Space Weather

RESEARCH ARTICLE

10.1029/2021SW002884

Key Points:

- A model for predicting low latitude scintillation occurrence after sunset is created based on gradient boosting algorithm
- The model with input of total electron content, equatorial hmF2 and foF2 before sunset, can well capture strong scintillation occurrence after sunset
- The gradient boosting algorithm is suggested to be effective in predicting low latitude strong scintillation occurrence on a daily basis

Correspondence to:

G. Li,
gzlee@mail.iggcas.ac.cn

Citation:

Zhao, X., Li, G., Xie, H., Hu, L., Sun, W., Yang, S., et al. (2021). The prediction of day-to-day occurrence of low latitude ionospheric strong scintillation using gradient boosting algorithm. *Space Weather*, *19*, e2021SW002884. https://doi.org/10.1029/2021SW002884

Received 16 AUG 2021
Accepted 30 NOV 2021

The Prediction of Day-to-Day Occurrence of Low Latitude Ionospheric Strong Scintillation Using Gradient Boosting Algorithm

Xiukuan Zhao[1,2,3], Guozhu Li[3,4,5], Haiyong Xie[2,3,4], Lianhuan Hu[3,4], Wenjie Sun[3,4,5], Sipeng Yang[3,4], Yi Li[2,3,4], Baiqi Ning[3,4,6], and Hisao Takahashi[7]

[1]Mohe Observatory of Geophysics, Institute of Geology and Geophysics, Chinese Academy of Sciences, Beijing, China, [2]Geophysics Center, National Earth System Science Data Center, Beijing, China, [3]Key Laboratory of Earth and Planetary Physics, Institute of Geology and Geophysics, Chinese Academy of Sciences, Beijing, China, [4]Beijing National Observatory of Space Environment, Institute of Geology and Geophysics, Chinese Academy of Sciences, Beijing, China, [5]College of Earth and Planetary Sciences, University of Chinese Academy of Sciences, Beijing, China, [6]Innovation Academy for Earth Science, Chinese Academy of Sciences, Beijing, China, [7]Divisão de Aeronomia, Instituto Nacional de Pesquisas Espaciais, São José dos Campos, Brazil

Abstract Ionospheric scintillations caused by equatorial plasma bubbles (EPBs) can seriously affect various high technology systems based on Global Navigation Satellite System (GNSS) signals at equatorial and low latitudes. A reliable prediction of ionospheric scintillation occurrence is critical to relieve the effect. Using the long-term ground-based GNSS receiver and ionosonde data collected in the Brazilian longitude sector during 2012–2020, an ionospheric strong scintillation prediction model based on the gradient boosting algorithms extreme gradient boosting (XGBoost), light gradient boosting machine (LightGBM), and CatBoost is created and tested. It is for the first time that the XGBoost, LightGBM, and CatBoost are utilized to predict the day-to-day occurrence of regional ionospheric scintillation during post-sunset hours. The relative importance of different parameters affecting EPB/scintillation occurrence for building the prediction model is examined. A comparison of daily scintillation occurrence from the modeled and observed results during 2014 (solar maximum) and 2020 (solar minimum) shows that the gradient boosting algorithms are effective for predicting strong scintillations over low latitude, with a prediction accuracy of ~85%. The results suggest that the trained model with input of total electron content, equatorial F layer peak height and critical frequency before sunset could be well employed to predict the occurrence/nonoccurrence of intense scintillations over low latitude after sunset on a daily basis.

Plain Language Summary Ionospheric scintillation is the rapid fluctuation of radio signals traversing through ionospheric irregularities. Severe scintillation can cause loss of lock for the systems using Global Navigation Satellite System signals. The dependences of scintillation on seasonal, solar and geomagnetic activities have been widely studied, but its day-to-day variability and prediction still remain a challenge. The relationship between scintillation occurrences and a variety of factors is complex. The machine learning algorithm could handle nonlinear problems and thus uncover the implicit correlations between multiple factors. The gradient boosting, which is a type of machine learning technique, has been demonstrated to be effective in many fields, such as in the field of intrusion detection. Here, we employ the gradient boosting algorithm, together with long-term observations in the Brazilian longitude sector to investigate if the day-to-day occurrence of low latitude ionospheric scintillation could be predicted. The results show that with limited input parameters, the prediction accuracy for scintillation occurrence on a daily basis reach ~85%, suggesting that the gradient boosting algorithms are effective for predicting strong scintillations over low latitude. This opens a possibility for scintillation forecasting with acceptable accuracy under the conditions without physical model and powerful computing capability.

1. Introduction

Ionospheric scintillation is one of the main error sources that can reduce the quality of communication and navigation. Ionospheric scintillations mainly occur at equatorial and low latitudes, and polar regions (e.g., Abdu et al., 1985; Alfonsi et al., 2011; Basu et al., 1999; Spogli, Alfonsi, Cilliers, et al., 2013). Equatorial and low latitude ionospheric scintillation occurs quite often between local sunset and midnight. A close relationship between the occurrences of Equatorial plasma bubbles (EPBs) and severe ionospheric scintillations has been observed at low latitudes (e.g., Manju et al., 2011; Xiong et al., 2016). Previous all-sky airglow imager observations showed that EPBs usually extend in a wide range, but they contain a lot of small-scale structures (e.g., Otsuka et al., 2002; Takahashi et al., 2015).

It is generally accepted that, the EPBs are generated through the generalized Rayleigh-Taylor (R-T) instability under favorable conditions. With their development initiated at the bottomside plasma density gradient region of a rapidly rising F layer after sunset, EPBs can rise to higher altitudes above the F layer peak to extend to a wide latitude band along magnetic field lines (Kelley, 2009). The pre-reversal enhancement (PRE) of the eastward electric field during sunset hours, which elevates the F layer to higher altitudes, contributes to the R-T instability growth. A strong correlation between the climatological behaviors of EPBs and vertical plasma drifts driven by the PRE was observed at specific locations (Fejer et al., 1999) and globally (Li et al., 2007). The growth rate of the R-T instability, together with the initial density perturbation at the bottomside of the F layer control the occurrence of EPBs (Abdu et al., 2009; Tsunoda et al., 2010). Despite that the dependency of EPB on seasonal variations, solar and magnetic activities has been well investigated, the day-to-day variability of EPB and ionospheric scintillation occurrences still remains a challenge (Li et al., 2021, and the references therein).

Due to the importance of ionospheric scintillation in space weather application, a lot of research works have been carried out on the modeling of ionospheric scintillations (e.g., Grzesiak et al., 2018; Materassi et al., 2020; Priyadarshi, 2015; Rezende et al., 2010; Spogli, Alfonsi, Romano, et al., 2013). de Lima et al. (2015) used the data obtained at São Luís, Brazil together with the neural network technique to build prediction model of ionospheric scintillation. Rezende et al. (2010) used data from two stations in Brazil, São Luís and São José dos Campos and the bagging-CART method to build ionospheric scintillation model. Grzesiak et al. (2018) used five days data collected by seven PolaRxS receivers over São Paulo state region in Brazil and continuity equation to forecast scintillation. On the other hand, Retterer (2010) and Nugent et al. (2021) used physics-based model to forecast low-latitude ionospheric scintillation. Retterer et al. (2005) built a theoretical model with the vertical plasma drift velocity observed by the Jicamarca incoherent radar in seven days to forecast the occurrence of spread F. Sousasantos et al. (2017) proposed a mathematical approach to the EPB forecasting based on five days of Digisonde data collected in São Luís, Brazil. However, due to the limited data employed in these studies, it's difficult to evaluate if the models are suitable for the prediction of scintillation occurrence on a daily basis.

