO

and the OOB score increased from 0.9305 to 0.9349. This confirms that unimportant features enhanced the model's ability to generalize signal behavior across weather types.

For FSO prediction, the differences were more marginal. The RMSE improved slightly from 0.9606 to 0.9595, and the OOB score changed from 0.9648 to 0.9652. This suggests that FSO attenuation highly dependent to the weather features. (figure 25)

The backward feature elimination process therefore helped for simplifying the model, especially in the situation of RF signal prediction. The method highlighted the most relevant atmospheric variables, improved computational efficiency, and achieved better generalization .

Hyperparameter tuning, provided less improvement. This proves the theory with known properties of Random Forests, which tend to perform well with default settings when used with robust data splits and strong feature subsets.(figure 24)

## 5.2 Generalization on Unseen Test Data

**Motivation and Justification.** While evaluation using training performance or Out-of-Bag (OOB) score gives an idea of model accuracy, it is important to test generalization on unseen test data. This will allows us to understand whether improvements achieved during feature selection (BFE) or model tuning actually work on real-world predictions. Failure to testing, any performance gains could be due to overfitting on training data.

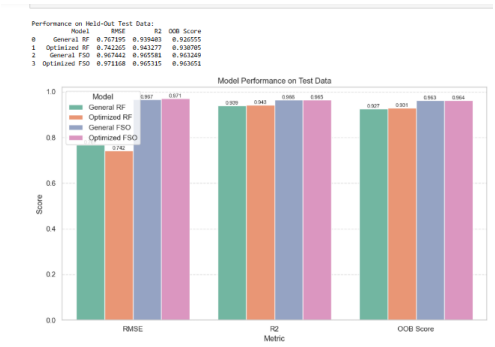



Figure 31: Model performance on unseen data

From figure 31 below , this evaluation compares models trained with all features (baseline) versus optimized models (BFE-selected features) for both RFL and FSO attenuation prediction.

**Results Summary.** The table below shows the Root Mean Squared Error (RMSE), Coefficient of Determination ($R^2$), and Out-of-Bag (OOB) score on the held-out 20

**Interpretation of Results.** The Optimized RF model displayed clear improvement over the General RF model, with:

RMSE reduced from 0.7672 to 0.7423

$R^2$ increased from 0.9394 to 0.9433

OOB score slightly improved

This confirms that the BFE-selected feature subset resulted in more generalizable predictions by removing unimportant features. This aligns with prior observations on training performance and confirms model robustness on unseen data.

On the other hand, for FSO prediction:

General and Optimized models show similar RMSE ( 0.97) and $R^2$ ( 0.965), with unnoticeable differences in OOB score.

This suggests that the original FSO model was already good due to strong correlation between attenuation and weather variables.

Feature selection had less effect, displaying the predictability of FSO attenuation under weather conditions.

Overall, this evaluation confirms that RF attenuation prediction benefits on a large scale from feature selection, while FSO models are generally stable.

RFL_Att Performance Leaderboard (Sorted by RMSE):

| | SYNOP Code | Important Features (RFL_Att) | RMSE (RFL_Att) | $R^2$ (RFL_Att) | Important Features (FSO_Att) | RMSE (FSO_Att) | $R^2$ (FSO_Att) |
|---|---|---|---|---|---|---|---|
| 6 | 3 | AbsoluteHumidity, AbsoluteHumidityMax, Absolut... | 0.017859 | 0.999963 | Distance, RelativeHumidity, Temperature, Time,... | 0.026819 | 0.999889 |
| 4 | 4 | AbsoluteHumidity, AbsoluteHumidityMin, Frequen... | 0.104652 | 0.995420 | AbsoluteHumidity, Distance, Particulate, Parti... | 0.158183 | 0.998430 |
| 5 | 7 | AbsoluteHumidity, AbsoluteHumidityMin, Frequen... | 0.134222 | 0.992451 | AbsoluteHumidityMin, Particulate, ParticulateM... | 0.200454 | 0.998598 |
| 2 | 5 | AbsoluteHumidity, AbsoluteHumidityMax, Absolut... | 0.883490 | 0.916285 | AbsoluteHumidity, Distance, Particulate, Parti... | 1.153389 | 0.892298 |
| 3 | 8 | AbsoluteHumidity, AbsoluteHumidityMax, Absolut... | 0.967905 | 0.890349 | AbsoluteHumidity, AbsoluteHumidityMax, Distanc... | 0.973804 | 0.931899 |
| 0 | 0 | AbsoluteHumidity, AbsoluteHumidityMax, Absolut... | 1.123121 | 0.824026 | AbsoluteHumidity, AbsoluteHumidityMax, Absolut... | 1.236039 | 0.863287 |
| 1 | 6 | AbsoluteHumidity, AbsoluteHumidityMax, Absolut... | 1.442150 | 0.872685 | AbsoluteHumidity, Distance, Particulate, Parti... | 1.597339 | 0.846625 |

