Investigating EMIC Wave Conjunctions: Linking Ground-Based and Space Observations Using Machine Learning

Marcela Moura[1], Hyomin Kim[1], Sung Jun Noh[2], Marc Lessard[3], Eun-Hwa Kim[4]

[1]New Jersey Institute of Technology, [2]Los Alamos National Laboratory, [3]University of New Hampshire, [4]Princeton Plasma Physics Laboratory

SM31C-2379

INTRODUCTION

Electromagnetic Ion Cyclotron (EMIC) waves are low-frequency plasma waves that travel through the Earth's magnetosphere and interact with high-energy particles. These interactions can scatter radiation belt particles into the ionosphere, affecting the dynamics of Earth's magnetosphere and contributing to space weather phenomena.

EMIC waves are typically measured by spacecraft in the magnetosphere and ground-based instruments. However, the occurrence rate obtained from spacecraft data is extremely low. Thus, a space-ground conjunction study is essential. This project aims to improve our understanding of EMIC wave occurrence by starting with ground-based magnetometer observations from Sanikiluaq station (SNK) and identifying corresponding space-based events using Geostationary Operational Environment Satellite (GOES) spacecraft data. **This is the inverse of the traditional method, which begins with space-based observations and looks for ground confirmations.**

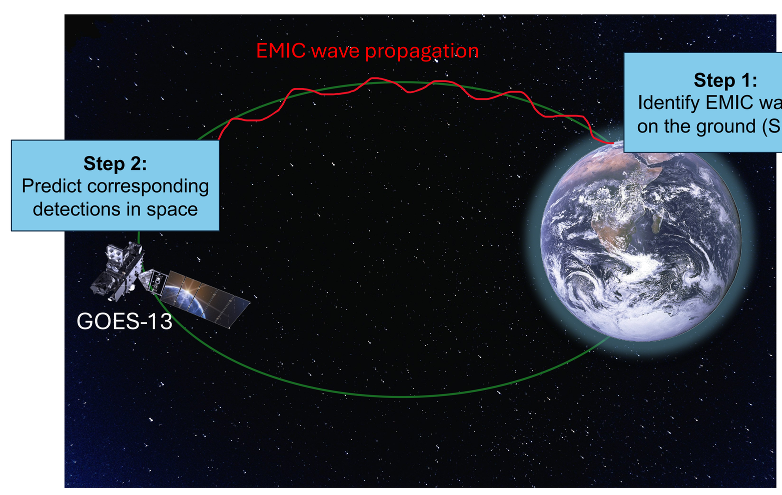


Figure 1: Illustration of EMIC wave propagation starting from ground-based detection at SNK to corresponding observations by the GOES-13 spacecraft.

