Atmosphere-Ocean

Taylor & Francis
Taylor & Francis Group

Weather Research and Forecasting (WRF) model and machine learning algorithms to improve marine fog predictions in areas offshore Atlantic Canada.

Journal:	*Atmosphere-Ocean*
Manuscript ID	AO-2025-0041.R1
Manuscript Type:	Applied Research / Recherche appliquée
Date Submitted by the Author:	n/a
Complete List of Authors:	Teeloku, Piyush; York University Faculty of Graduate Studies, Earth and Space Science and Engineering Chen, Zheqi; York University Taylor, Peter ; York University, Chen, Yongsheng; York University
Keywords:	WRF, NWP model, marine fog, machine learning, weather forecasting, North Atlantic, ERA5

Abstract:

In recent years, machine learning (ML) has gained popularity in the field of weather forecasting, particularly in areas where numerical weather prediction (NWP) models face challenges. One such area is fog prediction. Reduced visibility due to fog events poses serious challenges to transportation and public safety. A key factor in improving fog forecasts is the accurate representation of microphysics processes, which play a crucial role in fog formation and dissipation. Traditional NWP models face limitations in accurately forecasting fog due to those complex microphysical processes. In this study, we propose a post-processing approach that combines the Weather Research and Forecasting (WRF) model forecasts with a machine learning classifier to improve fog prediction by distinguishing between fog and no-fog conditions. Using twelve years of data (2012–2023) for St John's, Newfoundland and Labrador, and Yarmouth, Nova Scotia, Canada, the approach was tested on independent 2024 observations. Compared to forecasts based solely on liquid water content from WRF, the ML model achieved higher skill, with F1-score improvements of 13% at St John's and 18% at Yarmouth. According to the results obtained, this approach demonstrates the potential of ML techniques to enhance operational fog forecasting capabilities.

1. Introduction

The weather conditions for fog to occur can be quite mild, but when it happens, it can become quite disruptive due to the visibility drop, therefore aviation and marine industries are very interested in how fog will affect their operations (Flynn, 2018). It is estimated to be the second most common cause of weather-related aviation accidents behind strong winds (National Center for Atmospheric Research, 2022). Beyond its impact on transportation, fog plays a crucial role in atmospheric and ecological processes (Koračin and Dorman, 2017). It carries water droplets containing ions, aerosols, and microorganisms, which influence coastal ecosystems' hydrological, thermodynamic, nutrient, and toxicological properties. In addition, fog regulates temperature and moisture conditions, shaping vegetation patterns and forest development in coastal regions.

There have been various NWP approaches to modeling and predicting marine fog. Taylor et al. (2021) found that the amount of liquid water is too high in the WRF model. They proposed a roughness length for fog droplets and allowed them to deposit to the surface through turbulence

in the model. The modified model was run for North Atlantic during the summer 2018 on a domain extending from eastern Canada out beyond the Grand Banks and including Sable Island. They were able to have better liquid water content (LWC) as well as visibility representation compared with observations. However, they did not guarantee that this modification would improve the prediction of the onset and dissipation of fog. More recently, Gao et al. (2024) used another NWP approach where they used microwave radiometer-retrieved cloud water path observations to improve numerical simulations of sea fog over the Yellow Sea in May 2018. They used an ensemble method with WRF and a Grid-point Statistical Interpolation/EnKF assimilation system. Then, they had more accurate initial conditions by retrieving the cloud water content from the satellite. By doing so, they were able to make improvements on sea fog forecasting. Their method worked very well with areas of high chance of fog but had an average false alarm rate of 0.463, meaning that they were overestimating the number of fog occurrences by roughly 46%. So currently NWP models are overestimating the LWC and doing a poor job at predicting fog occurrence.

Gultepe et al. (2024) applied machine learning techniques to nowcast marine fog visibility using observational data collected during the Fog And Turbulence Interactions In the Marine Atmosphere (FATIMA) campaign in July 2022. The campaign took place in the Grand Banks region and near Sable Island off the northeast coast of Canada, using measurements from the research vessel Atlantic Condor. Key meteorological variables were used alongside derived droplet number concentration to analyze fog characteristics. Nowcasting was performed for lead times of 30 and 60 minutes using support vector regression (SVR), least-squares boosting (LSB), and deep learning models, targeting visibility thresholds below 1 km, and below 10 km. Their results emphasize the value of ship-based observations and ML models in supporting short-term fog forecasting in complex marine environments.

Related work has demonstrated the effectiveness of machine learning for short-term fog and low-visibility forecasting. Castillo-Botón et al. (2022) investigated low-visibility event prediction using 23 months of data from the Mondoñedo weather station in Galicia, Spain. They applied both regression and classification approaches for 30 minutes nowcasting of fog, demonstrating the utility of supervised learning techniques in both continuous and categorical forecasting contexts. Building on this, Peláez-Rodríguez et al. (2023) incorporated the ECMWF Reanalysis v5 (ERA5, Hersbach et al., 2020) data alongside local station observations to improve fog-related visibility predictions at the same site. Their study explored both traditional ML models

and evolutionary algorithms for short-term forecasts (1, 3, and 6 hours), reporting up to a 17% improvement in forecast accuracy. These findings support the value of integrating reanalysis data and advanced learning methods in visibility and fog prediction tasks.

Another sea fog ML approach was proposed by Chen et al. (2024) who developed a deep learning framework to improve the short-term prediction of sea fog regions in the Yellow Sea, with an emphasis on maritime navigational safety. The study introduced the Multivariable Sea Fog Forecast (MV-SFF) dataset, which contains 122 sea fog events from 2010 to 2020. Each event includes hourly geostationary satellite images and reanalysis-based meteorological variables, providing a rich data source for spatial fog forecasting. Using this dataset, the authors proposed the Rich-Element Aggregated (REA) model: a deep learning architecture designed to extract and integrate diverse meteorological and satellite-derived features across different times and spatial scales. The REA model includes a position-aware edge detection mechanism to accurately localize sea fog regions. A seven-hour forecast starting from 09:16 local time demonstrated strong performance, with the REA model outperforming other state-of-the-art deep learning methods in both accuracy and consistency. This work highlights the potential of combining remote sensing with deep learning to enhance spatial sea fog forecasting capabilities.

In this study, we decided to take an alternative approach by using both WRF and machine learning to improve the hourly forecast of fog. Binary classification was chosen as the machine learning framework to post-process numerical output from the WRF model at St John's Newfoundland and Labrador, and Yarmouth, Nova Scotia, Canada, due to its proven effectiveness in prior research. We used data from 2012 to 2023 since 2012 is the earliest time with available visibility reports at the weather stations and used only summer data from April to August for each year. This avoids complications with blowing snow/ice, and we are looking mainly at advection fog in the coastal areas where the highest frequency of fog in these regions is during the summer.

The binary approach was used to identify fog events (typically defined by visibility below 1 km) while accounting for the inherent variability in visibility measurements. From other research, we observed that using ERA5 data for the meteorological features helped increase the performance of the ML approaches and compared to other research that used features to complement their ML models, we used a post-processing approach. Many existing studies have employed

classification techniques to distinguish between fog and no-fog conditions and have reported reliable performance in identifying fog onset and duration. Classification is also particularly advantageous when the primary operational need is to issue timely alerts for fog events, rather than estimating precise visibility values. Our approach builds on this by integrating WRF model forecasts with site-specific visibility observations to improve local fog prediction. Unlike many previous works that focus on very short-term forecasts (e.g., 30 to 60 minutes or up to 6-hour lead time), our method extends the forecast horizon to 24 hours and provides predictions at an hourly resolution, offering greater lead time for decision-making. This also enables us to identify when fog occurs during the day and potentially how long it can last compared to using only time windows of 3 hours, 6 hours and 12 hours. Additionally, by using a high-resolution NWP model like WRF, our method benefits from better meteorological context, particularly in capturing key physical drivers of fog. While this approach does not resolve the full areal extent of fog, the combination of WRF-based physical guidance and site-specific ML classification offers an optimal balance between accuracy, interpretability, and operational utility for localized fog forecasting.

2. Data collection

2.1. Features

We chose features that are relevant to marine fog occurrence. Gultepe et al. (2024) tested ten different variables and performed a pairwise correlation analysis that was visualized on a heat map. So, we did the same process with a Pearson correlation analysis for 2 m temperature, 2 m relative humidity, 10 m U and V wind, land surface pressure, hour and month (Table 1). Values of the features are from the ERA5 dataset. ERA5 combines past observations with output from the European Centre for Medium-Range Weather (ECMWF) model to provide a consistent historical record of atmospheric, oceanic, and land variables from 1940 to the present. It has a resolution of 31 km and 137 vertical levels and is available every hour. For convenience, values are extracted using the WRF Preprocessing System (WPS) which is used for preparing input files for real-case WRF simulations. ERA5 does not contain 2 m relative humidity. Thus, it was calculated using an NCAR Command Language (NCL) function relhum_ttd with temperature and dew point from ERA5. Fig. 1 shows the domain settings used and the locations of the weather stations. The same domain setting is also used for running WRF for prediction.

We carefully selected a smaller set of input features (five variables in total) compared to previous studies. For instance, Gultepe et al. (2024) used ten features, including dewpoint temperature, dewpoint depression and wind direction. However, our heatmap correlation analysis showed that these variables were highly correlated with others such as relative humidity and the wind components (U10 and V10) and therefore added little new information. Another commonly used variable in earlier studies is visibility itself, often included as a lagged feature (e.g., visibility one or two hours prior, depending on the desired forecast lead time). We could have chosen this variable as a feature as well, but we were focusing on getting long-term forecasting periods and preliminary results from our models showed it benefited the nowcasting approach mainly (around 3-hour lead time).

In Yarmouth, most features show moderate to strong correlation with visibility, particularly relative humidity, while the U component at 10 m is weakly correlated. In St John's, correlations with temperature variables are generally weaker, though U10 and V10 wind components show moderate association. Relative humidity remains strongly correlated at both sites. Since we are dealing with time series data, time-based features are valuable to have and so we parsed the Time data into 2 variables: month and hour. By including these features, we were able to analyze the data based on different time components, such as looking at trends over specific hours of the day or different days. These features are useful in training ML models that need to capture time-based patterns, especially weather patterns that can last for hours, like fog.

From the heatmap on Fig. 2, we do not see much correlation between the time variables and visibility and so decided to look at the trends visually. From Fig. 3 and Fig. 4, we see that the monthly trends for both locations vary. St John's has the highest monthly occurrence in April to June, while Yarmouth has a higher occurrence in the later months, July and August. From this graph, we were able to identify that the month feature is important to have. Fig. 5 and Fig.6 show the line plots for fog occurrence count in a day for both locations. From both plots, we see the diurnal cycle influencing fog occurrence with daytime heating burning off most of the fog in the early mornings. The highest fog occurrence at both locations is around nighttime and very early mornings (till around 7 a.m. for both local time). Therefore, the monthly fog occurrence and the hour at which fog happens are valuable to learn the behavior of fog, and by having those time-based features, the ML model can identify any useful traits to make better predictions.