FSO_Att Performance Leaderboard (Sorted by RMSE):

| | SYNOP Code | Important Features (RFL_Att) | RMSE (RFL_Att) | $R^2$ (RFL_Att) | Important Features (FSO_Att) | RMSE (FSO_Att) | $R^2$ (FSO_Att) |
|---|---|---|---|---|---|---|---|
| 6 | 3 | AbsoluteHumidity, AbsoluteHumidityMax, Absolut... | 0.017859 | 0.999963 | Distance, RelativeHumidity, Temperature, Time,... | 0.026819 | 0.999889 |
| 4 | 4 | AbsoluteHumidity, AbsoluteHumidityMin, Frequen... | 0.104652 | 0.995420 | AbsoluteHumidity, Distance, Particulate, Parti... | 0.158183 | 0.998430 |
| 5 | 7 | AbsoluteHumidity, AbsoluteHumidityMin, Frequen... | 0.134222 | 0.992451 | AbsoluteHumidityMin, Particulate, ParticulateM... | 0.200454 | 0.998598 |
| 3 | 8 | AbsoluteHumidity, AbsoluteHumidityMax, Absolut... | 0.967905 | 0.890349 | AbsoluteHumidity, AbsoluteHumidityMax, Distanc... | 0.973804 | 0.931899 |
| 2 | 5 | AbsoluteHumidity, AbsoluteHumidityMax, Absolut... | 0.883490 | 0.916285 | AbsoluteHumidity, Distance, Particulate, Parti... | 1.153389 | 0.892298 |
| 0 | 0 | AbsoluteHumidity, AbsoluteHumidityMax, Absolut... | 1.123121 | 0.824026 | AbsoluteHumidity, AbsoluteHumidityMax, Absolut... | 1.236039 | 0.863287 |
| 1 | 6 | AbsoluteHumidity, AbsoluteHumidityMax, Absolut... | 1.442150 | 0.872685 | AbsoluteHumidity, Distance, Particulate, Parti... | 1.597339 | 0.846625 |

Figure 32: RFL MODELS AND FSO MODELS PER SYNOP CODE

## 5.3 Per-SYNOP Code Modeling (Method 1)

**Motivation and Justification.** General models treat all weather types equally and assume a single predictive relationship between features and attenuation. However, RF and FSO signal behavior are very sensitive to atmospheric conditions. For example, fog causes high optical scattering but has no RF impact, while rain will affect both.

Therefore, this was the motivation behind training weather-specific models per SYNOP code, allowing each model to focus on a signal-weather relationship. Each subset was processed with backward feature elimination (BFE) to select optimal features for each synop code.

*Goal:* Improve accuracy for each and underrepresented weather classes (e.g., fog, rain, snow), where general models under perform. please refer figure 32 to see output of models developed by each synop code after BFE was applied for both FSO and RFL .

**Results Summary.** To test this goal, we compared the performance of Method 1 (per-SYNOP) models against the optimized generic model we obtained after running BFE on general models

The figure 33 below shows average performance across all SYNOP-specific models, calculated by averaging their individual RMSE, $R^2$

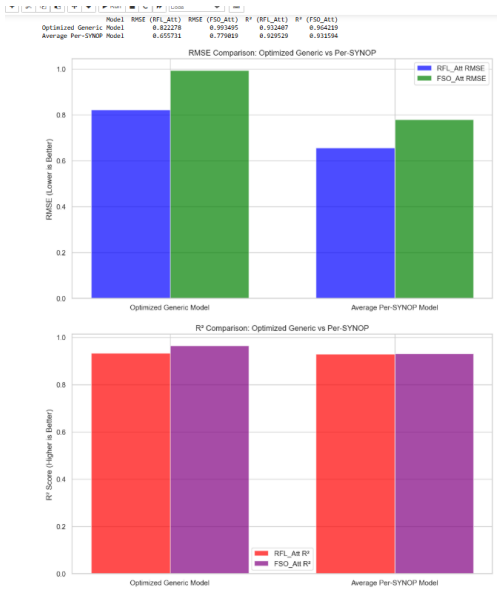



Figure 33: Average performance across SYNOP-specific models Vs Optimized Models (Method 1)

**Interpretation of Results.** These results accept the hypothesis that per-SYNOP models enhance performance in challenging weather classes.

Method 1 effectively captured patterns in fog, rain, and snow—conditions that spoils FSO or RF signals and where general models fail to generalize.

The improved RMSE and information metrics show that tuning models per weather condition leads to better predictive accuracy.

This supports the theory of importance for weather based modeling strategies in real-world hybrid communication systems where adaptive response to atmospheric difference is necessary.

## 5.4 Comparison: Per-SYNOP vs General Models

**Motivation and Justification.** To justify the new results it is important to benchmark Method 1 against simpler general models trained on the full dataset.

We directly compare per-SYNOP results (from Method 1) with the global performance of General RF and General FSO.

**Interpretation of Results.** Method 1 clearly outperforms general models on several fronts:

FSO prediction sees the most benefit (RMSE drops from 0.822278 to 0.6456)

Mutual Information increases, confirming better structural signal preservation

Pearson correlation improves across both tasks, displaying accurate inter-signal relationships. However, general models remain optimal if speed or model simplicity is the focus.