Considering the complex relationship between the occurrences of EPBs/ionospheric scintillations and various factors including external driving forces and background ionosphere (Abdu, 2001), the machine learning approach, which can deeply find implicit relationships between various variables and well handle the nonlinear problems (Camporeale, 2019), could be a promising method for predicting ionospheric scintillations. In this regard, the gradient boosting decision tree (GBDT) method has shown great prediction performance in many fields, such as in the field of intrusion detection (Tama & Rhee, 2019). Under the framework of GBDT, extreme gradient boosting (XGBoost; Chen & Guestrin, 2016), light gradient boosting machine (LightGBM; Ke et al., 2017) and unbiased boosting with categorical features (CatBoost; Prokhorenkova et al., 2017) have been proposed recently. To the best of our knowledge, the XGBoost, LightGBM and CatBoost algorithms have not been used to predict ionospheric scintillations. In this work, these boosting algorithms are used for the first time to predict amplitude scintillation (S_4) level over low latitudes in the Brazilian longitude sector. Using the long-term observations from eight GPS receivers, together with the equatorial ionosonde measurements at São Luiz, we built a new prediction model of strong scintillation (with $S_4 > 0.5$) occurrence. Comprehensive performance of the models using XGBoost, LightGBM and CatBoost algorithms is analyzed. A comparison of modeled results and observations on a daily basis during 2014 and 2020 is presented and discussed.

2. Input and Output Parameters

Based on the previous studies of factors affecting the generation of EPBs and scintillations, we employed the following items as input parameters to build the ionospheric scintillation prediction model, (a) local ionospheric total electron content (TEC), (b) equatorial F layer critical frequency (foF2), and (c) peak height (hmF2) over São Luiz, (d) solar activity characterized by F10.7, (e) geomagnetic activity characterized by K_P, (f) solar wind interplanetary magnetic field Bz component (IMF Bz), and (g) seasonal variation (day number). The output

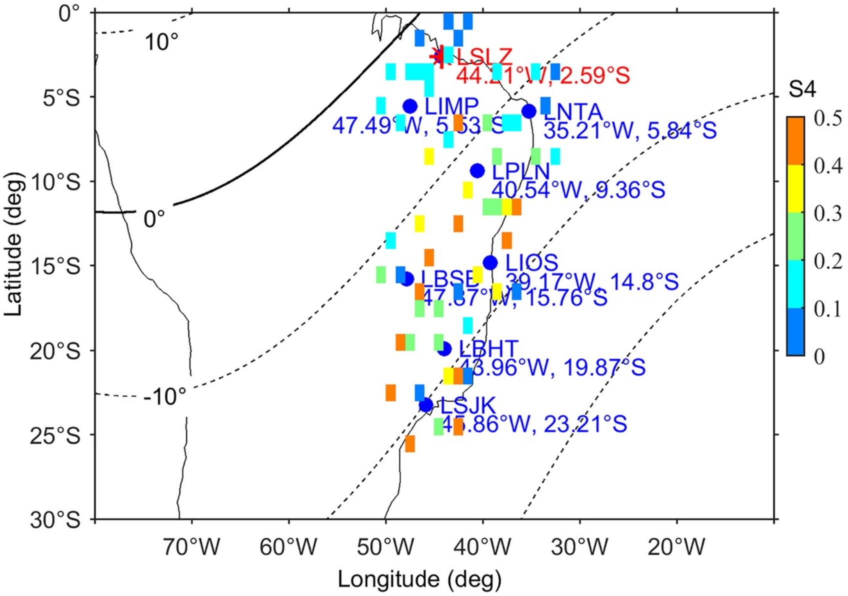


Figure 1. Geographic distributions of the Digisonde (red asterisk) and GPS receivers (blue circles). The station codes and geographic coordinates are superimposed on the figure. The magnetic dip equator, and dip latitudes of 10°, −10°, −20°, and −30°, calculated using the IGRF-13 (Alken et al., 2021), are shown as solid and dotted black curves, respectively. The S_4 index from 9:55 to 10:00 UT on 21 November 2015, are superimposed on the figure, binned into grids of 1° (latitude) × 1° (longitude).

parameter is the status of ionospheric amplitude scintillation, that is, with strong scintillation (S) or without strong scintillation (NS).

The ionospheric scintillation data used in this study were collected from eight GPS receivers belonging to the Low-Latitude Ionospheric Sensor Network (LISN; Valladares & Chau, 2012). These receivers are located at equatorial and low latitudes and their geographic distributions are shown in Figure 1. The average period used to calculate S_4 is 1 min, while the raw GNSS measurements are at 50 Hz. Only the scintillation data from GPS satellites with elevation angles higher than 30° are used. This angle threshold minimizes the effects of geometric factors in the S_4 index calculation due to tropospheric scattering and multipath propagation (Rezende et al., 2010). The S_4 index was binned into grids of 1° (latitude) × 1° (longitude) with 5-min resolution (an example shown in Figure 1). The S_4 indices obtained from all the satellites with ionospheric piercing point falling into the same grid are averaged in 5-min interval.

Through grouping the S_4 and TEC by dip latitude, we got the temporal variation of amplitude scintillation and TEC at different dip latitudes. Figure 2a shows an example of S_4 (gray dots) and its peak value (red dot) at dip lat. −12° during pre-midnight hours on 25 March 2015. Figure 2b shows the temporal variation of TEC (gray curve) and its moving average (thick gray curve) using a time window of 10 minutes. The Digisonde data from São Luís (2.6°S, 44.2°W, dip lat. −3°, marked as LSLZ in Figure 1), which is located close to the dip equator, was used to obtain the background ionospheric variation over the magnetic equator. The variability of the F layer critical frequency (foF2) and peak height (hmF2) as a function of local time (LT) is shown in Figures 2c and 2d. The values of the parameters TEC, foF2, and hmF2 obtained during the period of three hours before and one hour after sunset (highlighted by red thick curves) were employed to train the prediction model. Considering that the serious effect of scintillation on the transionospheric radio systems is signal loss, this effect occurs mainly when scintillation is strong. We used a threshold of 0.5, which is similar to previous studies (e.g., Jiao et al., 2017; Kil et al., 2002), to characterize the level of ionospheric scintillation into two classes: with strong scintillation (S) or without strong scintillation (NS).

For the input parameters F10.7, K_p and IMF Bz, the daily value of F10.7, the mean, maximum and minimum values of K_p and IMF Bz during the 24 hr period from the previous sunset to the present sunset, expressed as meanK_p, maxK_p, minK_p, meanBz, maxBz, and minBz respectively were used. The seasonal factors are described

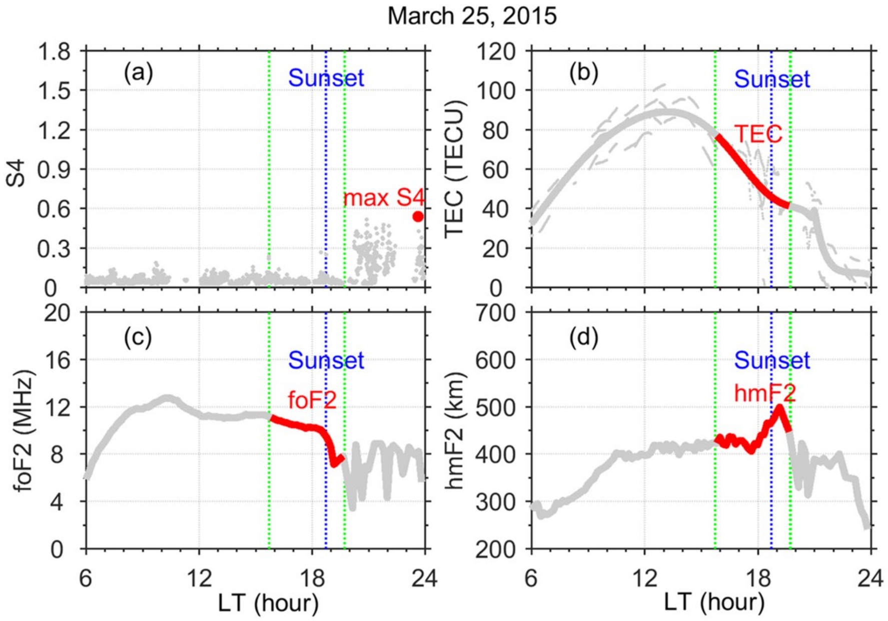


Figure 2. Examples of data used in the model building process. (a and b) S_4 and total electron content as a function of local time (LT) over dip latitude −12°. (c-d) foF2 and hmF2 as a function of LT from the station São Luiz (LSLZ). The red thick curves in b-d show the values before and after sunset employed to train the prediction model.

as DNS = sin (2π × DoY/DiY) and DNC = cos (2π × DoY/DiY), where DoY is day of year ($1 \leq \text{DoY} \leq 365$ or 366 for leap years) and DiY is total days in a year (365 or 366 for leap years) (Zhao et al., 2014).