Another important aspect is to use lagged features. Although data in our models at each hour are treated separately, lagged features are useful since they can capture historical dependencies and trends of the continuous meteorological conditions. Lagging a time series feature means shifting the current value forward in time by a desired number of steps. For each of the meteorological variables, its value at one hour before the predicted time (lag 1) is also used as a feature. It is possible since our goal is to use the 24-hour forecast of the variables from WRF, the lag 1 variables also come from WRF forecasts. To summarize, the features include five meteorological variables at the forecast time (T2, RH2, U10, V10 and P_sfc), those five meteorological variables 1 hour before the forecast time and 2 variables of time in UTC (month and hour).

2.2. Target

Visibility data we used are at St John's International Airport (47^{0}37'07" N, 52^{0}45'09" W with an elevation of 140.5 m) and at Yarmouth Airport (43^{0}49'37" N, 66^{0}05'17" W with an elevation of 42.9 m), collected by NAV CANADA (NAVCAN, 2017). These sites are influenced by marine air masses and complex coastal terrain, both of which are key contributors to fog formation. Both St John's and Yarmouth are situated along the Atlantic coast of Canada, where the interaction of major ocean currents, the warm Gulf Stream and the cold Labrador Current, creates ideal conditions for sea advection fog. In this region, warm, humid air masses transported by the Gulf Stream often move over the colder waters carried southward by the Labrador Current (Bullock et al., 2016). This temperature contrast is particularly pronounced during the spring and early summer months, when sea surface temperatures remain low while atmospheric temperatures begin to rise. As a result, dense fog frequently develops offshore and is advected inland by prevailing winds, especially under stable marine layer conditions. The convergence of these ocean currents is therefore a critical driver of marine fog formation and plays a central role in the fog climatology of both coastal sites. According to the reports, there are a lot of hours of rain combined with fog, therefore rain conditions were not excluded from this study.

From the Glossary of Meteorology (2025), fog is defined as a collection of tiny water droplets suspended in the air near the Earth's surface which reduces visibility to below 1 km (0.62 miles). For binary classification, by using the visibility reports collected, every time visibility dropped below 1 km (≤ 1 km), we classified it as 'fog' and the rest were classified as 'clear'. One important aspect of classification is figuring out how the categorical labels are distributed. Our dataset is an imbalanced one with fog being the minority class. For both locations, the percentage of fog data is around 10-12% so it is a moderate case of imbalanced data. This enables us to understand our visibility data better and allows us to take measures such as using oversampling and undersampling techniques or balancing the class weights of our labels when preparing our data to run our model properly.

Before training the machine learning model, we first considered whether to build a single model for both locations or treat them separately. Given the geographic and environmental differences, with Yarmouth's lower elevation (42.9 m) compared to St John's (140.5 m), we anticipated that oceanic and atmospheric influences could vary between the two sites. Heatmap analyses, along with observed differences in monthly and hourly weather patterns, supported this assumption. As a result, we chose to model the two locations separately, allowing each model to better capture the distinct weather dynamics of its respective region.

3. ML Model

3.1. ETClassifier

Extremely Randomized Trees (ExtraTrees) is an ensemble method proposed by Geurts Pierre (2006) that constructs multiple decision or regression trees, using more randomness in its technique to improve robustness and computational efficiency. ExtraTrees combine predictions from multiple unpruned (not optimized) trees for classification or regression but introduces additional randomness in the splitting process. In ExtraTrees, both the feature and the cut-point for splitting at each node are chosen at random. Even if the split threshold is chosen at random, the ExtraTrees will still try to get the best score. The ExtraTrees model can be used for

classification (ETClassifier). To optimize the ETClassifier model, we will use the RandomizedSearchCV function from Scikit-Learn (Pedregosa et al., 2011) with a TimeSeriesSplit cross-validation strategy tailored for time-series data. This approach ensures that the hyperparameter tuning respects the temporal order of the data, preventing data leakage from future observations that influence the training process. This comprehensive randomized search, combined with cross validation based on time series, will allow us to identify the best combination of hyperparameters for the ETClassifier.

As expected, the optimal hyperparameters for St John's and Yarmouth are different. The hyperparameters for both locations are listed in Table 2. The number of estimators (ensemble of trees built) is 105 for St John's and 125 for Yarmouth, suggesting a slightly more complex relationship between the training data for Yarmouth. The criterion used to split the nodes are log_loss for St John's and entropy for Yarmouth. We do not need the bootstrap method for both, meaning that all the training data were used for each tree. The class weight for St John's was balanced, and for Yarmouth, it was balanced_subsample. They work similarly by changing the weights so that both classes are weighed the same in the model. The maximum depth for the trees to grow are None for both locations, meaning that they will continue to grow until the threshold of the minimum sample split is met for a leaf node. This condition is 4 samples for St John's and 3 in Yarmouth. The minimum samples to determine a node is 8 for St John's and 4 for Yarmouth and finally, the maximum features selected for such splitting was square root (sqrt) of the number of features we had for St John's, and all the features were used (None) for Yarmouth.

3.2. LSTM

Hochreiter and Schmidhuber (1997) introduced the Long Short-Term Memory (LSTM) network, a type of Recurrent Neural Network (RNN) that uses gates (forget, input and output gates) to control information flow and manage long-term dependencies. For our bidirectional LSTM (biLSTM) model, input data are reshaped to [batch size, time steps, features], where time step is 1 and features total 14 (including lagged features). The model begins with an input layer followed by two BiLSTM layers with 64 and 32 units, respectively, both using tanh activation. Batch Normalization and Dropout (0.2) are applied after each LSTM and dense layer to improve stability and reduce overfitting. Two dense hidden layers follow, using LeakyReLU activation to maintain gradient flow. The final output layer is a single sigmoid-activated neuron for binary classification. The model has 115,861 parameters (38,513 trainable), well-matched to our dataset size of 33,000 samples.

We carefully configured the model by monitoring the loss function between the training and validation sets. For binary classification, we use the binary cross entropy loss function which calculates the log loss between the true labels and the predicted probabilities. We used the Adam optimizer with a learning rate of 0.0001 to update the weights based on the gradients calculated from the loss function. We found that a learning rate of 0.0001 ensured stability in the training process and it was not too small that the model took too long to converge. Since processing individual data samples is very expensive computationally, we use a batch size of 128 and we set the epochs to 200. Setting the epoch to 200 allowed the model to run through the entire training dataset enough time to learn more effectively and converge to a better solution. To reduce overfitting, we also used EarlyStopping which monitors the validation F1 score and halts the training if there is no improvement for 20 epochs.

3.3. XGBoost

XGBoost (eXtreme Gradient Boosting, Chen and Guestrin, 2016) is a model based on decision trees. It uses gradient boosting where in each iteration a new tree is constructed to fix the errors of the previous tree (Hsieh 2023). Logistic regression is chosen as the loss function. According to the recommended calculation (hours of clear divided by hours of fog), scale_pos_weight is 7.97 for St John's and 5.75 for Yarmouth. With this hyperparameter, the model will be punished more when it misclassifies a fog event as clear, so that it focuses more on the rarer cases of fog. Other hyperparameters are tuned by cross validation and are summarized in Table 3. Max_depth and min_child_weight are the two main hyperparameters of XGBoost. Max_depth controls the maximum number of layers that a tree can have. A larger value will lead to a more complex model and more likely to overfit. On the other hand, min_child_weight controls the threshold determining if a leaf will split further or not. A smaller value will make the model easier to get deeper and is more likely to overfit. Subsample lets the model use a randomly chosen portion of the dataset in each iteration, which is a method to prevent overfitting. Lambda, alpha and gamma are used in the regularization term to further avoid overfitting. Using early stopping, St John's and Yarmouth have 150 and 142 interactions respectively, which are the number of trees constructed in the model. The hyperparameters are applied to the structure and the loss function of all the trees. During prediction, each data point will go through all the trees, so each tree has one result. The final prediction of the model is based on all results.

Therefore, we selected ExtraTrees as a baseline due to its ensemble of decision trees, which aggregates multiple weak learners to improve predictive performance. XGBoost, a related model, was also included for comparison as it employs gradient boosting to iteratively build a strong learner. In addition, we explored neural networks for fog classification to take advantage of their strength in capturing nonlinear relationships in the data. More specifically, we implemented the biLSTM model to account for long-term temporal dependencies inherent in time series weather data. This allowed us to compare traditional tree-based ensemble methods with deep learning approaches and identify the most effective model for fog prediction.

4. Result

To summarize, the dataset included data from March to August 2012 to 2023, in which the features from ERA5 were downscaled using WPS, and the visibility reports were from the weather stations. Two locations are treated separately and have their own datasets. The dataset was split into three sets: training, validation and test sets, which were around 72%, 18%, 10% of the dataset, respectively. We ensured with the training and validation sets that both models were generalizing well and not overfitting. Since we split the data considering the nature of the time component, the test set was approximately the year 2023. This means that we tested our trained models with the 2023 dataset for both locations and saw how they performed.

We used the recall and precision metrics to evaluate the test dataset. The recall looks at the proportion of true positives (TP) that are correctly classified by the model:

$$Recall = \frac{TP}{TP + FN}$$

For fog occurrence, recall is how many fog classes the model got correctly out of the total number of fog cases. False negative (FN) , in this case, is the amount of fog our model falsely labeled as “clear”. False positive (FP) is the amount of clear incorrectly labeled as “fog” by our model and true negative (TN) is the amount of clear we got correctly. The next important metric is precision given by:

$$Precision = \frac{TP}{TP + FP}$$

Precision is the proportion of actual fog that our model got accurately from all the fog predictions it did. Using both metrics helps us to understand whether our model is just predicting a lot of fog and bumping up the performance or if it can differentiate between clear and fog well. We also used the F1-score, which is the harmonic mean of precision and recall, and offers a more balanced evaluation with one score. It ensures that both recall and precision are considered together, making it particularly useful when dealing with imbalanced datasets, where optimizing for just one metric can lead to misleading conclusions. The equation is given by:

$$F1\ score = 2\frac{precision \cdot recall}{precision + recall} = \frac{2TP}{2TP + FN + FP}$$

After training both models with the dataset collected at each location, we verified that it was generalizing well with the validation set. The next step was to see how our model did with the test data. Table 4 shows all the results obtained from the three models, ETClassifier, XGBoost and biLSTM, using a classification report showing the precision, recall, and F1-score for both clear and fog classes for St John's and Yarmouth for the 2023 dataset.