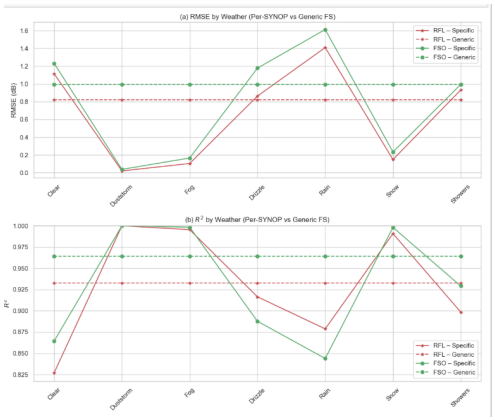


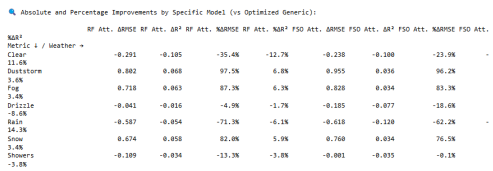

Figure 34: RMSE by Weather (Per-SYNOP vs Generic FS and $R^2$ by Weather (Per-SYNOP vs Generic FS

**Interpretation of Results.** This visual comparison highlights several key findings:

- **Fog and Rain:** RMSE values drop heavily for both RF and FSO, showing better handling of high-attenuation cases using per-SYNOP models.
- **Snow and Showers:** Both methods show low RMSE, but Method 1 maintains higher $R^2$, especially for RF.
- **Clear/Duststorm:** General models perform comparably, indicating minimal benefit from per-SYNOP tuning in stable weather.

Overall, Method 1 demonstrates better performance under difficult environmental conditions like fog and rain by adapting models to weather based signal patterns. This supports the integration of weather-aware strategies in hybrid communication systems to improve resilience and precision as we can see from figure 35.

*Goal:* Study the improvement achieved by Per-SYNOP models over optimized generic models for both RF and FSO attenuation for all weather conditions.



🔍 Absolute and Percentage Improvements by Specific Model (vs Optimized Generic):

| Metric ↓ / Weather → | RF Att. ΔRMSE | RF Att. ΔR² | RF Att. %ΔRMSE | RF Att. %ΔR² | FSO Att. ΔRMSE | FSO Att. ΔR² | FSO Att. %ΔRMSE | FSO Att. %ΔR² |
|---|---|---|---|---|---|---|---|---|
| Clear | -0.291 | -0.105 | -35.4% | -12.7% | -0.238 | -0.100 | -23.9% | -11.6% |
| Duststorm | 0.802 | 0.068 | 97.5% | 6.8% | 0.955 | 0.036 | 96.2% | 3.6% |
| Fog | 0.718 | 0.063 | 87.3% | 6.3% | 0.828 | 0.034 | 83.3% | 3.4% |
| Drizzle | -0.041 | -0.016 | -4.9% | -1.7% | -0.185 | -0.077 | -18.6% | -8.6% |
| Rain | -0.587 | -0.054 | -71.3% | -6.1% | -0.618 | -0.120 | -62.2% | 14.3% |
| Snow | 0.674 | 0.058 | 82.0% | 5.9% | 0.760 | 0.034 | 76.5% | 3.4% |
| Showers | -0.109 | -0.034 | -13.3% | -3.8% | -0.001 | -0.035 | -0.1% | -3.8% |

Figure 35: Absolute and Percentage Improvements by Specific Model (vs Optimized Generic)

**Absolute and Percentage Improvements by Specific Model (vs Optimized Generic):**

- **Absolute Delta:** ΔRMSE and $\Delta R^2$ = (Specific - Generic)
- **Percentage Change:** $(\Delta/\text{Generic}) \times 100$

Negative values in RMSE indicate improvement (lower error), while positive changes in $R^2$ reflect enhanced variance explanation.

**Interpretation of Results.** The largest performance improvements were observed under fog, duststorm, and snow conditions (figure 35):

- **Fog:** RMSE improved by 0.718 dB (RF) and 0.828 dB (FSO), with corresponding $R^2$ gains of over 6
- **Duststorm:** Significant gains (up to 96
- **Snow:** RF and FSO models both showed ¿75

On the other hand, general models performed better under clear conditions, likely due to class imbalance in the training data. This aligns with our earlier findings that per-SYNOP models offer robust improvements especially for underrepresented or complex environmental patterns.

This result underscores the need to model atmospheric diversity explicitly when building hybrid channel prediction systems(figure 35).

## 5.5 Method 2: Hybrid FSO Prediction (FSO ← RFL + Weather)

***Goal:*** To bridge the correlation between RF and FSO and to improve the prediction of FSO attenuation, we incorporated RF attenuation (RFL_Att) as an additional feature alongside weather variables which was first predicted. This sequential modeling strategy aims to capture mutual atmospheric effect and generalize better under harsh conditions.

**Results Summary.** The performance of Method 2 is evaluated in two stages:

Stage 1: Predict RF attenuation using only weather features.