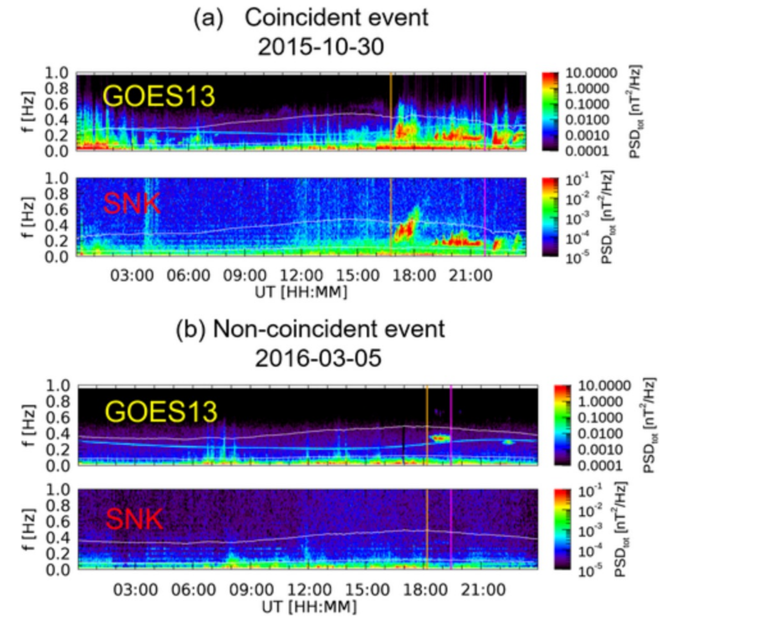


Figure 2: Spectrograms from two events: coincident (top) and non-coincident (bottom). The top spectrogram is from the GOES-13 spacecraft and the bottom spectrogram is from Sanikiluaq station.

METHODOLOGY

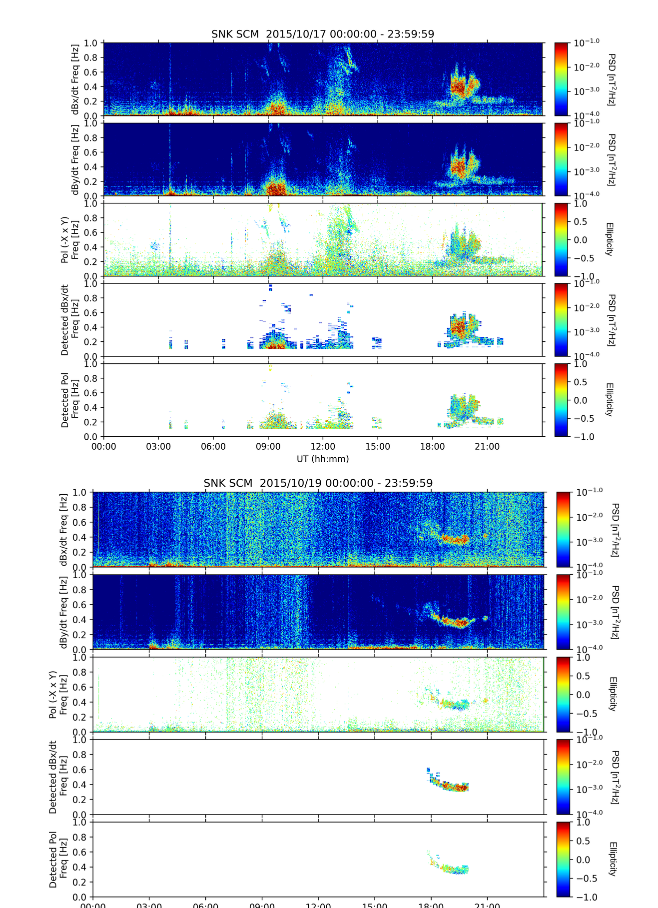


Figure 3: Examples of the wave detection algorithm output from Sanikiluaq (SNK) station. The top panel shows a broadband wave event, while the bottom panel shows a band-limited event.

To analyze these relationships, we applied a Random Forest binary classification model in Python (Fig. 4), using datasets that contain solar wind and geomagnetic information, as well as statistical surveys of wave occurrences in space and on the ground.

Wave events at SNK were detected using a custom-built algorithm (Fig. 3) and temporally aligned with OMNI and GOES parameters. The merged dataset served as the feature set for the classification model.

SMOTE (Synthetic Minority Oversampling Technique) was implemented to mitigate class imbalance due to the rarity of EMIC wave events. The model was optimized using the OPTUNA library to improve predictive performance.

By evaluating feature importance, this study identified key factors influencing EMIC wave generation, providing new insights into their occurrence, and potential for predictive modeling. Future work will include expanding the dataset with more ground stations and spacecraft, which will improve model accuracy.

RANDOM FOREST MODEL

Random Forest is a machine learning algorithm that combines the results of many decision trees to improve prediction accuracy and reduce overfitting. Each tree makes a classification based on a subset of the data and features, and the forest's final output is the majority vote. In this study, the Random Forest model predicts the probability of EMIC wave occurrence in space based on solar wind, geomagnetic indices, and ground wave detections.