For St John's, all three models showed comparable performance in predicting fog events, with F1-scores ranging from 0.71 to 0.76. XGBoost achieved the strongest results, with a precision of 0.76, recall of 0.77, and an F1-score of 0.76, reflecting a higher ability to correctly detect fog while limiting false positives. The biLSTM also performed well, with precision of 0.72 and recall of 0.74, resulting in an F1-score of 0.73 across 611 fog cases. The ExtraTrees model followed closely, with precision of 0.70, recall of 0.72, and an F1-score of 0.71. Overall, the ensemble tree models (ExtraTrees and XGBoost) demonstrated good performance with XGBoost having a slight advantage over the biLSTM for fog prediction at this site. The similarity between the models also indicates that there was minimal overfitting.

For Yarmouth, model performance was generally lower compared to St John's, with F1-scores ranging from 0.60 to 0.65. The ETClassifier performed best, achieving a precision of 0.63, recall of 0.68, and an F1-score of 0.65, showing stronger sensitivity to fog cases while maintaining reasonable precision. XGBoost produced balanced but lower values, with both precision and recall at 0.60, yielding an F1-score of 0.60. The biLSTM model performed similarly, with a

precision of 0.61, recall of 0.64, and an F1-score of 0.63. Overall, while the ETClassifier had a slight edge, none of the models achieved the same level of skill at Yarmouth as they did at St John's, suggesting greater challenges in predicting fog at this site. The results obtained also suggest that tree models and neural network models performed at a comparable level, with the tree models having a slight advantage. Now that we know that our ML models can predict fog, the next step was to see how they perform for forecasting.

There are many challenges in ML forecasting. Forecasting needs a lot of time series data availability that may contain missing values, have irregular sampling, or have abrupt changes due to external factors. Additionally, preparing the data is more complex due to more advanced feature engineering such as creating lag variables, rolling averages, or trend components. Therefore, instead of using the ML model for a time series forecast, we decided to approach the issue differently. The features we used for our ML models are the main meteorological variables and WRF does a good job of forecasting them, especially in summer with the largest error for wind components (Bughici et al., 2019 and Pappa et al., 2023). We decided to use the forecast values of the features from WRF and then use our ML models to predict fog occurrences from them. We used WRF v4.5.2 which ran pseudo-operationally 36 hours long for every day in the study period, April to August in 2024. Meteorological variables in the last 24 hours of each run were extracted, with the first 12 hours used as a spin-up time for WRF. This allowed the model to stabilize and adjust from the initial conditions (Liu et al., 2023). To simulate a forecast environment, the forecast data from the Global Forecast System (GFS) at a 0.25-degree resolution (National Centers for Environmental Prediction, 2015) were used to provide initial and boundary conditions to the model every three hours. Since the ML model was trained on the ERA5 data and the WRF model was initialized with the GFS forecast data, the differences between the two datasets were assessed. Table 5 shows the correlations and root mean squared errors (RMSE) of some variables of the two datasets, at 12Z every day from April to August 2024 when the model is initialized. The two datasets have high correlations of T2, U10 V10 and P_sfc, with RH2 being lower. Considering that GFS is only used to initialize the WRF model and there is a 12-hour spin-up time, the output of WRF is mostly based on the physics of the model itself, instead of the provided dataset. It may be preferable to use ECMWF real-time forecasts instead of GFS if the difference is of concern. However, we believe the different datasets used between training and forecast will not have a large impact on the results.

There are 51 vertical levels in the model with more levels closer to the surface. The Thompson aerosol-aware microphysics scheme (Thompson et al., 2008 and Thompson and Eidhammer, 2014) and the Mellor–Yamada Nakanishi Niino (MYNN) Level 2.5 planetary boundary layer (PBL) scheme (Nakanishi and Niino, 2006 and 2009) are chosen. The aerosol-aware option improves the representation of droplet activation and growth, a key factor in fog formation and visibility reduction. Meanwhile, the MYNN Level 2.5 PBL scheme better captures turbulence and vertical mixing in stable boundary layers, conditions under which fog often develops. These forecast variables were then processed using the same data preparation pipeline and fed into our trained ML models to evaluate their performance on new, unseen data. This approach allowed us to assess whether the models could be used operationally to forecast fog on an hourly basis for the next 24 hours. The physical parameterization setting of the model is summarized in Table 6.

The next crucial step is to compare performance with other forecasting methods, specifically the WRF model, which is commonly used for operational marine fog prediction based on surface liquid water content. For this comparison, the cloud water mixing ratio at the lowest model level at approximately 3.6 m was also extracted from the WRF runs mentioned above. To classify fog in a binary manner, we labeled any instance where the cloud mixing ratio was nonzero (indicating the presence of liquid water) as fog, and instances where it was zero as clear. This approach aligns with WRF's tendency to predict high liquid water content which always drops the visibility level in the fog zone. Table 7 shows the results for fog classification on the 2024 dataset for the post-processing ML models and the WRF model.

Table 8 shows the confusion matrices for all the models at both locations in 2024 to have a better understanding of how many fog labels we were able to capture. The confusion matrix for each location will indicate the amount of clear we got correctly (TN) and the amount we mislabeled as fog (FP). It will also indicate the amount of fog we got correctly (TP) and the amount we mislabeled as clear (FN) and will be given in this format, $\begin{pmatrix} TN & FP \\ FN & TP \end{pmatrix}$. The first row is the total number of clear cases (TN + FP), and the second row is the total number of fog cases (FN + TP).

At St John's, all models showed a slight decrease in performance compared to 2023, but the relative ranking of models changed with ETClassifier slightly performing better. The ETClassifier

achieved the highest F1-score (0.69) with balanced precision (0.68) and recall (0.70) across 565 fog cases. Both XGBoost and the biLSTM model performed comparably, each reaching an F1-score of 0.68, though XGBoost demonstrated stronger recall (0.74) at the cost of lower precision (0.64). The baseline WRF liquid water content approach performed the worst with an F1-score of 0.62. All three ML models, as a result, captured more fog events compared to WRF with XGBoost capturing the most fog events but overestimated the fog events due to lower precision. ETClassifier offered the most balanced tradeoffs, outperforming WRF on all aspects such as less false positives (191 vs. 229) and less false negatives (160 vs. 209).

Model performance was lower overall at Yarmouth compared to St John's. The ETClassifier again provided the best balance, achieving an F1-score of 0.61, with moderate precision (0.59) and recall (0.63). The biLSTM model performed similarly with an F1-score of 0.58, while XGBoost showed high recall (0.71) but low precision (0.47), leading to an F1-score of 0.56. The WRF LWC method had the weakest results with an F1-score of 0.51. Similarly to St John's, XGBoost captured the most fog events with 396 but also predicted twice the amount of fog events that happened in 2024 which aligns with the low precision. ETClassifier showed the most balanced tradeoffs while outperforming WRF in all aspects, and more significantly it was able to distinguish the clear events from fog (less false positives, 242 vs. 428). These findings confirm that post-processing with machine learning methods brings improvements in fog detection skill over WRF. However, while the performance gains at Yarmouth are bigger, they are more variable across model types compared to St John's. Additionally, the overall performance decreased more at Yarmouth from the 2023 dataset evaluation (e.g. ETClassifier went from 0.65 to 0.61).

In Fig. 7, we plotted the feature importance for both locations for the 2024 dataset to better understand which systematic biases in the WRF fields were being corrected by the machine learning models. At St John's, the ETClassifier identified relative humidity (RH2) and its lagged value as the most important predictors of fog, followed by the zonal wind (U and U_lag_1). This finding is consistent with the observed link between easterly winds and fog formation in this region, highlighting the role of moisture availability and wind-driven advection in controlling visibility. In contrast, at Yarmouth, RH2 and RH2_lag_1 overwhelmingly dominated the feature importance, indicating that near-surface moisture and its persistence were the primary drivers of fog events. Other predictors such as wind components, time of day, and temperature contributed modestly, suggesting that while circulation patterns influence fog, the local

predictability at Yarmouth relies more heavily on sustained humidity conditions than on wind forcing. This heavy reliance on RH2 could also explain why it is harder to predict fog using the ML models and the F1-score was lower than St John's. The patterns at both locations also align with the heatmap analysis.

To statistically compare the classification performance of the ExtraTrees classifier (ETClassifier) and the WRF-based liquid water content (LWC) predictions, we applied McNemar's test. This non-parametric test evaluates whether two classifiers differ significantly in terms of their error patterns, by focusing on the cases where the models disagree. A continuity correction was applied, and significance was assessed at the 0.05 level. For St John's, we obtained a χ^2 of 18.4 and p-value of 1.77e-05 and for Yarmouth, we got χ^2 to be 49.7 with a p-value of 1.84e-12. McNemar's test indicated a statistically significant difference between the ETClassifier and the WRF-based predictions as both p-values are significantly smaller than 0.05. This suggests that the ETClassifier made significantly fewer misclassifications than the WRF model, confirming that the observed performance improvements are unlikely to be due to chance.

Our post-processing approach not only improves hourly fog prediction performance over a 24-hour horizon but also compares favorably with other ML-based classification studies, despite the differences in datasets and lead times. This suggests that tree-based ensemble methods like ETClassifier provide a reliable and efficient solution for operational fog forecasting when used alongside traditional numerical models. To better understand the performance of our 24-hour fog forecast at an hourly resolution, we evaluated model skill over time by dividing the forecast horizon into four 6-hour bins. In many machine learning approaches, model performance tends to degrade as the forecast lead time increases due to accumulated uncertainty. However, in our case, we believe that the use of post-processing techniques would help stabilize accuracy across the forecast period. While we acknowledge that the underlying WRF model can introduce increasing error over time, we aimed to determine whether our post-processing could mitigate this effect.

Fig. 8 and Fig. 9 present the F1-scores for fog detection in 6-hour bins for both St John's and Yarmouth, based on forecasts from the 2024 dataset. The number of fog events occurring is shown for each UTC 6h bin. For both locations, the first 12 hours had the highest fog counts compared to the last 12 hours. Since the times are in UTC, the first 12 hours represent late night to early mornings (21:30 to 09:30 for St John's and 21:00 to 09:00 for Yarmouth) which

aligns with radiation fog events. At St John's, the F1-scores remain very stable across all bins, with the 06-11hr bin being slightly lower. In Yarmouth, performance is much lower at the 12-17hr mark but shows minimal variation across the other time bins, with no clear trend of degradation. This decrease in performance at the 12-17hr mark can be due to it representing daytime where fog occurrence can be harder to predict and coupled with a lower fog count compared to the first 12 hours (87 vs. 233).

These results support our assumption that the post-processing approach helps maintain consistent fog classification skills across the 24-hour forecast window. Nevertheless, we note that the WRF model's intrinsic error growth with lead time was not explicitly assessed here, and future work should investigate the upper limits of this post-processing approach for longer forecast horizons.