3. Construction of Prediction Model

In general, the construction process of the prediction model is shown in Figure 3. First, the data set is built with 82 values of the input parameters and 1 output. The 82 values include 24 TEC values, 24 foF2 values, 24 hmF2 values and 10 other values (dip latitude, F10.7, meanKP, maxKP, minKP, meanBz, maxBz, minBz, DNS, and DNC). All these parameters are described in the previous section. The output is the level of ionospheric amplitude scintillation, that is, with or without scintillation above the specific threshold, that is, 0.5. Second, the data obtained are separated as training set and test set. Third, a five-fold cross validation (CV) approach is utilized to tune the model hyperparameters. Fourthly, using the optimal hyperparameters configuration, the optimal prediction model is fitted based on the training set. Finally, the test set is adopted to evaluate model performance according to the prediction results. The machine learning algorithms, hyperparameter optimization process and model evaluation methods are descripted as follows.

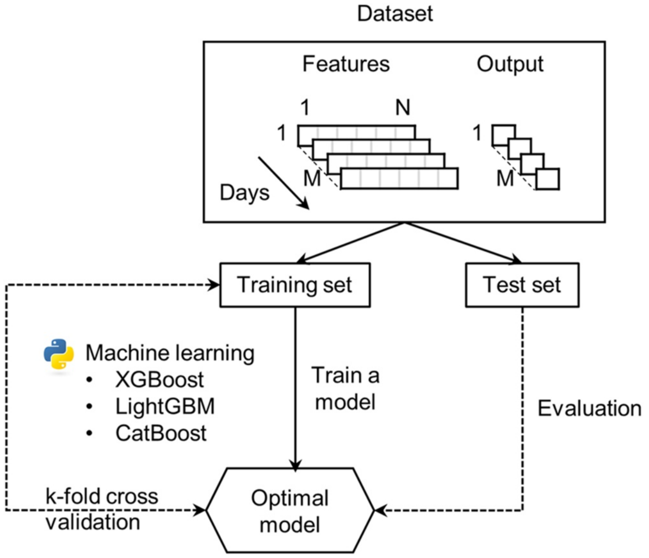


Figure 3. Construction process of the prediction model.

3.1. Machine Learning Algorithms

Previous studies showed that the XGBoost, LightGBM, and CatBoost algorithms have their own advantages in different data sets (Bentéjac et al., 2020). To investigate which algorithm is more suitable for building the ionospheric scintillation prediction model, all the three algorithms are employed and their modeled results are compared.

The XGBoost algorithm, which was proposed by Chen and Guestrin (2016), has received widespread attentions because of its excellent performance in Kaggle's machine learning competitions. Similar to the gradient boosting, XGBoost builds an additive expansion of the objective function by minimizing a loss function. Considering that XGBoost is focused only on decision trees as base classifiers, a variation of the loss function is used to control the complexity of the trees (Bentéjac et al., 2020). XGBoost applies level-wise tree growth and uses pre-sorted algorithm and Histogram-based algorithm for computing the best split.

The LightGBM is an extensive library that implements gradient boosting (Ke et al., 2017). LightGBM is unique in the aspect that it can construct trees using gradient-based one-sided sampling (GOSS). GOSS keeps all the instances with large gradients and performs random sampling on the instances with small gradients. GOSS allows LightGBM to quickly find the most influential cuts. Another feature of LightGBM is exclusive feature bundling (EFB). EFB technique bundles sparse features into a single feature. This can be done without losing any information when those features do not have non-zero values simultaneously. Both GOSS and EFB provide further training speed gains. Unlike the XGBoost which applies level-wise tree growth, the LightGBM applies leaf-wise tree growth. Level-wise approach grows horizontal while leaf-wise grows vertical.

The CatBoost, which is also based on gradient boosting, was developed by Prokhorenkova et al. (2017). Two critical algorithmic advances introduced in CatBoost are the implementation of ordered boosting, a permutation-driven alternative to the classic algorithm, and an innovative algorithm for processing categorical features. Both techniques were created to fight a prediction shift caused by a special kind of target leakage present in all currently existing implementations of gradient boosting algorithms. CatBoost distinguishes itself from LightGBM and XGBoost by focusing on optimizing decision trees for categorical variables. It can deal with categorical features during training time instead of preprocessing time.

3.2. Hyperparameter Optimization

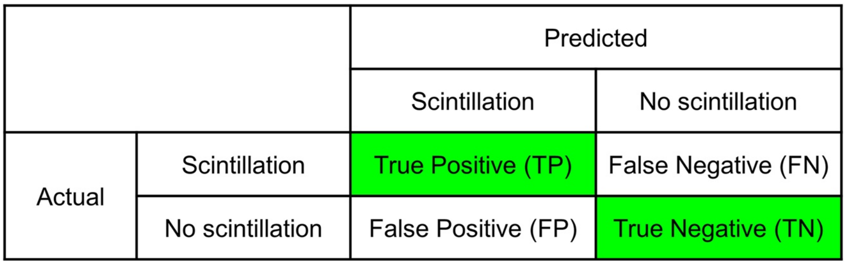
Most machine learning algorithms contain hyperparameters that need to be tuned. Generally, the hyperparameters search methods include grid search, Bayesian optimization, heuristic search and randomized search. In this study, grid search and five-fold CV approach were utilized for hyperparameters configuration. For five-fold CV, the original data set is randomly split into five equal size subsamples. Among them, a singular subsample is adopted as the validation set, and the other four subsamples are used as the training subset. This procedure is repeated five times, until each subsample is selected as a validation set once. Thereafter, the average accuracy of these five validation sets is used to determine the optimal hyperparameter values (Liang et al., 2020).

3.3. Model Evaluation

Robust quantitative measures are essential to compare the performance of different prediction methods. In machine learning there are many metrics for measuring classifier performance over a set of data (Camporeale, 2019). Since the receiver operating characteristic (ROC) curve was proposed as a comparison metric in machine learning, it has become the most commonly used way to visualize the performance of a binary classifier. A machine learning classification model can be used to predict the actual class of the data point directly or predict its probability of belonging to different classes. The latter provides more control over the predicted result. We can determine a threshold to interpret the result of the classifier. Through changing threshold values, a ROC curve can be created. The area under the curve (AUC), that is, the area between the ROC curve and the coordinate axis, is the best way to summarize classifier performance in a single number. The higher the AUC, the better the performance of the model for distinguishing between the positive and negative classes.

Figure 4 shows the confusion matrix containing information about the actual and predicted classes generated by a classification system. True negative (TN) and false negative (FN) represent the conditions where no scintillation above the threshold occurs and no scintillation is predicted, and scintillation occurs but no scintillation is predicted, respectively. True positive (TP) and False Positive (FP) represent the conditions where scintillation occurs and scintillation is predicted, and no scintillation occurs but scintillation is predicted, respectively. These four values provide the basis from which all prediction metrics are derived. The ROC curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings, where TPR is calculated as TP/(TP + FN) and FPR is calculated as FP/(FP + TN).