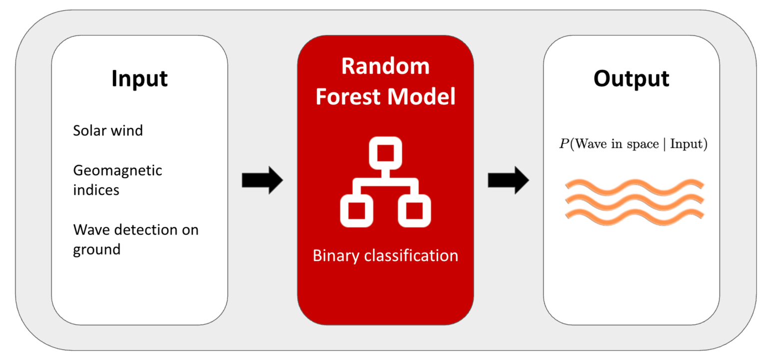


Figure 4: Workflow of the Random Forest binary classification model.

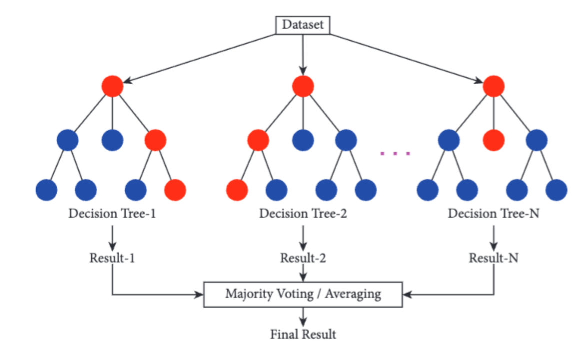


Figure 5: Conceptual diagram of the Random Forest classifier. (Khan, et al. 2021)

Wave Events (1)	Non-wave Events (0)
72044	1803398

Table 1: Number of wave and non-wave points in the dataset, demonstrating the significant class imbalance between wave and non-wave occurrences.

The dataset contains a significant class imbalance, with EMIC wave events representing only a small fraction of the total observations. As shown in Table 1, the number of wave events is much lower than the number of non-wave events, resulting in a dataset dominated by quiet conditions.

This imbalance can cause bias in machine-learning models toward predicting the majority class, leading to high accuracy but poor detection of actual wave events. To address this issue, we applied the Synthetic Minority Oversampling Technique (SMOTE) to increase the representation of wave events in the training set. We also used the OPTUNA library to tune the model's hyperparameters and identify the optimal parameter configuration. Together, these techniques aimed to improve the model's overall performance.