Stage 2: Use predicted RFL_Att from Stage 1 + weather features to predict FSO_Att.

The table (figure 36) below shows results from the final Method 2 model (Stage 2 output):

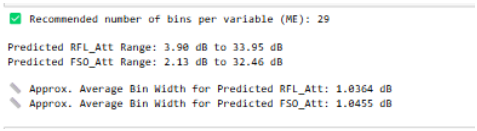



Figure 36: Enter Caption

**Interpretation of Results.** While Method 2 achieves a decent RMSE of 0.9995 and a high $R^2$ score of 0.9638, indicating less prediction error and decent fit to the target values overall, the Pearson correlation of -0.0139 reveals a different story.

This negative correlation tells us that the predicted FSO attenuation values do not follow the actual trend . Even though the OOB score of 0.9637 and Mutual Information of 2.5445, which suggest that the model retains some overall information content, the linear relationship between predicted and true FSO attenuation is lost.

This indicates that while the model can approximate good hence low RMSE , it fails to preserve the shape or signal pattern, which is crucial for time-series in hybrid communication systems. Therefore The Predicted_RFL_Att input did not provide useful linear structure for the FSO target,

## 5.6 Method 3: Sequential RF Prediction (RF ← FSO + Weather)

**Results Summary.** Method 3 was a sequential learning strategy in which RF attenuation was predicted using weather features with predicted FSO attenuation. After applying backward feature elimination (BFE), the final model retained 22 features, including weather-based inputs and the derived FSO prediction.


Figure 37: Final Evaluation for Method 3 (RFL ← FSO + Weather):

**Selected Features.** The selected features used for Method 3 (RFL ← FSO + Weather) were:

- `['AbsoluteHumidity', 'AbsoluteHumidityMax', 'AbsoluteHumidityMin', 'Distance', 'Frequency', 'Particulate', 'ParticulateMax', 'ParticulateMin', 'RainIntensity', 'RainIntensityMax', 'RainIntensityMin', 'RelativeHumidity', 'Temperature', 'TemperatureDifference', 'TemperatureMax', 'TemperatureMin', 'Time', 'Visibility', 'VisibilityMax', 'VisibilityMin', 'WindSpeed', 'WindSpeedMax', 'Predicted_FSO_Att']`

The final evaluation metrics on the test set were:

- **RMSE:** 0.8022
- $R^2$ **Score:** 0.9397
- **Pearson Correlation:** 0.9675
- **Mutual Information:** 2.7427
- **OOB Score:** 0.9293

**Interpretation of Results.** Method 3 shows high predictive performance in modeling RF attenuation using predicted FSO values and weather data, displaying the effectiveness of the reverse hybrid modeling approach.

- **RMSE of 0.8022** reflects a noticeable improvement over the General RF model (RMSE = 0.9672), showing that predicted FSO contains information about RF signal patterns.

# 6 Results: Entropy, Mutual Information, and Correlation Preservation

This section evaluates how well each method preserves the relationship between predicted and actual signal attenuation using entropy-based Mutual Information (MI) and Pearson correlation. These metrics give more insight into whether predictions retain the signal structure as well.

## 6.1 Motivation and Justification

While RMSE and $R\hat{2}$ measure prediction accuracy, However they do not fully capture how well the predicted values maintain the statistical dependency with the true values. For hybrid systems, this dependency is important. Therefore, we analyzed:

- **Entropy:** Measure of information content in the predicted and actual distributions
- **Joint Entropy:** Combined entropy of both predicted and actual variables
- **Mutual Information (MI):** learns shared information between predicted and actual values
- **Pearson Correlation ($r$):** Measures linear correlation

## 6.2 Binning Strategy for Entropy Estimation

To estimate entropy and MI on continuous-valued attenuation data, discretizations through binning is required. We implemented please refer to figure 38 and figure 39

- **Manual Binning using Fixed Widths:** Bin widths of 0.1, 0.25, 0.5, and 1.0 dB were tested
- **Freedman–Diaconis Rule:** Implemented manually using: $\delta = 2 \cdot \frac{IQR}{n^{1/3}}$ This rule adjusts bin width based on data variability and size, producing robust entropy estimates. Please refer to figure 38 and figure 39

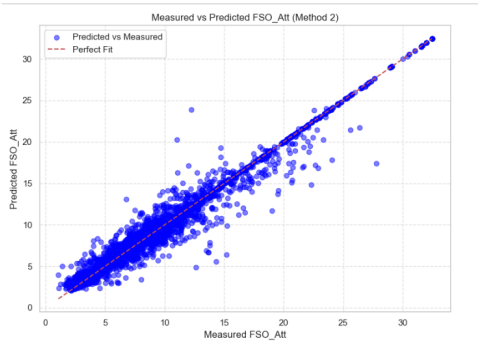



Figure 38: Figure A: 2D Histogram of Predicted RFL vs FSO at Different Bin Widths