To further compare the modeled results from the different algorithms, the accuracy and F1, which are generally used in machine learning, were also computed. The accuracy is the proportion of successful predictions, which is calculated as (TP + TN)/(TP + FN + FP + TN). F1 can be interpreted as a weighted average of precision and



		Predicted	
		Scintillation	No scintillation
Actual	Scintillation	True Positive (TP)	False Negative (FN)
	No scintillation	False Positive (FP)	True Negative (TN)

Figure 4. The confusion matrix showing correct (TP and TN) and incorrect (FN and FP) predictions made by the model compared to the observational data (i.e., with or without strong scintillations).

recall, $F1 = 2 \times (\text{precision} \times \text{recall})/(\text{precision} + \text{recall})$, where the precision is calculated as TP/(TP + FP) and the recall is calculated as TP/(TP + FN).

4. Results and Discussion

In this study, the XGBoost (version 1.2.0), LightGBM (version 3.0.0), and CatBoost (version 0.24.2) libraries were used. These Python libraries can be installed via pip.

4.1. Relative Importance of Input Parameters

To study the influence of different parameters on scintillation occurrence, the input parameters were classified into five groups, that is, GLO, TEC, foF2, DAY, and hmF2. GLO is for GLObal parameters which includes F10.7, meanKP, maxKP, minKP, meanBz, maxBz, minBz, and Dip latitude. The groups TEC, foF2, and hmF2 include themselves and Dip latitude. The group DAY includes DNS, DNC, and Dip latitude.

We used randomly 70% of the data during 2012–2019 as training set to build the model and the remaining 30% data as test set. The XGBoost, LightGBM, and CatBoost with default hyper-parameters were trained and expressed as D-XGB, D-LGB, and D-Cat. The accuracy, AUC and F1 of the prediction models trained with different groups of parameters are shown in Figure 5. From the top panel of Figure 5, we can see that the models using the DAY group parameters have worst accuracy, which is 79.3% for D-XGB, 78.4% for D-LGB and 78.6% for D-Cat. This indicates that the DAY group parameters are less important to the scintillation prediction model. The accuracy rank of the AVE (average value of D-XGB, D-LGB, and D-Cat) is hmF2 > foF2 > GLO > TEC > DAY. The AUC and F1 also show similar pattern to the accuracy. The highest accuracy of the hmF2 group indicates that, out of a total of five groups, the variation of equatorial ionospheric F layer peak height during the evening hours plays a dominant role in the scintillation occurrence. Furthermore, if we use all these parameters, we can get the best outcomes, with accuracy up to 93.0% for AVE.

The prediction results of D-XGB using different groups of parameters along with the observation are shown in Figure 6. The groups of parameters employed to derive the results shown in Figures 6b–6g are DAY, TEC, GLO, foF2, hmF2, and all the parameters, respectively. In Figure 6a, the observational results clearly show the seasonal, latitudinal and solar activity dependences of scintillation. In Figure 6b, using the DAY group parameters, the model captures the seasonal and latitudinal variations. However, the solar activity dependence is not presented. From Figures 6c–6g, the models capture the seasonal, latitudinal and solar activity dependences, and the consistency between the modeled and observational results is getting much better. Comparing Figure 6a with Figure 6g, it is clear that the model built with all the parameters shows a good agreement with the observation.

4.2. Comparison of Results Using Different Algorithms and Different S_4 Thresholds

For the XGBoost, LightGBM, and CatBoost, a lot of hyperparameters need to be tuned to get a better model. In this study, some critical hyperparameters in XGBoost, LightGBM, and CatBoost algorithms were tuned. As shown in Table 1, the search values of different hyperparameters are specified. In particular, for different algorithms, the search range of the same hyperparameter is kept consistent. We used randomly 70% of the data

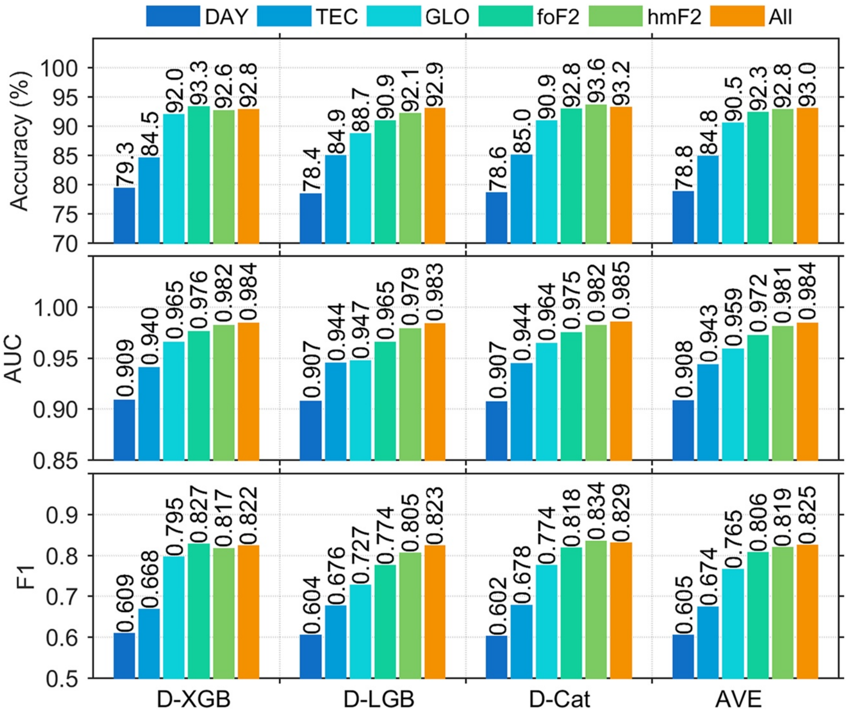


Figure 5. The accuracy, area under the curve and F1 of prediction models using different groups of parameters. The AVE shown in the right part of each panel represents the average value of the three algorithms (D-XGB, D-LGB, and D-Cat).

containing all parameters as training set and the remaining 30% of the data as test set. Based on the grid search and five-fold cross validation, the optimal values for each set of hyperparameters were obtained (Table 1).

Using the tuned and default hyperparameters of the corresponding Python packages, the performance of the three algorithms was analyzed. Table 2 displays the evaluation metrics of different models. The best results are highlighted using a bold font. The models based on the XGBoost, LightGBM and CatBoost algorithms with tuned hyper-parameters are expressed as T-XGB, T-LGB, and T-Cat, respectively. It can be seen from Table 2 that the T-Cat achieves best performance, with the accuracy of 93.73%, AUC is 0.9860 and F1 is 0.8407. It is relevant to mention that the difference in performance of the six models is very small. In fact, the difference in model performance between the tuned and default hyper-parametrizations is also not significant. The result indicates that all the six models can work well to predict scintillation.

Figure 7a shows the performance for T-Cat. The ROC curve was created by plotting the TPR against the FPR. In a ROC curve, a higher *x*-axis value indicates a higher number of FP than TN, while a higher *y*-axis value indicates a higher number of TP than FN. The optimal point, which is nearest to the top left corner (as shown in Figure 7a), represents the optimal compromise between TPR and FPR. It has the highest TPR together with the lowest FPR. The associated confusion matrices of the optimal point are shown in the embedded chart of Figure 7a. It is clear from the confusion matrices that the TP is 2,415, the FN is 119, the FP is 796 and the TN is 11,269. TPR = TP/(TP + FN) = 2,415/(2,415 + 119) = 0.9530. FPR = FP/(FP + TN) = 796/(796 + 11,269) = 0.0660. TPR (FPR) is the y-value (x-value) of the optimal point. The accuracy is calculated as (TP + TN)/(TP + FN + FP + TN), 93.73%.

Besides the threshold ($S_4 > 0.5$) employed in the above analysis, we also used different thresholds of S_4 (from 0.2 to 0.7 with an interval of 0.1) for characterizing the occurrence or nonoccurrence of scintillations with different strength and rebuilt the T-Cat model. The prediction accuracy of T-Cat as a function of S_4 threshold is shown in Figure 7b. We can see that for the threshold 0.2, the accuracy is ∼85%, which is obviously less than the accuracy (>90%) for the thresholds 0.3–0.7. This suggests that weak scintillations are more difficult to be forecasted than the moderate and strong scintillations. Since the main topic of this study is to predict strong scintillations which have more significant impact on satellite communication and navigation quality, only the threshold 0.5 is employed in the following analysis.