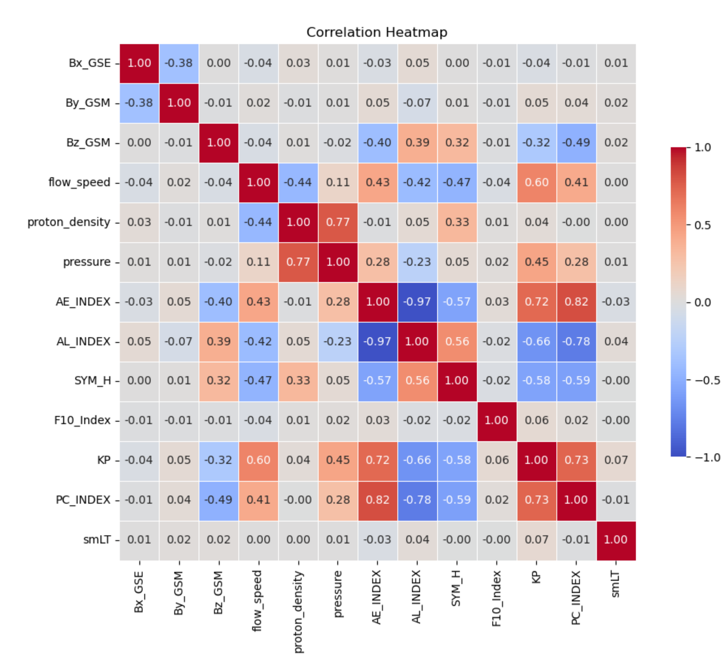


Figure 6: Correlation heatmap of input features used in the model

- Strong correlations between proton density, pressure, and flow speed (solar wind properties)
- AE, AL, Kp, and PC indices show high inter-correlation, reflecting linked geomagnetic activity
- Used to ensure data consistency before training the model

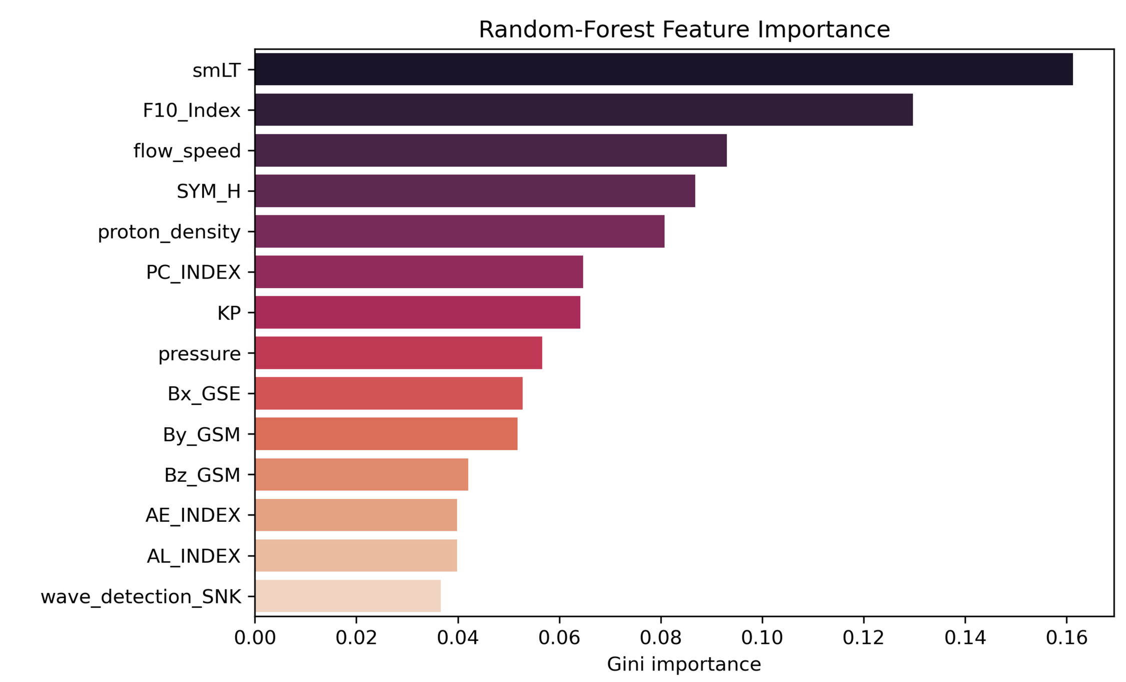
Feature	Units	Explanation
datetime	yyyy-mm-dd HH:MM:SS	Measurement of time
wave_detection_SNK	0 / 1	Ground-station wave flag from Sanikiluaq (SNK): 1 = EMIC wave detected
Bx (nT), GSE	nT	Interplanetary magnetic field (IMF) in Geo solarecliptic coordinate system (GSE) coordinates
By (nT), GSM	nT	IMF in GSM coordinates
Bz (nT), GSM	nT	IMF in GSM coordinates
flow_speed	km/s	Solar wind velocity
proton_density (Np)	cm^3	Solar-wind proton number density
pressure	nPa	Solar-wind dynamic pressure
AE_Index	nT	Auroral Electrojet index
AL_Index	nT	Aurora index
SYM_H	nT	Ring current index
F10_Index	sfu	F10.7 cm solar radio flux
Kp	-	Planetary Kp index (3-hour geomagnetic activity)
PC_index	-	Polar Cap index
sMLT	hours (0-24)	Satellite magnetic local time at GOES-13
wave_detection_GOES	0 / 1	Ground-station wave flag at GOES-13: 1 = EMIC wave detected