We found that the results were relatively consistent across the 24-hour forecast horizon and that the ETClassifier significantly outperformed the WRF baseline according to McNemar's test. However, because the 2024 dataset contained a relatively small number of fog events (565 at St John's and 558 at Yarmouth), it was important to assess the uncertainty in the reported metrics. To address this, we applied a bootstrap resampling approach to compute 95% confidence intervals (CIs) for precision, recall and F1-score. For each metric, we generated 10,000 bootstrap samples by resampling with replacement from the 2024 test dataset. The metric was recalculated for each resample, and the 2.5th and 97.5th percentiles of the resulting distribution were taken as the 95% confidence interval. The results for each location are given in Table 9.

Reporting bootstrap confidence intervals provides an assessment of the reliability of the observed improvements. Unlike single-point estimates, CIs show the plausible range of metric values that could be expected if the experiment were repeated. From Table 9, we see that the relatively narrow confidence intervals suggest stable performance estimates, while still accounting for sampling variability due to the limited number of fog events. All 3 metrics for both locations fall into the range of the 95% confidence interval (as shown in Table 7). We observe that even at the lower bound of the 95% CI, the ETClassifier maintains better performance than the WRF baseline, strengthening the conclusion that the improvement is robust.

While the statistical tests provide confidence in the overall results, they do not illustrate how the models behave during specific forecast periods. To provide further insight into the strengths and limitations of ML post-processing, we present a short case study for both locations. We have a 5-day window for both locations. For St John's, we have from May 10th to 15th and for Yarmouth, it is from July 5th to 10th. These periods were chosen because they contain a mix of foggy and clear conditions, allowing for a clearer assessment of how the WRF model and the ETClassifier perform. The corresponding time series are shown in Fig. 10 and Fig. 11.

For St John's (Fig. 10), the case study highlights both the challenges of forecasting rapidly evolving fog events and the advantages of applying the ML post-processing. Over the 5-day window from May 10 to 15, several short fog events and quick transitions between fog and clear conditions were observed. The raw WRF output (blue dashed line) generally captured the broader fog periods but tended to overestimate their persistence, missing several short-lived clear intervals, most notably during May 11-13. The ETClassifier (red dashed line) offered a slightly improved representation overall, with better agreement during the prolonged fog episode near the end of the case study (May 14-15). However, it still struggled to resolve short-lived fog fluctuations, capturing only part of the frequent transitions evident in the observations (black line). Thus, while the ML approach reduced some false persistence compared to WRF, both systems exhibited limitations in reproducing rapid temporal variability.

For Yarmouth (Fig. 11), the ETClassifier demonstrated a clearer advantage in capturing both the timing and duration of fog events compared to WRF. Even if WRF captured the onset of the first fog formation on July 7 more accurately, the ML model was able to predict the duration better. The difference became more pronounced during the July 8 event, where WRF missed the fog onset by a few hours and failed to sustain the fog through its observed duration. The ETClassifier, on the other hand, closely followed the observed fog evolution, even detecting a secondary fog episode following a brief clearing period late on July 8. Toward the end of the period, the ETClassifier continued to respond effectively to rapid transitions that WRF largely failed to capture. While not perfect, the ML approach reduced false alarms compared to WRF alone and demonstrated its potential to better resolve the variability in local fog conditions.

Beyond the comparison with the baseline WRF forecasts, it is also important to evaluate how the performance of our approach relates to existing studies in fog prediction. While direct quantitative comparisons with prior studies are constrained by differences in data sources and

forecast settings, the relative performance of our approach aligns well with or in many cases exceeds the results reported in similar classification-based ML studies, despite those often focusing on shorter lead times. Castillo-Botn et al. (2022) conducted a comprehensive comparison of supervised ML approaches for fog prediction at the Mondonedo weather station in Galicia, Spain. Their work focused on a very short-term nowcasting task, with a lead time of 30 minutes, and evaluated both classification and regression methods. For classification, they assessed model performance using accuracy and the weighted F1-score, with Gradient Boosting yielding the best results, obtaining an accuracy of 0.826 and a weighted F1-score of 0.828. The weighted F1-score is the combined F1-scores for clear and fog and for St John's, ETClassifier obtained a 0.91 for both metrics and 0.87 for Yarmouth. A higher performance was achieved for both locations with a much longer forecast horizon.

Miao et al. (2020) used LSTM for short-term fog forecasting using weather stations in Anhui, China and provided the F1-score for hourly forecasts up to 4 hours. They were able to obtain an F1-score of 0.758 in the first hour, 0.66 in the next hour, 0.53 in hour 3 and 0.474 in the final hour. These scores reflect a declining trend over time but in our case, we had a stable F1-score for the 24-hour prediction at an hourly resolution. For St John's, we obtained an F1-score of 0.70 and for Yarmouth we got 0.61. While Miao et al. (2020) achieved higher performance in the very short term (first hour for St John's and first two hours for Yarmouth), our method maintained a more stable F1-score across the entire 24-hour period, offering a longer and more operationally useful forecast horizon.

Kamangir et al. (2021) proposed FogNet which gave the forecast of fog over the coast of Texas. They also used a post-processing approach with NWP and sea surface temperature satellite observations to give fog forecast at the Mustang Beach Airport in Port Aransas and chose three different lead times (6 hours, 12 hours and 24 hours). For the evaluation of their approach, they used different metrics such as the Probability of Detection (POD) which measures the likelihood that the ML model correctly classifies the intended target and False Alarm Ratio (FAR) which is

the fraction of forecasted non-occurrence of the event over the occurrence of the event. For POD, a value closer to 1 is better and for the FAR, a value closer to 0 is better. We calculated both metrics for each location on the 2024 dataset using the ETClassifier score. While they categorized visibility different from us, with fog events happening when visibility drops to 1.6 km, we did a relative comparison. For the 6-hour forecasting window, they obtained a POD of 0.61 and a FAR of 0.40. For St John's, we had a POD of 0.71 and a FAR of 0.32 for the 24-hourly forecast and for Yarmouth, we had a POD of 0.63 and a FAR of 0.41. We were able to have a higher time resolution (1-hour compared to 6-hour) for the 24-hour prediction and for both locations we obtained a higher POD. We also obtained a higher FAR for St John's with Yarmouth slightly trailing behind by 0.01. However, the results were much better when compared to the 12 and 24 windows since our results maintained a relatively stable performance for the 24-hour forecast.

In summary, while direct comparisons are limited by methodological differences, our results show that the proposed ETClassifier approach achieves performance comparable to or exceeding that of previous ML-based fog prediction studies, particularly in maintaining stable forecast accuracy over a 24-hour horizon.

5. Discussion

While a slight decrease in model performance is expected when evaluating on new, unseen data, our results show that the decline was very small at both locations. For St John's, the F1-score decreased from 0.71 on the 2023 dataset to 0.70 on the 2024 dataset using the ETClassifier model. At Yarmouth, the F1-score dropped slightly more compared to St John's, but within range of new unseen data, from 0.65 in 2023 to 0.61 in 2024 for ETClassifier. These small differences indicate that the ETClassifier generalizes well over the years, maintaining consistent improvements over the baseline WRF model.

We were able to improve on the WRF LWC model at both locations, but we did get a lower F1-score at Yarmouth compared to St John's, even with training them separated. We observed that

WRF followed the same trend which suggests that the input data fed into the machine learning model may have already carried inherent limitations. If the forecasted features, such as temperature, relative humidity, or wind components are less accurate in Yarmouth due to model bias, resolution, or local meteorological complexities, this could significantly hinder the model's ability to generalize. Therefore, the observed drop in ML model performance may not reflect a shortcoming of the classification model itself but rather a degradation in input feature quality, which propagates through the prediction pipeline.

Furthermore, to make sure we got the best possible trained ML model, we had to ensure that we prepared the data correctly, optimized the model carefully, and tested it on different types of data. Besides, we implemented specific strategies to refine our approach, ensuring that our model was well-suited for the task and capable of delivering the most accurate fog predictions. We evaluated which input data to use for our ML models between ERA5 reanalysis and GFS Final Analysis data. We used ETClassifier since it was the better performing model. We trained and did a comparison on the 2023 dataset. Since we were only able to collect data from 2016 to 2023 from GFS Final Analysis, compared to 2012 for ERA5, the performance when using features from the ERA5 data was better on all metrics. We also evaluated the importance of our lagged features, which showed slightly better or similar performance in terms of precision and F1-score but had higher recall at each location. Although the overall improvement was not drastic, incorporating a one-hour lag to the features helped the model to distinguish more accurately between fog and clear cases, which is crucial for operational forecasting.

One can argue that by having more data to train, the ML model may learn more about the patterns to classify fog or will generalize better. Therefore, we tested the case that combined both locations together and trained a single ETClassifier ML model. The data were prepared in the same way as before: splitting the data while maintaining the time series, normalizing the features, using lagged features. After concatenating the datasets from both locations, we used OneHotEncoder from Scikit-Learn to transform the names of the locations into a binary format without any ordinality. After training the model, we tested it with the 2023 dataset for both St John’s and Yarmouth. When compared to when we trained the locations individually, we obtained an overall lower F1-score at each location.Therefore, we were able to confidently say that choosing a model for each location was the correct approach to get the best results as the model was able to learn the weather patterns of that location better, thus learning and generalizing better.

We have also tested the case of multi-classification. Between clear and fog (≤ 1 km), we added one more label of mist with visibility from 1 km to 5 km. A key advantage of multi-class classification is that if the model predicts mist instead of fog, it still serves as an early warning, bringing us closer to identifying fog conditions. Additionally, having mist as a separate category provides more nuanced predictions and better reflects real-world conditions. We prepared the data exactly as before and used RandomSearchCV to get the best hyperparameters for this task. However, since mist was severely under-represented in our dataset, our model could not predict “mist” classes at all and performed worse.

Regression is another supervised ML technique that models the relationship between a dependent variable (target) and one or more independent variables (features), similar to classification, but the model predicts continuous values instead of discrete labels (Sagar 2025). It is possible for forecasting visibility since we have visibility reports that range from 0 km to around 24 km. There are various types of regression, with the simplest being linear regression (Altman and Krzywinski, 2015), which assumes a linear relationship between the input variables and the target. We have tried two models, the ExtraTrees Regressor (ETRegressor) and Quantile Regression with Random Forests. We initially used ETRegressor as the ExtraTrees model was the better performing model for classification but had to use a more robust model (Quantile Regression) to get better results.

To evaluate the regression approach, we employed several standard metrics: Mean Absolute Error (MAE), which quantifies the average magnitude of errors between predicted and observed values; Mean Squared Error (MSE), which penalizes larger errors by averaging the squared differences; and the coefficient of determination (R^2), which measures how well the predicted values align with the actual data. After training the model, we evaluated its performance on the 2023 test dataset for St John’s. While the model was occasionally able to capture low-visibility events (below 1 km), it frequently failed to detect sudden drops in visibility associated with fog formation or rapid increases during fog dissipation. These missed transitions highlighted the limitations of regression in identifying abrupt changes, which are critical for fog forecasting.