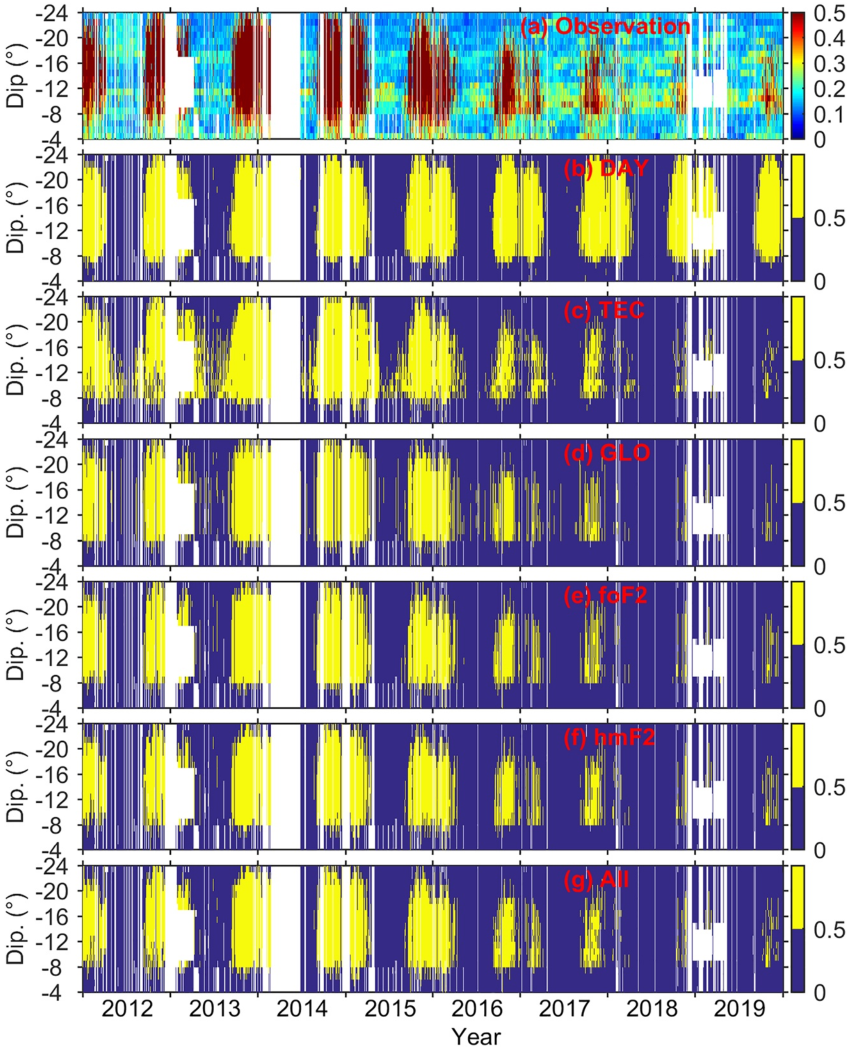


Figure 6. Comparison of (a) observations and (b–g) prediction results of D-XGB using different groups of parameters. The blank area in the plots represent that there is no data for part of the input parameters.

4.3. Scintillation Prediction on a Daily Basis Using Equatorial Ionosonde Observations

To evaluate the prediction capability of the T-Cat model for strong scintillations on a daily basis, we took the equatorial F layer peak height (hmF2) and critical frequency (foF2) over São Luiz as the input parameters and used the data during 2012–2013 and 2015–2019 to retrain the model. The data in 2014 (solar maximum) and 2020 (solar minimum) were employed as test data set.

Figures 8a and 8b shows the observations in 2014 and the prediction results of strong scintillation. In general, the prediction results show a good agreement with the observations in the seasonal and latitudinal dependences. Figure 8c shows the FP (blue) and FN (red) events as functions of month/day and latitude. The ROC curve for the T-Cat in 2014 is shown in Figure 8d. It can be noted from the confusion matrix that out of a total of 1,514 strong scintillation events, 1,247 (267) events were successfully (not) predicted. For the events without strong scintillations, 1,740 events were identified, while 275 events were misclassified. The corresponding accuracy is 84.64%.

Figure 9 shows the results in 2020. In general, the prediction results in the seasonal and latitudinal dependences also exhibit good agreement with the observations. In Figure 9d, the confusion matrix shows that 142 (18) out of 160 strong scintillation events were successfully (or not) predicted. For the events without strong scintillations, 3,618 events were identified, while 668 events were misclassified. The corresponding accuracy is 84.57%. The

Table 1
The Default Values, Possible Values for the Grid Search and Optimal Values of Hyperparameters for the XGBoost, LightGBM, and CatBoost Algorithms

Algorithm	Hyperparameter	Default value	Grid search values	Optimal value
XGBoost	learning_rate	0.3	0.025, 0.05, 0.1, 0.2, 0.3	0.1
	gamma	0	0, 0.1, 0.2, 0.3, 0.4, 1.0, 1.5, 2.0	0
	max_depth	6	3, 6, 9	9
	colsample_bytree	1	0.5, 0.7, 0.9, 1.0	0.9
	subsample	1	0.15, 0.5, 0.75, 1.0	1.0
LightGBM	learning_rate	0.1	0.025, 0.05, 0.1, 0.2, 0.3	0.1
	min_split_gain	0	0, 0.1, 0.2, 0.3, 0.4, 1.0, 1.5, 2.0	0.3
	num_leaves	31	7, 31, 127	127
	colsample_bytree	1	0.5, 0.7, 0.9, 1.0	0.9
	subsample	1	0.15, 0.5, 0.75, 1.0	0.15
CatBoost	learning_rate	Auto	0.025, 0.05, 0.1, 0.2, 0.3	0.05
	max_depth	6	3, 6, 9	9
	leaf_estimation_iterations	Auto	1, 10	10
	rsm	1	0.5, 0.7, 0.9, 1.0	0.7
	l2_leaf_reg	3	1, 3, 6, 9	9

model achieves a good prediction (with accuracy of ~85%) in both 2014 (solar maximum) and 2020 (solar minimum).

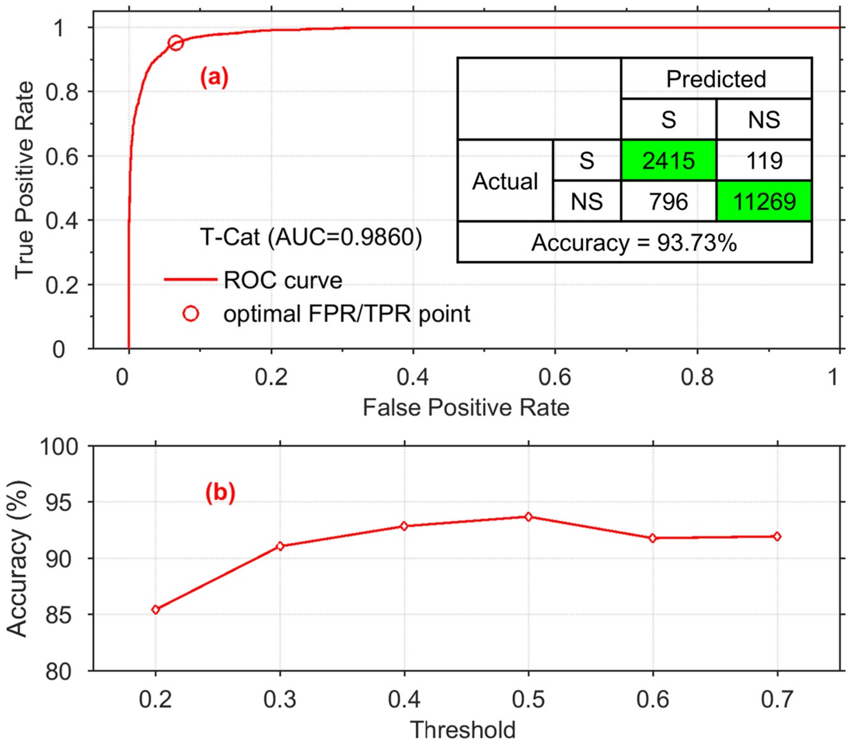
Regarding the discrepancies between the predicted results and observations, there could be several possibilities. Generally, to successfully predict the occurrences of ionospheric scintillations over a given longitude sector, Li et al. (2021) suggested that it is required to determine if EPBs will be generated locally or be generated at neighboring longitudes and travel into the given sector through zonally drifting, and if EPBs will cause scintillations. For the present FN condition where strong scintillations occurred but were not predicted by the model, one possibility is that some of the EPBs producing scintillations were not generated at the longitudes near São Luiz, where the ionosonde measurements were used as input parameters. Previous studies have shown that there could be extremely large differences in the generation rates of EPBs at closely located longitudes (Li et al., 2016). Whereas the input parameter hmF2 from the São Luiz ionosonde measurements during sunset hours could represent the variation of PRE over a large longitude region, some local processes responsible for the initial seeding of EPBs, for example, the F region bottomside wave structure in the neighboring longitudes, are difficult to be caught by the ionosonde measurements. On the other hand, previous studies showed that if there is a strong seeding source, EPBs could be generated even without PRE (Tsunoda et al., 2010). In a more recent study, Takahashi et al., (2020) observed a possible source of EPB seeding by 2,500 km south away from the equator. Under such conditions, the model may not predict well the occurrence of the EPBs and associated scintillation, only with the input parameters from the ionosonde measurements.