Table 2. Description of input parameters used in the model

ACKNOWLEDGEMENTS

The work at the New Jersey Institute of Technology was supported by NSF under grants AGS-2133837 and AGS-2247398, as well as the Grace Hopper Research Institute (GHRI) Artificial Intelligence Summer Fellowship Program at New Jersey Institute of Technology.

RESULTS AND DISCUSSION



The feature importance ranking indicates that the model relies mostly on magnetic local time (sMLT) and several solar wind and geomagnetic activity parameters (F10.7, solar wind speed, SYM-H, proton density, Kp, and PC index). These results are consistent with known EMIC wave drivers: waves are more likely to occur during periods of elevated geomagnetic activity and solar wind pressure.

Figure 7: Random Forest feature importance scores for predicting EMIC wave occurrences.

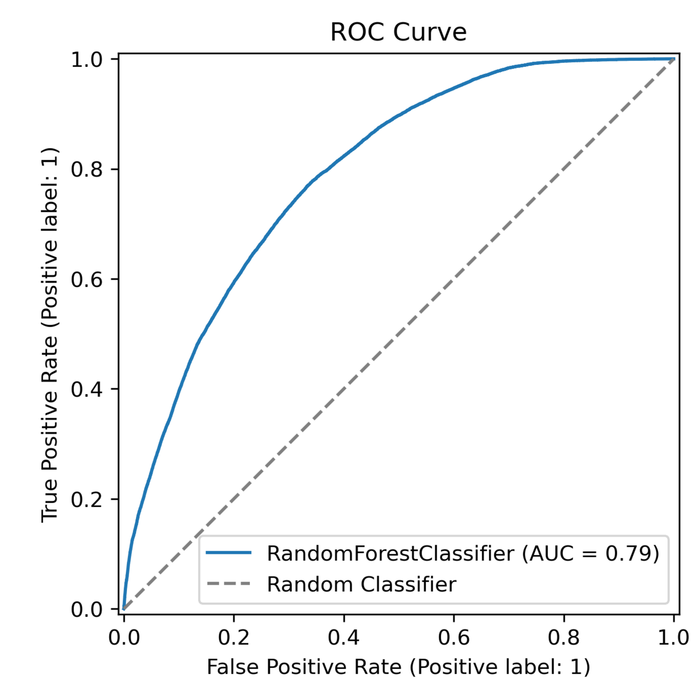


Figure 8: Receiver Operating Characteristic (ROC) curve for the Random Forest classifier. The model achieves an AUC of 0.79, indicating good overall separability between wave and non-wave events compared to a random model.