Although the overall regression performance was consistent with previous studies on marine fog visibility, it became evident that binary classification was more effective at capturing the onset and dissipation of fog events.

6. Conclusion

This study explored the challenges of marine fog prediction by integrating physical and data-driven approaches. Initially, we used the WRF model to simulate marine fog for the Fog And Turbulence Interactions In the Marine Atmosphere (FATIMA) project. Recognizing the limitations of physical models in fog forecasting, we implemented a machine learning approach using binary classification to predict fog occurrence in coastal regions of Canada (St John's and Yarmouth). We trained the models using ERA5 and labeled fog events using NAV Canada visibility reports (visibility ≤ 1 km). Among the models tested, ETClassifier consistently outperformed WRF-based predictions, achieving an 13% improvement in fog classification performance based on F1-score at St John's and 18% at Yarmouth. The binary classification approach proved most effective, aligning with previous studies such as Lam et al. (2023). Additionally, this approach enabled us to do 24-hour forecasts at hourly resolution, which is an advancement over typical 1–7-hour lead times reported in earlier works.

Improvements could be made by incorporating alternative forecast datasets. Using ECMWF's Integrated Forecast System (IFS) or processing forecasted GFS data directly through WPS may enhance 24-hour prediction quality. While ERA5 and GFS perform well for surface variables, discrepancies with station observations (Velikou et al., 2022, Haiden et al., 2021) suggest that increasing observational input for training could further improve accuracy. Standalone AI forecasting systems also present a promising avenue. ECMWF's Artificial Intelligence Forecasting System (AIFS), operational since February 2025, provides high-resolution deterministic forecasts and could be used to drive post-processed ML fog predictions. Our

models, when coupled with such advanced inputs, may yield higher accuracy and faster outputs.

Additionally, nowcasting presents a promising opportunity to enhance fog prediction. This approach involves using lagged visibility observations as input features, enabling the model to learn from recent visibility trends and make more accurate short-term forecasts. Though not the focus of this study, preliminary results using visibility as a feature for short-term forecasts (1-6 hours) showed substantial gains, with an F1-score of around 0.85 for 1-hour lead times. Expanding this approach could be beneficial for real-time applications. Finally, the broader context of climate change must be considered. Recent studies (e.g., Johnstone and Dawson, 2010, Taha and Abduljabbar, 2024) indicate a declining trend in marine fog frequency linked to rising sea surface temperatures and weakened air-sea temperature contrasts. For regions like St John's and Yarmouth, shifts in wind patterns, ocean currents (e.g., the Labrador Current), and aerosol concentrations could further impact fog occurrence. These evolving dynamics emphasize the need for adaptive, ML-enhanced forecasting systems capable of responding to changing environmental conditions.

References

Altman, N. and Krzywinski, M., 2015: Simple linear regression. *Nat Methods*, **12**, 999–1000, https://doi.org/10.1038/nmeth.3627.

Bughici, T., Lazarovitch, N., Fredj, E. and Tas, E., 2019: Evaluation and Bias Correction in WRF Model Forecasting of Precipitation and Potential Evapotranspiration. *J. Hydrometeor.*, **20**, 965–983, https://doi.org/10.1175/JHM-D-18-0160.1.

Bullock, T., Isaac, G. A., Beale, J. and Hauser T., 2016: Improvement of visibility and severe sea state forecasting on the grand banks of Newfoundland and Labrador. Paper presented at the Arctic Technology Conference, St. John's, Newfoundland and Labrador, Canada. https://doi.org/10.4043/27406-MS.

Castillo-Botón, C., Casillas-Pérez, D., Casanova-Mateo, C., Ghimire, S., Cerro-Prada, E., Gutierrez, P.A., Deo, R.C. and Salcedo-Sanz S., 2022: Machine learning regression and classification methods for fog events prediction. Atmospheric Research, **272**, https://doi.org/10.1016/j.atmosres.2022.106157.

Chen, K., Zhou, Y., Ren, T., and Li, X., 2024: Short-term sea fog area forecast: A new data set and deep learning approach. *Journal of Geophysical Research:Machine Learning and Computation*, **1**, e2024JH000230, https://doi.org/10.1029/2024JH000230.

Chen, T., and Guestrin, C., 2016. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16). Association for Computing Machinery, New York, NY, USA, 785–794, https://doi.org/10.1145/2939672.2939785

Flynn, C., 2018: The physics of fog. https://blog.metservice.com/Physics-of-Fog. [Online; accessed 25-December-2024].

Gao, X., Bao, X., Ma, S., Chen, Q. Wang B., 2024: An online assimilation method to improve the numerical forecast of sea fog using microwave radiometer-retrieved cloud water path. *JGR Atmospheres*, **129**(1), https://doi.org/10.1029/2023JD040229.

Geurts P., Ernst, D. and Wehenkel, L., 2006: Extremely randomized trees. *Machine Learning*, **63**, 3-42, https://doi.org/10.1007/s10994-006-6226-1.

Glossary of Meteorology (2025). Glossary of meteorology- fog. https://glossary.ametsoc.org/wiki/Fog

Gultepe E., Wang S., Blomquist B., Fernando H. J. S., Kreidl O. P., Delene D.J. and Gultepe I., 2024: Machine learning analysis and nowcasting of marine fog visibility using FATIMA Grand Banks campaign measurements. *Front. Earth Sci*. **11**, https://doi.org/10.3389/feart.2023.1321422.

Haiden, T., Janousek, M., Vitart, F., Ben-Bouallegue, Z., Ferranti, L. and Prates F., 2021: Evaluation of ECMWF forecasts, including the 2021 upgrade. https://doi.org/10.21957/90pgicjk4.

Hersbach H. et al., 2020: The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, **146**, 1999–2049, https://doi.org/10.1002/qj.3803.

Hsieh, W. W., 2023: Introduction to environmental data science (1st ed.). Cambridge University Press.

Hochreiter, S. and Schmidhuber, J., 1997: Long Short-Term Memory. *Neural Computation*, **9**(8), 1735-1780, https://doi.org/10.1162/neco.1997.9.8.1735.

Johnstone, J. A. and Dawson, T. E., 2010: Climatic context and ecological implications of summer fog decline in the coast redwood region. *Proceedings of the National Academy of Science of the United States of America*, **107**(10), 4533-4538, https://doi.org/10.1073/pnas.0915062107.

Kamangir, H., W. Collins, P. Tissot, S. A. King, H. T. H. Dinh, N. Durham, and J. Rizzo (2021). Fognet: A multiscale 3d cnn with double-branch dense block and attention mechanism for fog prediction. https://www.sciencedirect.com/science/article/pii/S2666827021000190. DOI: https://doi.org/10.1016/j.mlwa.2021.100038.

Koračin, D., Dorman, C. E., Lewis, J. M., Hudson, J. G., Wilcox, E. M., and Torregrosa A., 2014: Marine fog: A review. *Atmospheric Research*, **143**, 142-175, https://doi.org/10.1016/j.atmosres.2013.12.012.

Lam, R. et al., 2023: Learning skillful medium-range global weather forecasting. *Science,* **382**,1416-1421, https://doi.org/10.1126/science.adi2336.

Liu, Y., Zhuo, L. and Han D., 2023: Developing spin-up time framework for WRF extreme precipitation simulations. *Journal of Hydrology*, **620**, https://doi.org/10.1016/j.jhydrol.2023.129443.

Miao, K.-c., T.-t. Han, Y.-q. Yao, H. Lu, P. Chen, B. Wang, and J. Zhang (2020). Application of lstm for short term fog forecasting based on meteorological elements. Neurocomputing 408, 285–291.

Nakanishi, M., and Niino, H., 2006: An Improved Mellor–Yamada Level-3 Model: Its Numerical Stability and Application to a Regional Prediction of Advection Fog. *Boundary-Layer Meteorology,* **119**, 397-407, https://doi.org/10.1007/s10546-005-9030-8

Nakanishi, M., and Niino, H., 2009: Development of an Improved Turbulence Closure Model for the Atmospheric Boundary Layer. *Journal of the Meteorological Society of Japan*, **87**, 895-912. https://doi.org/10.2151/jmsj.87.895.

National Center for Atmospheric Research, 2022: Fog forecasting to avoid delays and accidents. https://ral.ucar.edu/pressroom/features/fog-forecasting-to-avoid-delays-accidents.[Online; accessed 25-December-2024].

National Centers for Environmental Prediction, National Weather Service, NOAA, U.S. Department of Commerce. 2015, updated daily. NCEP GDAS/FNL 0.25 Degree Global Tropospheric Analyses and Forecast Grids. NSF National Center for Atmospheric Research. https://doi.org/10.5065/D65Q4T4Z. Accessed† 30 08 2024.

NavCanada, 2017: Aviation weather services guide for aviation users. https://www.navcanada.ca/en/aviation-weather-services-guide.pdf.

Pappa, A., Siouti, E., Pandis, S. N. and Kioutsioukis I., 2023: High-resolution WRF forecasts in the SmartAQ system: Evaluation of the meteorological forcing used for PMCAMx predictions in an urban area. *Atmospheric Research*, **296**, https://doi.org/10.1016/j.atmosres.2023.107041.

Pedregosa, F. et al., Scikit-learn: Machine Learning in Python, JMLR 12, pp. 2825-2830, 2011.

Peláez-Rodríguez, C., Pérez-Aracil, J., Casanova-Mateo, C. and Salcedo-Sanz S., 2023: Efficient prediction of fog-related low-visibility events with Machine Learning and evolutionary algorithms, *Atmospheric Research,* **295**, https://doi.org/10.1016/j.atmosres.2023.106991.

Ratner, B., 2009: The correlation coefficient: Its values range between +1/−1, or do they?, *Journal of Targeting, Measurement and Analysis for Marketing*, **17**, 139-142, https://doi.org/10.1057/jt.2009.5.

Sagar, S., 2025: Regression in machine learning. https://www.geeksforgeeks.org/regression-in-machine-learning/. [Online; accessed 19-February-2025]

Taha, N. W. and Abduljabbar H. S., 2024: The relationship between climate change and fog phenomenon in southern Iraq, E3S Web of Conferences, **583**, 02014, https://doi.org/10.1051/e3sconf/202458302014.

Taylor, P. A., Chen, Z., Cheng, L., Afsharian, S., Weng, W., Isaac, G. A., Bullock, T. W. and Chen Y., 2021: Surface deposition of marine fog and its treatment in the Weather Research and Forecasting (WRF) model. *Atmospheric Chemistry and Physics*, **21**(19), 14687–14702, https://doi.org/10.5194/acp-21-14687-2021.

Thompson, G., Field, P. R., Rasmussen, R. M. and Hall, W. D., 2008: Explicit Forecasts of Winter Precipitation Using an Improved Bulk Microphysics Scheme. Part II: Implementation of a New Snow Parameterization. *Mon. Wea. Rev.*, **136**, 5095–5115, https://doi.org/10.1175/2008MWR2387.1.