For the FP condition where strong scintillations above the threshold were not observed but were predicted by the model, one possibility is that the EPBs did not produce strong scintillations in there. It is evident from Figures 9a and 9c that for the FP conditions, scintillations were observed but with relatively weak intensity. In fact,

Table 2
The Accuracy, AUC, and F1 of Different Models

Model metric	D-XGB	T-XGB	D-LGB	T-LGB	D-Cat	T-Cat
Accuracy (%)	92.84	92.51	92.94	93.40	93.18	**93.73**
AUC	0.9841	0.9845	0.9833	0.9853	0.9852	**0.9860**
F1	0.8222	0.8156	0.8226	0.8329	0.8290	**0.8407**

Note. The bold font in Table 2 highlights are the best results for different metrics among the models.



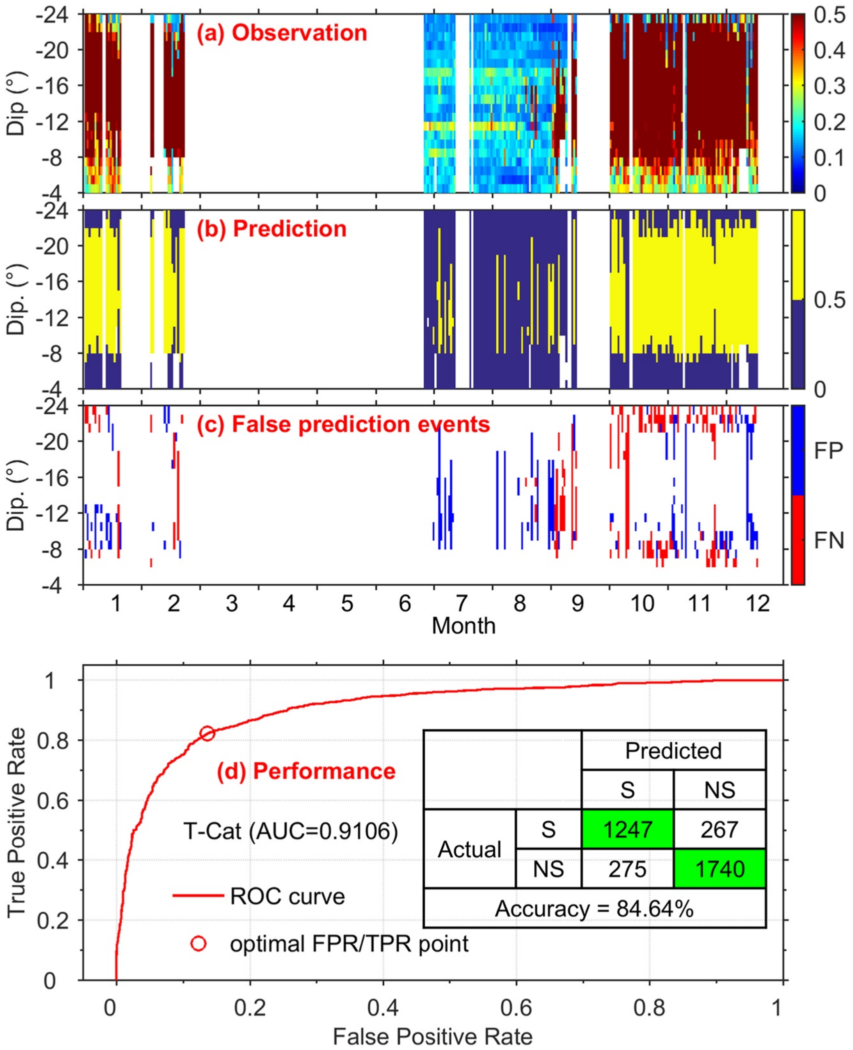
		Predicted	
		S	NS
Actual	S	2415	119
	NS	796	11269
Accuracy = 93.73%			

Figure 7. (a) The performance of T-Cat for the S_4 threshold 0.5. The circle superimposed on the receiver operating characteristic curve is the optimal point and its corresponding confusion matrices are shown in the embedded chart. (b) The prediction accuracy of T-Cat as a function of S_4 threshold.

the ionospheric scintillation is produced by plasma irregularities of scales sizes of hundreds of meters, generated through secondary instability processes within the EPB depletion structure (Haerendel, 1974). The scintillation intensity is dependent on the strength of plasma density fluctuation (ΔN) when the satellite-receiver radio links traverse through the irregularity structure embedded in EPBs. Even if the EPB depletion structures extend to low latitude, they may not produce serious scintillations under relatively low background ionospheric density condition. For the present scintillation prediction on a daily basis, except the foF2 from the equatorial ionosonde, no plasma density measurements at other latitudes were included as input parameters. By including the background ionospheric density measurements at different latitudes, and the location and the intensity of the equatorial ionization anomaly in the daily forecast, the prediction accuracy of strong scintillations could be improved.

5. Conclusions

A prediction model of low latitude ionospheric strong scintillation has been built based on the gradient boosting algorithms using long-term ground-based GPS receiver and Digisonde data over the Brazilian longitude sector. It is for the first time that the algorithms XGBoost, LightGBM, and CatBoost are utilized to build prediction models for regional occurrence of strong scintillations. Generally, the XGBoost, LightGBM, and CatBoost with default and tuned hyperparameters can forecast the scintillation with a good performance. Based on the long-term observational data during 2012–2019, the relative importance of different input parameters causing strong scintillations was investigated. The results revealed that the model trained with the equatorial hmF2 may be extended to other parameters in predicting the low latitude strong scintillations. By using the equatorial ionosonde hmF2 and foF2 measurements as input parameters of the trained model, which was trained with the data from 2012 to 2013 and 2015 to 2019, the performance for predicting the scintillation occurrence on a daily basis was examined in 2014 (solar maximum) and 2020 (solar minimum). The prediction accuracy is ~85%. It was suggested that the trained model with the input of equatorial hmF2 and foF2 before sunset could well predict the occurrence/non-occurrence of low latitude scintillations on a daily basis. Because this study mainly focuses on predicting strong



		Predicted	
		S	NS
Actual	S	1247	267
	NS	275	1740
Accuracy = 84.64%			

Figure 8. A comparison of (a) observations and (b) prediction results on a daily basis using T-Cat, (c) the false positive and false negative events, (d) the performance of T-Cat in 2014. The circle superimposed on the receiver operating characteristic curve is the optimal point and its corresponding confusion matrices are shown in the embedded chart. The blank area in the top two panels represent data gap.

scintillations characterized by the threshold 0.5, the prediction of weaker scintillations with the current model and input parameters may deteriorate.