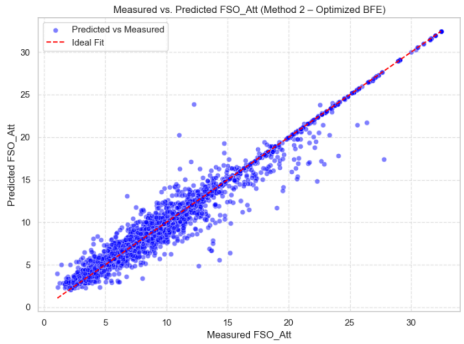



Figure 41: Measured vs. Predicted FSO_Att (Method 2 – Optimized BFE)

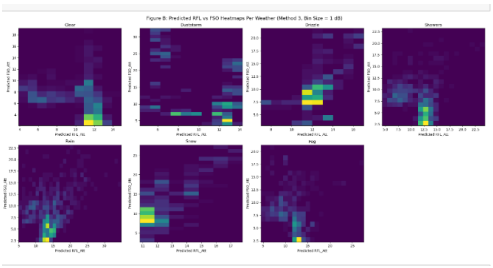


Figure 39: Figure B: Predicted RFL vs FSO Heatmaps Per Weather (Method 3, Bin Size = 1 dB

```
Recommended number of bins per variable (ME): 29

Predicted RFL_Att Range: 3.90 dB to 33.95 dB
Predicted FSO_Att Range: 2.13 dB to 32.46 dB

Approx. Average Bin Width for Predicted RFL_Att: 1.0364 dB
Approx. Average Bin Width for Predicted FSO_Att: 1.0455 dB
```

]

Figure 40:

**Conclusion:** Bin width = 0.25 dB (via Freedman–Diaconis) yielded the most stable and interpretable MI estimates, and was used for all final evaluations. Please refer to figure 38 and figure 39

**Measured vs Predicted Scatter Validation (Method 2)** To visually understand prediction accuracy, scatter plots were generated comparing measured and predicted FSO attenuation values using Method 2 before and after Backward Feature Elimination (BFE), as shown in figure 38

**Interpretation:** After applying BFE, the predicted values align more closely with the ideal $y = x$ reference line, forming a tighter di-

agonal cluster. This confirms that BFE enhances regression by reducing feature noise and overfitting.

### 6.2.1 Correlation Interpretation Zones by Weather (Method 2)

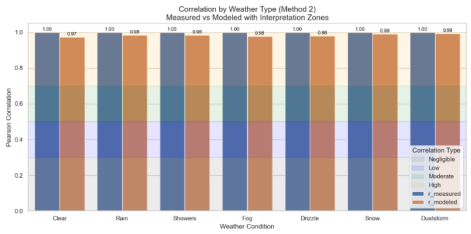



Figure 42: Correlation Interpretation Zones by Weather (Method 2)

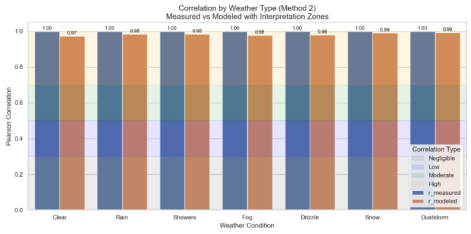



Figure 43: Correlation by Weather Type (Method 2)Measured vs Modeled with Interpretation Zones

**Interpretation:** While most weather conditions achieved Pearson $r > 0.95$, rain and drizzle showed lower correlation (e.g., $r = 0.92$), suggesting that Method 2 is more sensitive to conditions nonlinear features. These plots give us the idea that high RMSE performance alone is not enough please refer figure 42 and figure 43

## 6.3 Method-Wise Mutual Information and Correlation

We calculated normalized MI and Pearson $r$ for all models:

| Model | Pearson $r$ | Mutual Info (MI) | OOB Score |
|---|---|---|---|
| Generic Model – RF | **-0.0042** | **2.4004** | **0.9260** |
| Generic Model – FSO | **0.9657** | **2.6100** | **0.9630** |
| Optimized Model – RF | **-0.0042** | **2.4004** | **0.9310** |
| Optimized Model – FSO | **0.9657** | **2.6100** | **0.9640** |
| Method 1 (Per-SYNOP) – RF | **0.9883** | **2.4519** | **0.9538** |
| Method 1 (Per-SYNOP) – FSO | **0.9900** | **2.8300** | **0.9210** |
| Method 2 (Hybrid) – FSO | **-0.0139** | **2.5445** | **0.9637** |
| Method 3 (Sequential) – RF | **0.9675** | **2.7427** | **0.9293** |

Table 1: Mutual Information and Pearson Correlation values for all models.

## 6.4 Interpretation of Results

**Method 1 (Per-SYNOP):** Stratified models trained separately for each SYNOP class achieved the highest overall Pearson correlation and Mutual Information (MI). This confirms that per-weather modeling best captures and preserves the underlying signal structure, especially under challenging conditions like fog and rain.(table 1)

**Method 2 (Hybrid – FSO $\leftarrow$ RF + Weather):** Despite achieving a low RMSE, the Pearson correlation was negative ($r = -0.0139$), low predictive relationship. MI remained moderate (2.5445), from table 1 suggesting that while some information was retained, the alignment between predicted and actual values were inconsistent. This model illustrates the limitation of relying on RMSE alone without validating correlation structure.