Thompson, G. and Eidhammer, T., 2014: A study of aerosol impacts on clouds and precipitation development in a large winter cyclone. *J. Atmos. Sci.*, **71**(10), 3636-3658. https://doi.org/10.1175/JAS-D-13-0305.1.

Velikou, K., Lazoglou, G., Tolika, K. and Anagnostopoulou C., 2022: Reliability of the ERA5 in Replicating Mean and Extreme Temperatures across Europe, *Water*, **14**(4), 543, https://doi.org/10.3390/w14040543.

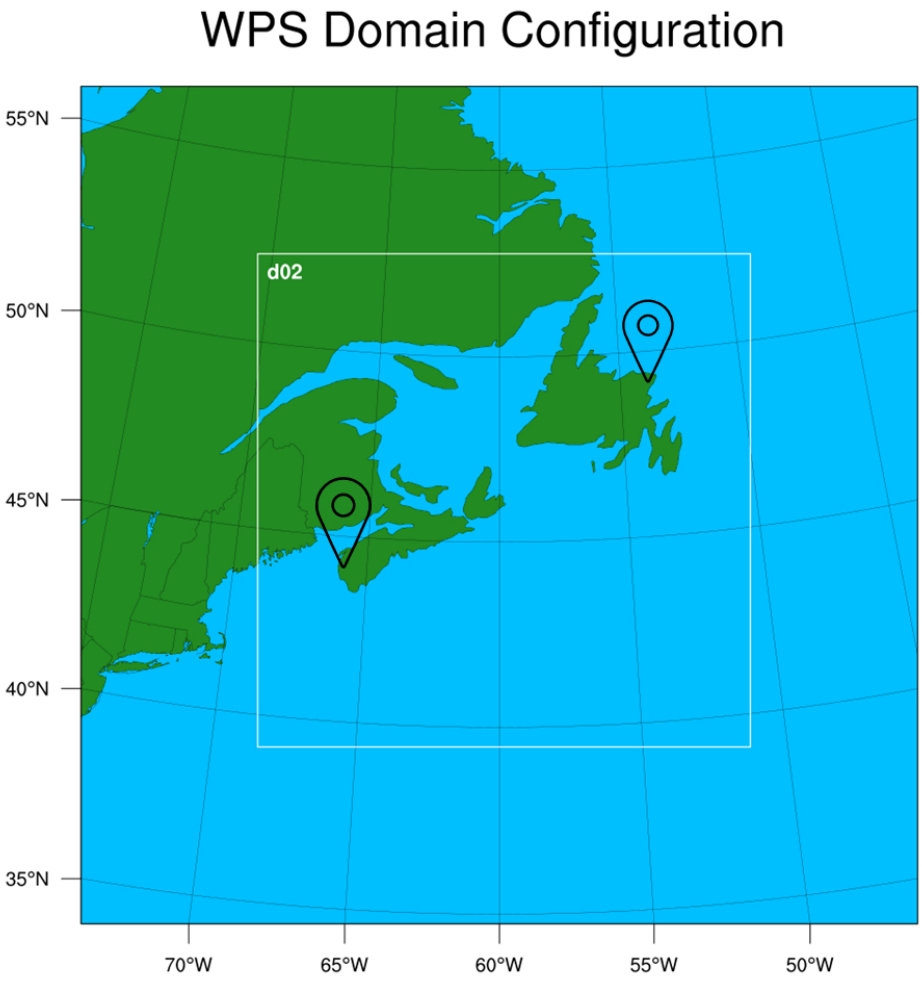

Figure 1: Two-nested WRF domain configuration with 27 km and 9 km resolutions with the locations of the weather stations pinned. The pin on the right-hand side in d02 represents St John's and that on the left-hand side represents Yarmouth.

414x414mm (59 x 59 DPI)

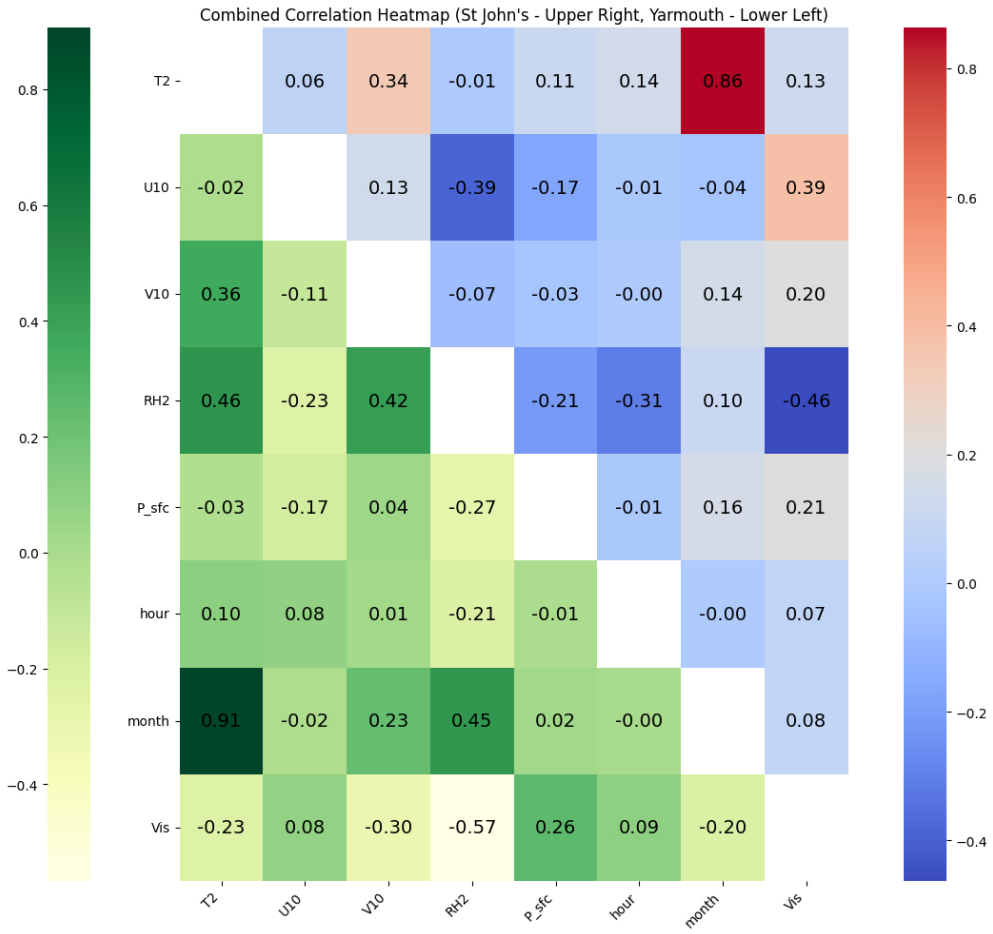

Figure 2: Correlation heatmap of the features relevant for fog prediction using the ERA5 data from 2012 to 2023 for St John's, Newfoundland and Labrador on the upper right in cool warm colors and Yarmouth, Nova Scotia on the lower left in green colors. Values in the diagonal are hidden because the correlation of a variable with itself is always equal to 1.

683x644mm (39 x 39 DPI)

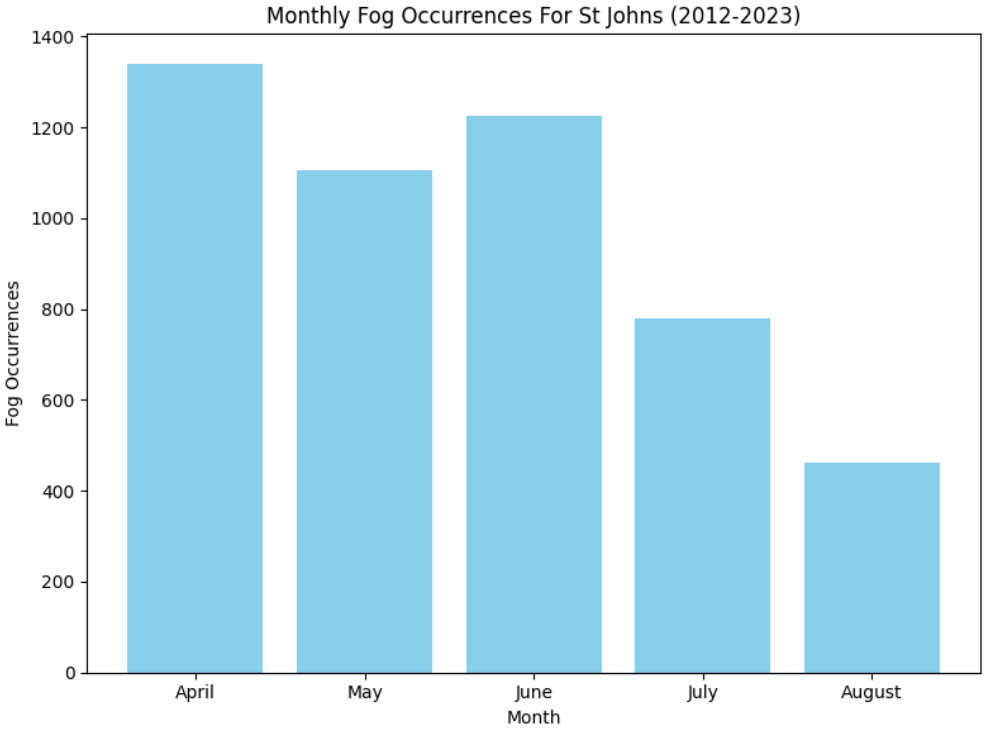

Figure 3: Hourly fog occurrence classified monthly for St John's, with the peak month being April.