Whereas some efforts have been made in the prediction of ionospheric strong scintillations, there are still some discrepancies between the forecasted results and observations. The possibilities responsible for the discrepancies, including the drifting EPB which may produce scintillation but the factors affecting its generation in the neighboring longitudes are not included to train the model, and the background plasma density which may affect the strength of plasma density fluctuation and thus the scintillation intensity, have been discussed. In the future, we hope to train and test the model for predicting the strength of ionospheric scintillation in the East and Southeast Asia sector, where a lot of facilities including big radars, regional dense GNSS TEC and scintillation receivers, and ionosonde oblique and vertical sounding networks have been deployed at equatorial and low latitudes in a wide longitude region.

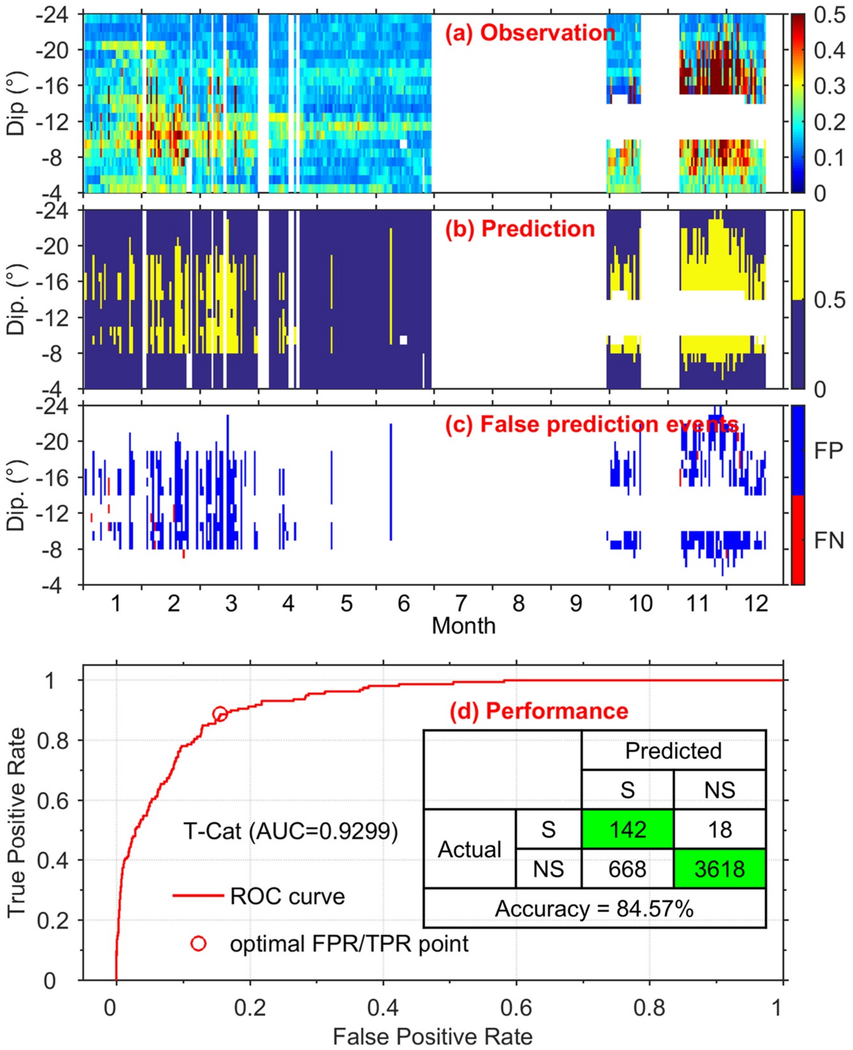


Figure 9. Same as Figure 8 but for 2020.

Data Availability Statement

The GPS scintillation and TEC data were obtained from the LISN (http://lisn.igp.gob.pe/jdata/database/, registration required). The Digisonde data was obtained from the Instituto Nacional de Pesquisas Espaciais, São José dos Campos, Brazil (http://www2.inpe.br/climaespacial/SpaceWeatherDataShare/). The K_p index was obtained from the GFZ German Research Centre for Geosciences (http://www.gfz-potsdam.de/en/kp-index/). The F10.7 data was obtained from the Canada.ca (https://www.spaceweather.gc.ca/solarflux/sx-3-en.php). The IMF data was obtained from the OMNI (https://omniweb.gsfc.nasa.gov/ow_min.html). All the data used in this study can be accessed at the WDC for Geophysics, Beijing (https://doi.org/10.12197/2021GA020).

Acknowledgments

This work was supported by the National Natural Science Foundation of China (42020104002, 41727803, 42074190, and 41904141), the Project of Stable Support for Youth Team in Basic Research Field, CAS (YSBR-018), the International Partnership Program of Chinese Academy of Sciences (Grant No. 183311KYSB20200003), the Solar-Terrestrial Environment Research Network (STERN) of the Chinese Academy of Sciences, and the Chinese Meridian Project. The authors acknowledge Prof. M. A. Abdu and one anonymous reviewer for the very helpful comments and suggestions that greatly improved the quality of the paper.

References

Abdu, M. A. (2001). Outstanding problems in the equatorial ionosphere thermosphere electrodynamics relevant to spread F. *Journal of Atmospheric and Solar-Terrestrial Physics*, *63*(9), 869–884. https://doi.org/10.1016/S1364-6826(00)00201-7

Abdu, M. A., Batista, I. S., Reinisch, B. W., de Souza, J. R., Sobral, J. H. A., Pedersen, T. R., et al. (2009). Conjugate Point Equatorial Experiment (COPEX) campaign in Brazil: Electrodynamics highlights on spread F development conditions and day-to-day variability. *Journal of Geophysical Research*, *114*, A04308. https://doi.org/10.1029/2008JA013749

Abdu, M. A., Sobral, J. H. A., Nelson, O. R., & Batista, I. S. (1985). Solar cycle related range type spread-F occurrence characteristics over equatorial and low latitude stations in Brazil. *Journal of Atmospheric and Terrestrial Physics*, *47*(8), 901–905. https://doi.org/10.1016/0021-9169(85)90065-0

Alfonsi, L., Spogli, L., De Franceschi, G., Romano, V., Aquino, M., Dodson, A., & Mitchell, C. N. (2011). Bipolar climatology of GPS ionospheric scintillation at solar minimum. *Radio Science*, *46*(3). https://doi.org/10.1029/2010RS004571

Alken, P., Thébault, E., Beggan, C. D., Amit, H., Aubert, J., Baerenzung, J., et al. (2021). International geomagnetic reference field: The thirteenth generation. *Earth Planets and Space*, *73*, 49. https://doi.org/10.1186/s40623-020-01288-x

Basu, S., Groves, K. M., Quinn, J. M., & Doherty, P. (1999). A comparison of TEC fluctuations and scintillations at Ascension Island. *Journal of Atmospheric and Solar-Terrestrial Physics*, *61*(16), 1219–1226. https://doi.org/10.1016/S1364-6826(99)00052-8

Bentéjac, C., Csörgő, A., & Martínez-Muñoz, G. (2020). A comparative analysis of gradient boosting algorithms. *Artificial Intelligence Review*, *54*, 1937–1967. https://doi.org/10.1007/s10462-020-09896-5

Camporeale, E. (2019). The challenge of machine learning in space weather: Nowcasting and forecasting. *Space Weather*, *17*(8), 1166–1207. https://doi.org/10.1029/2018SW002061

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Paper presented at the *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. https://doi.org/10.1145/2939672.2939785

de Lima, G. R. T., Stephany, S., de Paula, E. R., Batista, I. S., & Abdu, M. A. (2015). Prediction of the level of ionospheric scintillation at equatorial latitudes in Brazil using a neural network. *Space Weather*, *13*(8), 446–457. https://doi.org/10.1002/2015SW001182

Fejer, B. G., Scherliess, L., & de Paula, E. R. (1999). Effects of the vertical plasma drift velocity on the generation and evolution of equatorial spread F. *Journal of Geophysical Research*, *104*(A9), 19859–19869. https://doi.org/10.1029/1999JA900271

Grzesiak, M., Cesaroni, C., Spogli, L., De Franceschi, G., & Romano, V. (2018). Regional short-term forecasting of ionospheric TEC and scintillation. *Radio Science*, *53*(10), 1254–1268. https://doi.org/10.1029/2017RS006310

Haerendel, G. (1974). *Theory of equatorial spread F*. Max-Planck Inst. für Extraterr. Phys.