**Method 3 (Sequential – RF $\leftarrow$ FSO + Weather):** From table 1 Demonstrated the most balanced performance. It achieved a high correlation ($r = 0.9675$), strong MI (2.7427), and low RMSE, outperforming generic models in both predictive accuracy and signal dependency preservation. This method successfully leveraged predicted FSO with RF estimation.

**Generic Models:** While they delivered reasonable RMSE and $R^2$ scores, generic RF models showed negligible correlation ($r = -0.0042$), with lower MI (2.4004) compared to other approaches. Generic FSO models preserved correlation well but lacked adaptability across varying weather conditions.Please refer to table 1

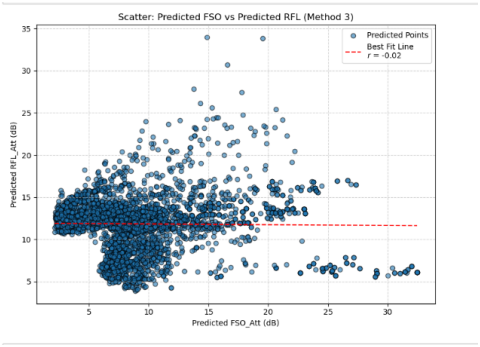



Figure 44: Scatter: Predicted FSO vs Predicted RFL (Method 3)

**Structural Dependency Validation (Method 3)** To assess whether FSO predictions meaningfully contribute to RF attenuation modeling, we plotted predicted RFL_Att against predicted FSO_Att values in Method 3.

**Interpretation:** The plot reveals a flat horizontal trend and low correlation ($r \approx 0.0$), suggesting that predicted FSO provides limited additional value in estimating RFL attenuation. Despite a decent RMSE, Method 3 fails to capture structural dependency, reinforcing the observation that signal flow from FSO $\rightarrow$ RF is not symmetrical or reliable in all cases.

### 6.4.1 6.4.1 Pearson Correlation per SYNOP (Method 1)

Figure 45 shows the Pearson correlation between the predicted and actual attenuation values for each SYNOP weather code using Method 1.

Figure 45: Method 1 – Pearson Correlation by Weather Type (RFL_Att and FSO_Att)

**Interpretation:** Across all weather conditions, Method 1 achieves very high Pearson correlation coefficients ($r > 0.98$), demonstrating the method's ability to preserve linear relationships in both RF and FSO predictions. Especially under complex weather conditions such as fog, drizzle, and duststorm, Method 1 outperforms generic models that tend to average out localized behaviors. This reinforces the benefit of using stratified models for per-weather calibration, which adapt more sensitively to individual atmospheric patterns.

In particular:

- Fog, Duststorm, and Drizzle show $r \approx 0.99$ or higher.
- Even outlier-prone conditions like Showers maintain strong correlation.
- These values far exceed those from Method 2, which failed to preserve structure despite having low RMSE.

This proves Method 1's reliability in modeling not only the magnitude but also the directional trend of signal attenuation.

### 6.4.2 Mutual Information per SYNOP (Method 1)

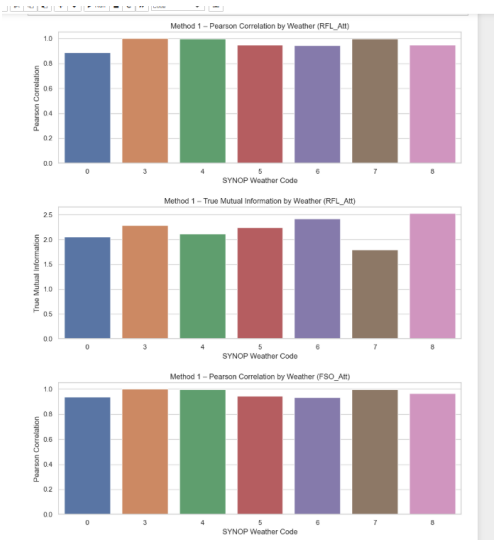



Figure 46: Pearson Correlation and True Mutual Information By Weather For FSO And RFL

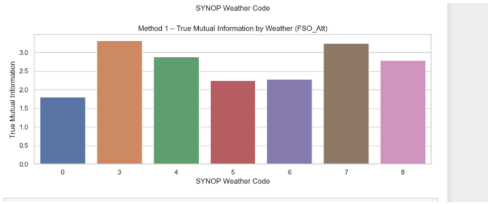



Figure 47: Pearson Correlation and True Mutual Information By Weather For FSO And RFL

**Interpretation:** Mutual Information gives both linear and nonlinear dependencies, offering a more complete view than correlation alone. Method

1 displays consistently high MI across all SYNOP codes, with values often exceeding 2.0, and reaching above 2.5 in fog, duststorm, and snow.