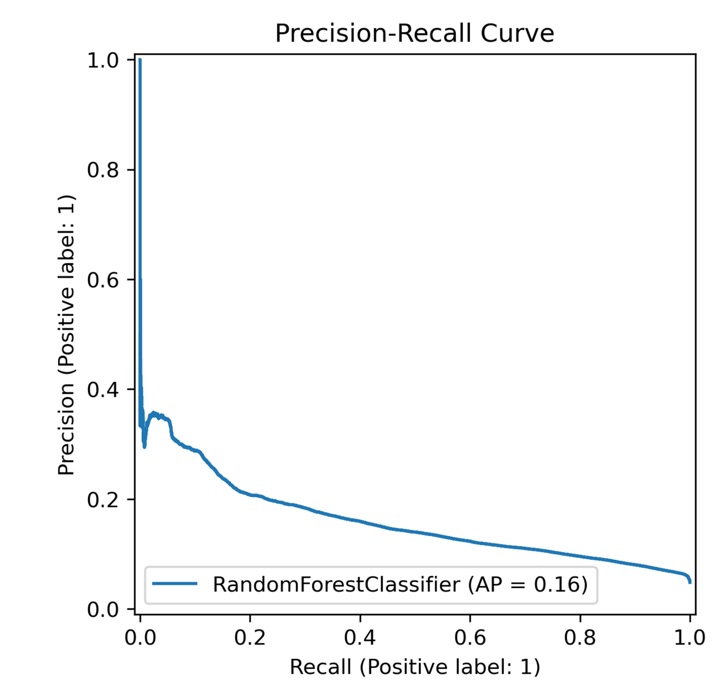


Figure 9: Precision-recall curve for the Random Forest classifier. The model achieves a low precision of 0.16, showing the model's limited ability to correctly identify positive EMIC wave events.

The event timeline (Fig. 10) highlights why prediction is challenging, with wave events representing only 3.8% of the total dataset. Conjunction events between SNK and GOES are rare, while single-site detections occur far more frequently. In this dataset, space-based detections greatly outnumber ground detections, indicating (1) the need for improved ground-based wave detection methods and (2) the importance of addressing severe class imbalance in the training data.

Due to the sporadic nature of EMIC waves, the model has a low ability to uniquely identify conjunctions based on solar wind and geomagnetic parameters alone. Waves observed on the ground may also originate from different magnetospheric regions as compared to GOES detections.

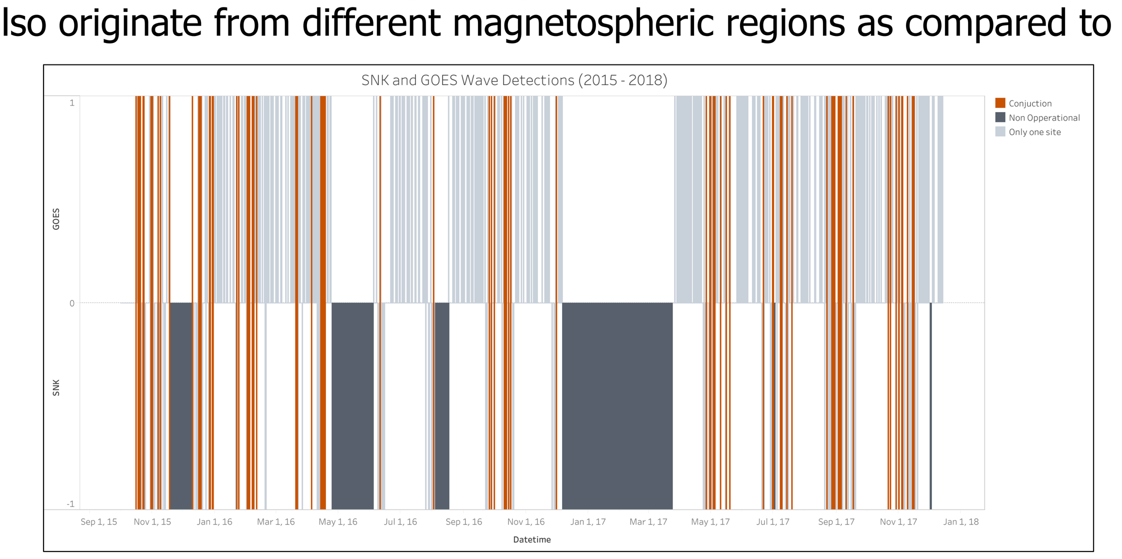


Figure 10: Timeline of SNK and GOES wave detections from 2015–2018. Red lines mark conjunction events where waves were detected at both sites, while gray lines indicate single-site detections. Black lines indicate times where the SNK magnetometer was non-operational.