Jiao, Y., Hall, J. J., & Morton, Y. T. (2017). Automatic equatorial GPS amplitude scintillation detection using a machine learning algorithm. *IEEE Transactions on Aerospace and Electronic Systems*, *53*(1), 405–418. https://doi.org/10.1109/TAES.2017.2650758

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., et al. (2017). LightGBM: A highly efficient gradient boosting decision tree. Paper presented at the *Proceedings of the 31st International Conference on Neural Information Processing Systems*.

Kelley, M. C. (2009). *The Earth's Ionosphere: Plasma Physics and Electrodynamics* (Vol. 43). Academic Press.

Kil, H., Kintner, P. M., de Paula, E. R., & Kantor, I. J. (2002). Latitudinal variations of scintillation activity and zonal plasma drifts in South America. *Radio Science*, *37*(1), 6-1–6-7. https://doi.org/10.1029/2001RS002468

Li, G., Ning, B., Liu, L., Ren, Z., Lei, J., & Su, S. Y. (2007). The correlation of longitudinal/seasonal variations of evening equatorial pre-reversal drift and of plasma bubbles. *Annals of Geophysics*, *25*(12), 2571–2578. https://doi.org/10.5194/angeo-25-2571-2007

Li, G., Ning, B., Otsuka, Y., Abdu, M. A., Abadi, P., Liu, Z., et al. (2021). Challenges to Equatorial plasma bubble and ionospheric scintillation short-term forecasting and future aspects in east and southeast Asia. *Surveys in Geophysics*, *42*(1), 201–238. https://doi.org/10.1007/s10712-020-09613-5

Li, G., Otsuka, Y., Ning, B., Abdu, M. A., Yamamoto, M., Wan, W., et al. (2016). Enhanced ionospheric plasma bubble generation in more active ITCZ. *Geophysical Research Letters*, *43*(6), 2389–2395. https://doi.org/10.1002/2016GL068145

Liang, W., Luo, S., Zhao, G., & Wu, H. (2020). Predicting hard rock pillar stability using GBDT, XGBoost, and LightGBM algorithms. *Mathematics*, *8*(5), 765. https://doi.org/10.3390/math8050765

Manju, G., Sreeja, V., Ravindran, S., & Thampi, S. V. (2011). Toward prediction of L band scintillations in the equatorial ionization anomaly region. *Journal of Geophysical Research*, *116*(A2). https://doi.org/10.1029/2010JA015893

Materassi, M., Alfonsi, L., Spogli, L., & Forte, B. (2020). Chapter 18 – Scintillation modeling. In M. Materassi, B. Forte, A. J. Coster, & S. Skone (Eds.), *The dynamical ionosphere* (pp. 277–299). Elsevier. https://doi.org/10.1016/b978-0-12-814782-5.00018-2

Nugent, L. D., Elvidge, S., & Angling, M. J. (2021). Comparison of low-latitude ionospheric scintillation forecasting techniques using a physics-based model. *Space Weather*, *19*, e2020SW002462. https://doi.org/10.1029/2020SW002462

Otsuka, Y., Shiokawa, K., Ogawa, T., & Wilkinson, P. (2002). Geomagnetic conjugate observations of equatorial airglow depletions. *Geophysical Research Letters*, *29*(15), 43-41–43-44. https://doi.org/10.1029/2002GL015347

Priyadarshi, S. (2015). A review of ionospheric scintillation models. *Surveys in Geophysics*, *36*(2), 295–324. https://doi.org/10.1007/s10712-015-9319-1

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2017). CatBoost: Unbiased boosting with categorical features. *Advances in Neural Information Processing Systems*, *31*.

Retterer, J. M. (2010). Forecasting low-latitude radio scintillation with 3-D ionospheric plume models: 2. Scintillation calculation. *Journal of Geophysical Research*, *115*(A3). https://doi.org/10.1029/2008JA013840

Retterer, J. M., Decker, D. T., Borer, W. S., Daniell, R. E., Jr, & Fejer, B. G. (2005). Assimilative modeling of the equatorial ionosphere for scintillation forecasting: Modeling with vertical drifts. *Journal of Geophysical Research*, *110*(A11). https://doi.org/10.1029/2002JA009613

Rezende, L. F. C., de Paula, E. R., Stephany, S., Kantor, I. J., Muella, M. T. A. H., de Siqueira, P. M., & Correa, K. S. (2010). Survey and prediction of the ionospheric scintillation using data mining techniques. *Space Weather*, *8*(6). https://doi.org/10.1029/2009SW000532

Sousasantos, J., Kherani, E. A., & Sobral, J. H. A. (2017). An alternative possibility to equatorial plasma bubble forecasting through mathematical modeling and Digisonde data. *Journal of Geophysical Research: Space Physics*, *122*(2), 2079–2088. https://doi.org/10.1002/2016JA023241

Spogli, L., Alfonsi, L., Cilliers, P. J., Correia, E., De Franceschi, G., Mitchell, C. N., et al. (2013). GPS scintillations and total electron content climatology in the southern low, middle and high latitude regions. *Annals of Geophysics*, *56*(2). https://doi.org/10.4401/ag-6240

Spogli, L., Alfonsi, L., Romano, V., De Franceschi, G., Joao Francisco, G. M., Shimabukuro, M. H., et al. (2013). Assessing the GNSS scintillation climate over Brazil under increasing solar activity. *Journal of Atmospheric and Solar-Terrestrial Physics*, *105–106*, 199–206. https://doi.org/10.1016/j.jastp.2013.10.003

Takahashi, H., Wrasse, C. M., Figueiredo, C. A. O. B., Barros, D., Paulino, I., Essien, P., et al. (2020). Equatorial plasma bubble occurrence under propagation of MSTID and MLT gravity waves. *Journal of Geophysical Research: Space Physics*, *125*, e2019JA027566. https://doi.org/10.1029/2019JA027566

Takahashi, H., Wrasse, C. M., Otsuka, Y., Ivo, A., Gomes, V., Paulino, I., et al. (2015). Plasma bubble monitoring by TEC map and 630 nm airglow image. *Journal of Atmospheric and Solar-Terrestrial Physics*, *130–131*, 151–158. https://doi.org/10.1016/j.jastp.2015.06.003

Tama, B. A., & Rhee, K.-H. (2019). An in-depth experimental study of anomaly detection using gradient boosted machine. *Neural Computing & Applications*, *31*(4), 955–965. https://doi.org/10.1007/s00521-017-3128-z

Tsunoda, R. T., Bubenik, D. M., Thampi, S. V., & Yamamoto, M. (2010). On large-scale wave structure and equatorial spread F without a post-sunset rise of the F layer. *Geophysical Research Letters*, *37*(7). https://doi.org/10.1029/2009GL042357

Valladares, C. E., & Chau, J. L. (2012). The low-latitude ionosphere sensor network: Initial results. *Radio Science*, *47*(4). https://doi.org/10.1029/2011RS004978

Xiong, C., Stolle, C., & Lühr, H. (2016). The Swarm satellite loss of GPS signal and its relation to ionospheric plasma irregularities. *Space Weather*, *14*(8), 563–577. https://doi.org/10.1002/2016SW001439

Zhao, X., Ning, B., Liu, L., & Song, G. (2014). A prediction model of short-term ionospheric foF2 based on AdaBoost. *Advances in Space Research*, *53*(3), 387–394. https://doi.org/10.1016/j.asr.2013.12.001