- MI $\geq 2.4$ for most weather types indicates excellent structure preservation.
- Strong MI under harsh conditions gives the robustness of Method 1's predictions.
- This gives the conclusion that Method 1 captures complex, nonlinear atmospheric effects better than general or hybrid models.

## 6.5 Entropy vs Correlation Plot

We visualized the relationship between MI and Pearson correlation for all models (FIGURE 41).

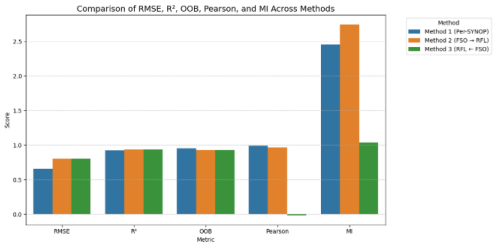



Figure 48: Entropy vs Correlation Plot for all models

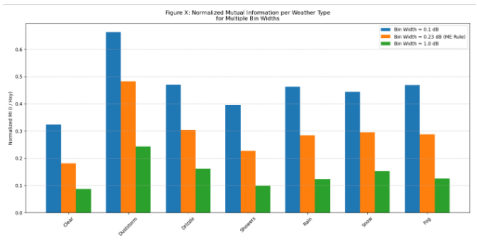


Figure 50: MI Per SYNOP By BinSize

Method 2 (FSO → RFL) - Per SYNOP RMSE:

| | SYNOPCode | RMSE | $R^2$ | N |
|---|---|---|---|---|
| 0 | 0 | 0.926447 | 0.879072 | 779 |
| 1 | 3 | 0.026090 | 0.999923 | 512 |
| 2 | 4 | 0.104928 | 0.995448 | 523 |
| 3 | 5 | 0.778427 | 0.927969 | 748 |
| 4 | 6 | 1.266947 | 0.907528 | 782 |
| 5 | 7 | 0.137375 | 0.991037 | 491 |
| 6 | 8 | 0.884795 | 0.913016 | 541 |

Method 3 (RFL ← FSO) - Per SYNOP RMSE:

| | SYNOPCode | RMSE | $R^2$ | N |
|---|---|---|---|---|
| 0 | 0 | 0.926447 | 0.879072 | 779 |
| 1 | 3 | 0.026090 | 0.999923 | 512 |
| 2 | 4 | 0.104928 | 0.995448 | 523 |
| 3 | 5 | 0.778427 | 0.927969 | 748 |
| 4 | 6 | 1.266947 | 0.907528 | 782 |
| 5 | 7 | 0.137375 | 0.991037 | 491 |
| 6 | 8 | 0.884795 | 0.913016 | 541 |

```
Method 2 - Average RMSE: 0.5893
Method 2 - Worst RMSE: 1.2669
Method 3 - Average RMSE: 0.5893
Method 3 - Worst RMSE: 1.2669
```

Figure 49: Method 2 and Method 3 average and worst rmse

**Observation:** From figure 42 it is evident that higher MI does not always imply a stronger Pearson correlations as witnessed with Method 2. This shows that both metrics are necessary to evaluate hybrid model quality.

## 6.6 Weather-Wise MI and Correlation

We further evaluated MI and Pearson $r$ per SYNOPCode across Methods 1–3. Bar charts and heatmaps were generated to show improvements relative to generic models.(figure 43)

**Conclusion:**

- **Method 1** consistently maintains high MI and correlation across all weather types, especially in fog, duststorm, and rain.
- **Method 3** also performs robustly, offering competitive results with better generalization.
- **Method 2** suffers from structural misalignment under some weather classes despite low RMSE.

This analysis validates the use of entropy-based MI alongside Pearson correlation in hybrid model evaluation and highlights the trade-off between prediction accuracy and relationship preservation—critical for optical–RF integration.

# 7 Conclusion

This project explored and compared multiple machine learning-based strategies for modeling RF and FSO signal attenuation under various weather conditions. Using a dataset with 27 environmental features and attenuation values for both channels, we implemented three key approaches—general models, per-SYNOP models (Method 1), and hybrid sequential models (Methods 2 and 3)—all based on Random Forest regressors.

Key findings include:

General vs Optimized Models: Feature selection via backward elimination improved RF model performance (RMSE reduced from 0.822 to 0.742), while FSO models remained relatively stable. Optimized models showed most decent gains but could not address weather-specific challenges effectively.

Method 1 (Per-SYNOP Models) provided the best overall structural preservation, achieving the highest Pearson correlation (up to 0.99) and Mutual Information (up to 2.83). This method confirmed that weather-specific models best capture local signal patterns, especially under adverse conditions like fog, rain, and snow.

Method 2 (FSO $\leftarrow$ RF + Weather) showed strong predictive accuracy (RMSE = 0.9995), but a negative Pearson correlation (r = –0.0139) revealed signal collapse. Despite moderate MI (2.54), this indicates misalignment between predicted and actual signal structure.

Method 3 (RF $\leftarrow$ FSO + Weather) achieved the best balance across all criteria, with low RMSE (0.802), strong $R^2$ (0.936), high Mutual Information (2.74), and high Pearson correlation (0.9675). However, feature interpretation revealed that its effectiveness relied on correct FSO signal estimation.