278x208mm (72 x 72 DPI)

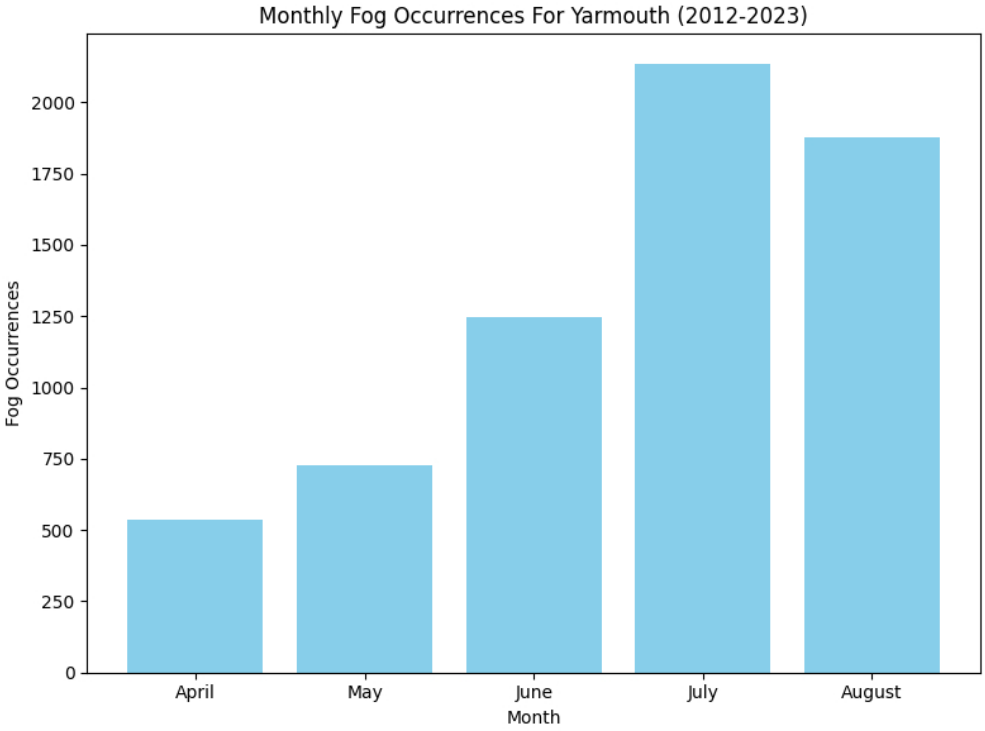

Figure 4: Hourly fog occurrence classified monthly for Yarmouth, with the peak month being July.

278x208mm (72 x 72 DPI)

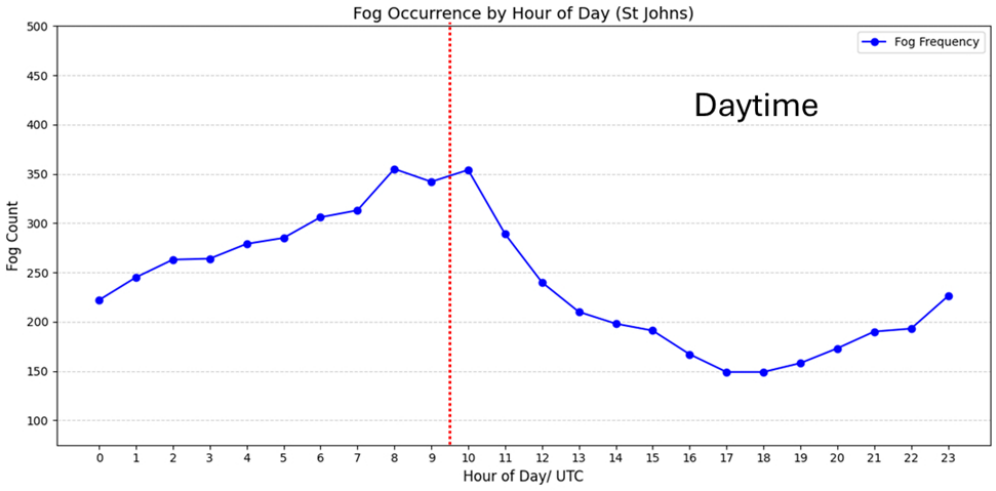

Figure 5: Line plots of hourly fog occurrences in a day during March to August, from 2012 to 2023 for St John's. Since the time is in UTC, this indicates higher fog during nighttime and early morning.

419x207mm (59 x 59 DPI)

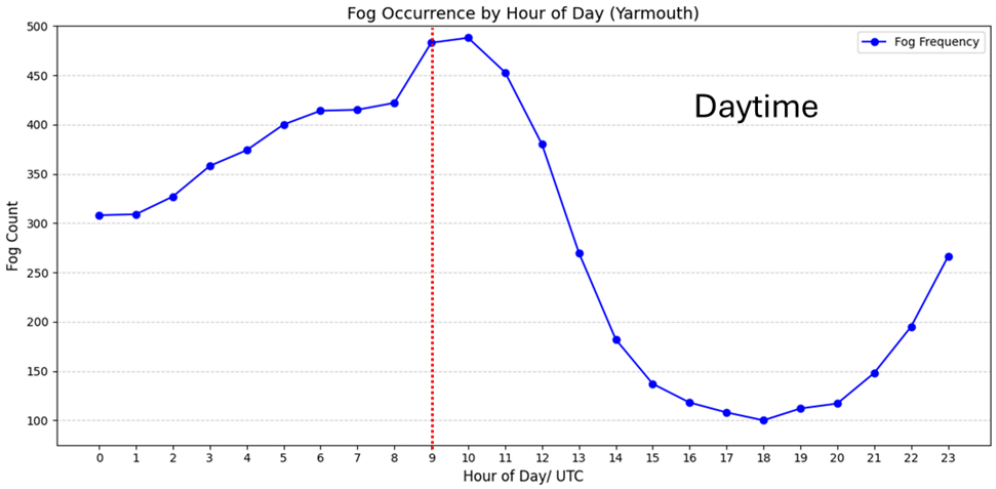

Figure 6: Line plots of hourly fog occurrences in a day during March to August, from 2012 to 2023 for Yarmouth. Similar to St John's, this indicates higher fog during nighttime and early morning.

419x207mm (59 x 59 DPI)

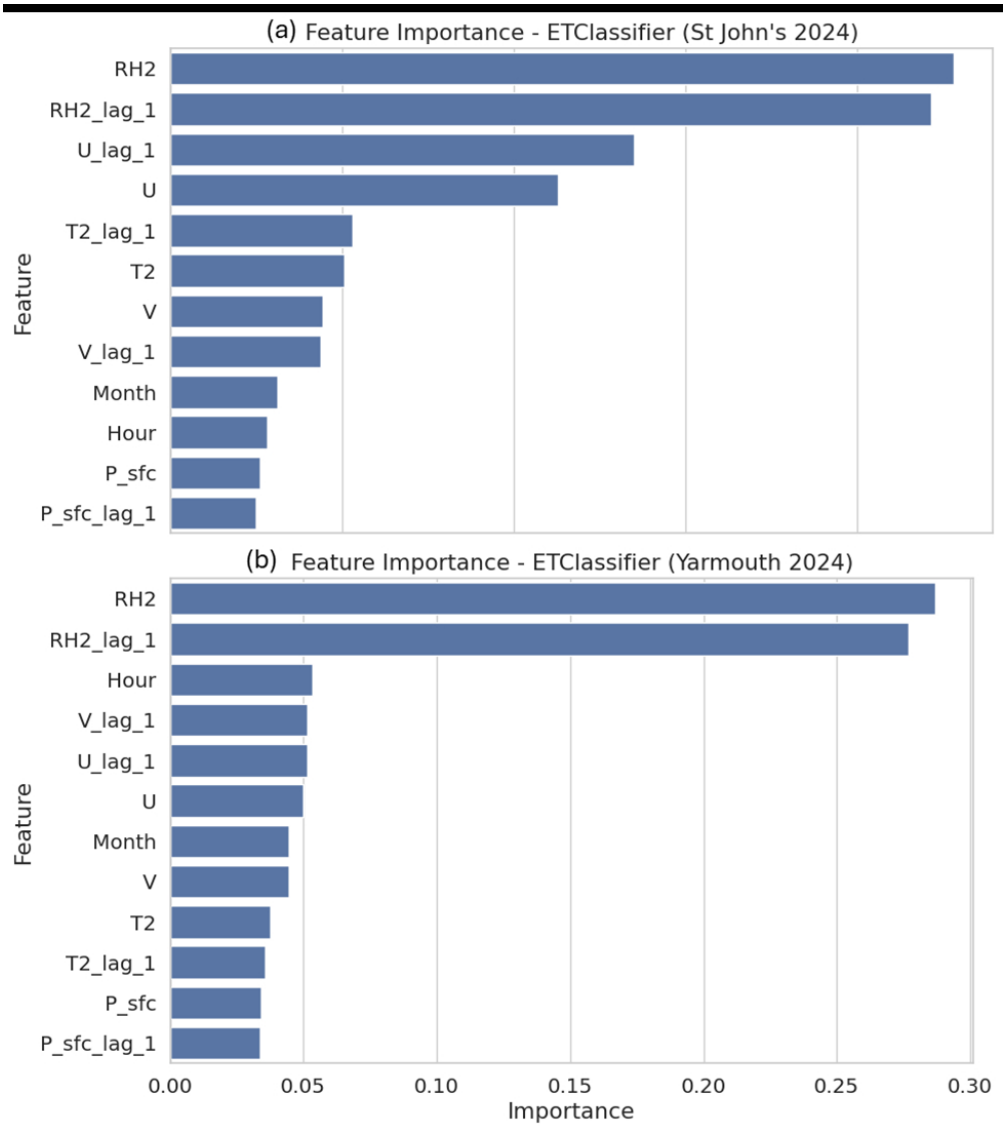

Figure 7: Relative importance of the WRF-derived predictors in the ETClassifier model for (a) St John's and (b) Yarmouth for the 24hr forecast during Summer 2024.

419x473mm (59 x 59 DPI)

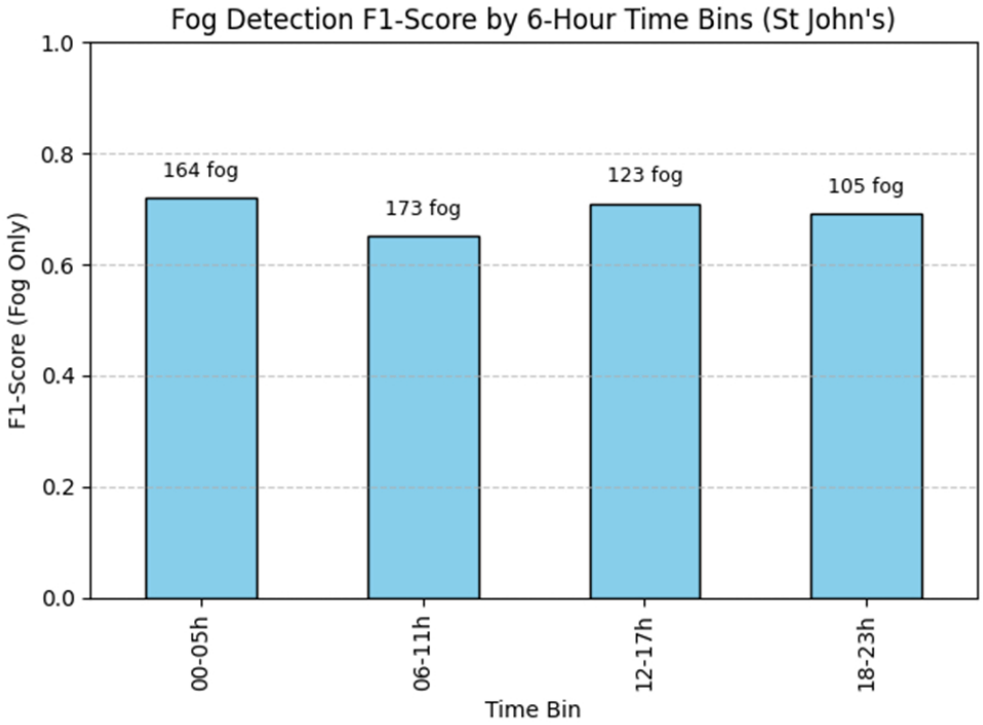

Figure 8: Hourly fog classification performance evaluated using F1-score in 6-hour time bins for St John's, based on 24-hour forecasts for the summer 2024 dataset.