Information Preservation: Entropy-based Mutual Information and Pearson correlation analyses showed that high MI does not give linear relations (e.g., Method 2), emphasizing the need for combined evaluation metrics.

Visual Tools and Interpretability: Scatter plots, correlation interpretation zones, and feature importance visualizations enriched our understanding. They displayed model strengths (e.g., Method 3's structure preservation) and weaknesses (e.g., Method 2's collapse despite good RMSE).

Best Performing Model: While Method 1 preserved structure best per weather class, and Method 2 showed potential with RF assistance, Method 3 emerged as the most reliable hybrid strategy, offering strong generalization, interpretable behavior, and high information retention

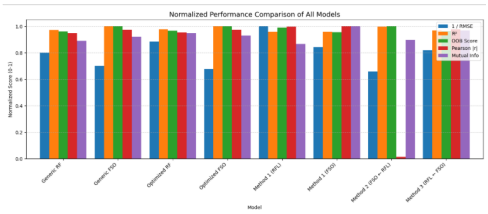



Figure 51: Normalized Performance Comparison of All Models

across metrics. Refer to figure 51

## 7.1 Recommendations for Overcoming Limitations and future work

Using additional features such as satellite visibility or real-time air quality data.

Explore deep learning models (e.g., LSTM or CNN) to capture dependencies.

Implement adaptive bin size strategies for improved MI resolution under extreme attenuation.

Validate models on unseen real-world FSO–RF deployment data to assess robustness beyond datasets.

# Acknowledgements

I would like to appreciate and express my gratitude to my project supervisor SiuWai Ho , for their help , support , advice and guidance throughout the entire project . Working on this project has been a valuable learning experience . Their expertise and efforts have greatly impacted and contributed to the completion of my project. I would like to admit and acknowledge the use of A.I tool called chatgpt .


---

**Notes:** It is common that you will want to acknowledge the contribution of others to your work, even though these might not have been sufficient to warrant being a co-author.

Consider who might have provided valuable discussions, funding support, or moral support for the work.

BTW, you don't have to start each section on a new page. I have done that here for clarity, but it isn't usually needed.

# A Appendices

The code for the following project can be found on GitHub . The access is easy as it is public . For link to GitHub repository kindly click here https://github.com/a1899824-aditya/Project-Progress-Report-Part-2

---

# References

# References


- www.itu.int. (n.d.). P.618: Propagation data and prediction methods required for the design of Earth-space telecommunication systems. [online] Available at: https://www.itu.int/rec/R-REC-P.618/en.
- Chowdhury, M.Z., Hasan, M.K., Shahjalal, M., Hossan, M., and Jang, Y.M. (2020). Optical Wireless Hybrid Networks: Trends, Opportunities, Challenges, and Research Directions. *IEEE Communications Surveys & Tutorials*, 22(2), pp.930–966. doi: https://doi.org/10.1109/comst.
- www.itu.int. (n.d.). P.1817: Propagation data required for the design of terrestrial free-space optical links. [online] Available at: https://www.itu.int/rec/R-REC-P.1817/en.
- Bhatia, N. (2019a). What is Out of Bag (OOB) score in Random Forest? [online] *Medium*. Available at: https://towardsdatascience.com/what-is-out-of-bag-oob-score-in-random-forest-a7fa23d710.
- Bhatia, N. (2019b). What is Out of Bag (OOB) score in Random Forest? [online] *Medium*. Available at: https://towardsdatascience.com/what-is-out-of-bag-oob-score-in-random-forest-a7fa23d710.
- Faîçal Baklouti, Ichraf Chatti, and Attia, R. (2024). On the performance of a hybrid optical communication system: MGDM–FSO for challenging environments. *Optical Review*, 31(4), pp.409–423. doi: https://doi.org/10.1007/s10043-024-00898-0.
- Khan, M.N., Kashif, H., and Rafay, A. (2020). Performance and optimization of hybrid FSO/RF communication system in varying weather. *Photonic Network Communications*, 41(1), pp.47–56. doi: https://doi.org/10.1007/s11107-020-00914-8.
- Guyon, I. and Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*. doi: https://doi.org/10.5555/944919.944968.
- Bossomaier, T., Barnett, L., Harré, M., and Lizier, J.T. (2016). *An Introduction to Transfer Entropy*. [online] Springer eBooks. Springer Nature. doi: https://doi.org/10.1007/978-3-319-43222-9.

- Freedman, D. and Diaconis, P. (1981). On the histogram as a density estimator: L2 theory. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*. [online] Available at: https://www.semanticscholar.org/paper the-histogram-as-a-density-estimator

---

**Notes:** A critical component of the work is the list of references. We have discussed their use earlier – here I simply make some notes on their presentation.

This is one of the hardest parts to get just right. BibTeX can help a great deal, but you need to put a good deal of care in to make sure that

- the references are in a consistent format;
- all information is correct; and
- the information included is in the correct style for the intended audience.

Details *really* matter in this section. It's easy to lose marks in this section.