419x312mm (59 x 59 DPI)

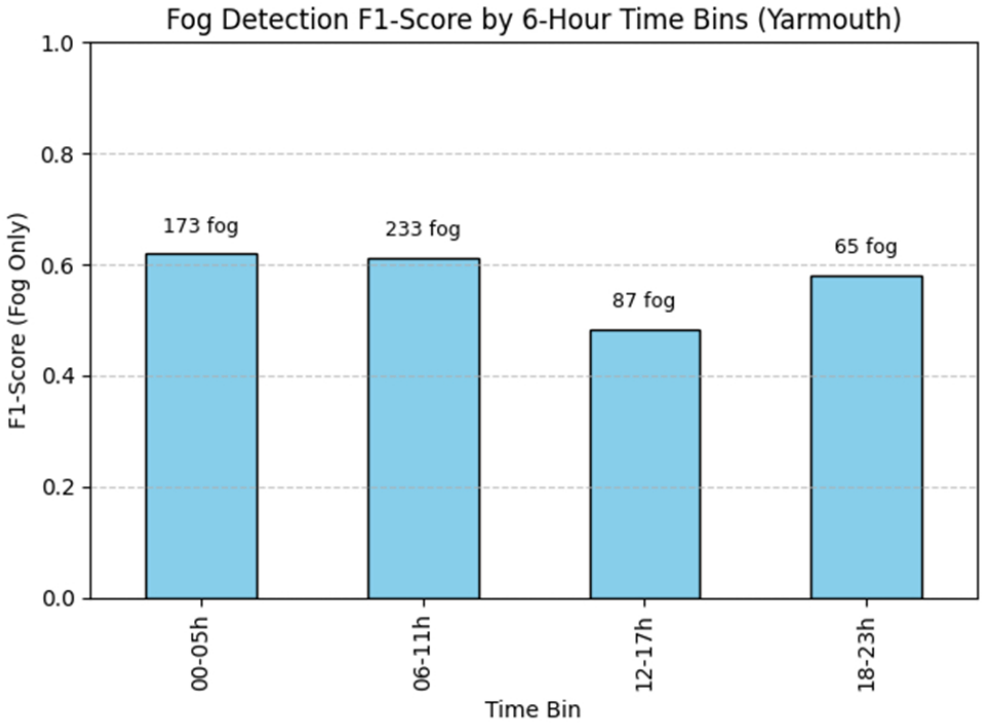

Figure 9: Hourly fog classification performance evaluated using F1-score in 6-hour time bins for Yarmouth, based on 24-hour forecasts for the summer 2024 dataset.

419x312mm (59 x 59 DPI)

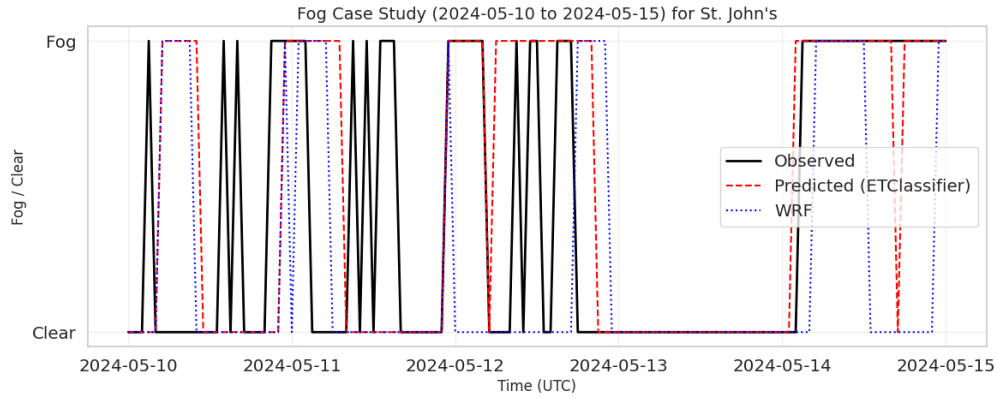

Figure 10: Fog Case Study from May 10th to 15th, 2024 for St John's comparing ETClassifier and WRF.

760x308mm (39 x 39 DPI)

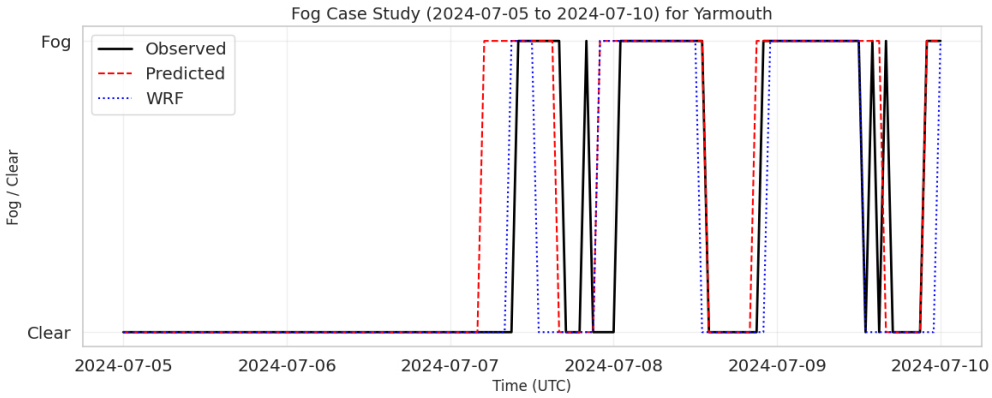

Figure 11: Fog Case Study from July 5th to 10th, 2024 for Yarmouth comparing ETClassifier and WRF.

760x308mm (39 x 39 DPI)

Variable	Abbreviation	Units
Temperature at 2m	T2	K
Zonal wind component at 10m	U10	m s^{-1}
Meridional wind component at 10m	V10	m s^{-1}
Relative Humidity at 2m	RH2	%
Surface Pressure	P_sfc	Pa

Table 1: Predictive variables collected hourly using the ERA5 dataset for each location used to predict fog occurrence.

414x143mm (57 x 57 DPI)

Hyperparameters	St John's	Yarmouth
bootstrap	False	False
class_weight	balanced	balanced_subsample
criterion	log_loss	entropy
max_depth	None	None
max_features	sqrt	None
min_samples_leaf	4	3
min_samples_split	8	4
n_estimators	105	125

Table 2: Best parameters for the ETClassifier model for both locations using RandomSearchCV.

461x213mm (57 x 57 DPI)

Hyperparameter	St John's	Yarmouth
max_depth	2	2
min_child_weight	20	12
subsample	0.6	1
lambda	18	6
alpha	24	18
gamma	0	8

Table 3: Best parameters for XGBoost for both locations using RandomSearchCV.

461x167mm (57 x 57 DPI)

Location	Class	Model	Precision	Recall	F1-score	Labels count
St John's	Clear	ETClassifier	0.95	0.95	0.95	3794
		XGBoost	0.96	0.95	0.95	
		biLSTM	0.96	0.95	0.96	
	Fog	ETClassifier	0.70	0.72	0.71	611
		XGBoost	0.76	0.77	**0.76**	
		biLSTM	0.72	0.74	0.73	
Yarmouth	Clear	ETClassifier	0.93	0.92	0.93	3629
		XGBoost	0.92	0.93	0.93	
		biLSTM	0.93	0.92	0.92	
	Fog	ETClassifier	0.63	0.68	**0.65**	746
		XGBoost	0.60	0.60	0.60	
		biLSTM	0.61	0.64	0.63	

Table 4: Classification Report for St John's and Yarmouth on the 2023 test data using ETClassifier, XGBoost and biLSTM.

461x416mm (57 x 57 DPI)

		T2 (K)	RH2 (%)	U10 (ms^{-1})	V10 (ms^{-1})	P_sfc (Pa)
St John's	correlation	0.990	0.902	0.979	0.982	0.998
	RMSE	0.980	6.53	0.690	0.870	2.57
Yarmouth	correlation	0.974	0.845	0.943	0.972	0.999
	RMSE	1.09	8.51	1.01	0.844	0.457

Table 5: Correlations and root mean squared errors of T2, RH2, U10, V10 and P_sfc at St John's and Yarmouth, at 12Z every day from April to August 2024.

467x151mm (57 x 57 DPI)

Domains	D01	D02
Resolution (km)	27	9
No. of grid points (lon × lat)	91 × 91	160 × 160
ETA levels	51	51
Microphysics	Thompson Aerosol-Aware	Thompson Aerosol-Aware
Longwave Radiation	RRTMG	RRTMG
Shortwave Radiation	RRTMG	RRTMG
Land Surface Model	Unified Noah	Unified Noah
Surface Layer Model	MYNN	MYNN
Cumulus Parameterization	Tiedtke	Off
Planetary Boundary Layer	MYNN 2.5	MYNN 2.5

Table 6: Configuration of the WRF model for the pseudo-operational simulation from April to August 2024.

460x257mm (57 x 57 DPI)

Location	Model	Precision	Recall	F1-score	Fog Count
St John's	ETClassifier	0.68	0.72	**0.70**	565
	XGBoost	0.64	0.74	0.68	
	biLSTM	0.65	0.70	0.68	
	WRF only	0.61	0.63	0.62	
Yarmouth	ETClassifier	0.59	0.63	**0.61**	558
	XGBoost	0.47	0.71	0.56	
	biLSTM	0.55	0.60	0.58	
	WRF only	0.41	0.61	0.51	

Table 7: Classification Report for St John's and Yarmouth on the 2024 forecast data using ETClassifier, XGBoost and biLSTM as a post-processing method, and the liquid water from WRF.

396x285mm (57 x 57 DPI)

ETClassifier		XGBoost		biLSTM		WRF only	
St John's							
2915	191	2871	236	2896	219	2878	229
160	**406**	148	**417**	170	**395**	209	**356**
Yarmouth							
2871	242	2659	455	2842	271	2686	428
208	**350**	162	**396**	221	**337**	219	**339**

Table 8: Confusion matrices of each model at St John's and Yarmouth for the 2024 dataset.

461x183mm (57 x 57 DPI)

	St John's	Yarmouth
Precision CI	0.644 - 0.719	0.523 - 0.601
Recall CI	0.662 - 0.739	0.594 - 0.674
F1-score CI	0.660 - 0.722	0.562 - 0.628

Table 9: Bootstrap confidence intervals (CIs) for precision, recall and F1-score at each location for the 2024 dataset.

461x98mm (57 x 57 DPI)