Shidler and Rodrigues
Earth, Planets and Space (2020) 72:102
https://doi.org/10.1186/s40623-020-01227-w

Earth, Planets and Space

FULL PAPER Open Access

Modeling equatorial ionospheric vertical plasma drifts using machine learning

S. A. Shidler* and F. S. Rodrigues

Abstract

We present the results of an effort to model quiet-time vertical plasma drifts in the low-latitude *F*-region ionosphere using the random forest machine learning technique. The model is capable of describing the climatological variation of the drifts as a function of universal time, day of the year, solar flux, and altitude (200–600 km). The model has been trained using measurements of the vertical plasma drifts made by the incoherent scatter radar of the Jicamarca Radio Observatory (11.95° S, 76.87° W, ∼ 1° dip lat). In our analysis, we compare our machine learning model results with the Scherliess and Fejer (J Geophys Res 104:6829–6842, 1999) model (SF99 model), a widely used empirical model of the vertical drifts developed using a different set of Jicamarca measurements. We find that the machine learning model is able to capture the overall features of the diurnal variation of the equatorial drifts for different seasonal and solar flux conditions. The model is also capable of capturing the mean height variation of the drifts, particularly the height gradient enhancements that have been observed near sunrise and sunset. Finally, the model can easily be expanded and improved as more drift measurements are made and become available for training.

Keywords: Ionosphere, Drifts, Equatorial, Model, Machine learning, Random forest

Introduction

The zonal component of the ionospheric electric field in the magnetic equatorial region plays an important role in the dynamics of the geospace environment with implications for space weather. For instance, this electric field is one of the main drivers of ionospheric plasma transport at low- and mid-latitude regions (e.g., Klobuchar et al. 1991). During daytime, the zonal component of the equatorial ionospheric electric field is typically eastward, which creates upward $E \times B$ drifts of the ionospheric plasma (Scherliess and Fejer 1999). These upward drifts lift the equatorial ionospheric plasma to higher altitudes. Then, pressure and gravitational forces will cause the plasma to diffuse poleward along magnetic field lines. This "fountain" effect results in ionization peaks at higher magnetic latitudes, a phenomenon referred to as the equatorial ionization anomaly (EIA) or Appleton anomaly (Schunk and Nagy 2009).

*Correspondence: sas141430@utdallas.edu
W. B. Hanson Center for Space Sciences, University of Texas at Dallas, Richardson, TX, USA

The zonal electric field is also one of the main drivers of ionospheric plasma instabilities leading to the development of severe ionospheric irregularities at low latitudes (e.g., Fejer et al. 1999; Abdu 2001; Smith et al. 2015). For instance, the linear growth rate of the ionospheric Generalized Rayleigh–Taylor (GRT) instability is proportional to the magnitude of the vertical drifts (Sultan 1996). Near sunset, when drifts reverse from upward to downward, a pre-reversal enhancement (PRE) of the drifts is commonly observed (Eccles et al. 2005) producing favorable conditions for the GRT instability and for the development of ionospheric irregularities. These irregularities are capable of disrupting radio-based systems used for communications, navigation, and remote sensing (Basu et al. 1998; Carrano et al. 2012; Kintner et al. 2007).

Therefore, a description of the equatorial vertical drifts is useful when trying to understand or even predict ionospheric behavior. A climatological description of the drifts, in particular, is useful when trying to understand the average behavior of phenomena observed in the low- and mid-latitude ionosphere. Scherliess and Fejer (1999) developed an empirical, climatological global model of

quiet-time equatorial vertical plasma drifts which is commonly used in ionospheric studies. The model, hereby referred to as the SF99 model, was developed using measurements made by the Jicamarca incoherent scatter radar (ISR) between 1968 and 1992. It was expanded to longitudes outside the Peruvian sector using measurements made by in situ sensors on the Atmospheric Explorer-E (AE-E) satellite between 1977 and 1979.

To develop their model, Scherliess and Fejer (1999) considered the variability of quiet-time drifts with respect to local time, day of the year, longitude, and solar flux. The SF99 model uses univariate normalized cubic-B splines of order 4 to describe the local time and longitudinal variability of the vertical drifts, as well as a linear dependence to describe the variation of the vertical drifts with solar flux. The SF99 model uses a simple linear interpolation scheme to transition between seasons, and coefficients that represent the best-fit model to the data were determined using a least-squares procedure. Finally, the SF99 model provides a height-averaged estimate of the drifts.

In addition to the SF99 model, other empirical models of the vertical drifts have been derived using artificial neural networks applied to magnetometer measurements (e.g., Anderson et al. 2004; Anghel et al. 2007; Dubazane and Habarulema 2018; Chaitanya and Patra 2020). These models relate measurements made by latitudinally spaced magnetometers to vertical drift measurements made by an independent sensor. The independent sensor is, in most cases, a coherent scatter radar system capable of estimating vertical drifts from the so-called 150-km echoes (e.g., Anderson et al. 2004; Anghel et al. 2007; Chaitanya and Patra 2020). The main advantage of such an approach is that an expensive, high power ISR system is not needed to provide the data for a model. The 150-km echoes and drifts derived from them, however, are limited in range (around 150 km altitude) and to daytime hours (e.g., Kudeki and Fawcett 1993; Chau and Woodman 2004). Dubazane and Habarulema (2018) have also used in situ drift measurements made by sensors on the C/NOFS Low-Earth-Orbit satellite (de La Beaujardiére 2004) over the African sector. While satellite measurements can provide nighttime measurements, there are limitations associated with the orbit (e.g., limited passes over site, different altitudes, etc.). Finally, the observations available did not allow these studies to address the height variability of the drifts.

Motivated by the increasing use of machine learning in various fields of study, we present here the results of the application of the random forest technique to the empirical modeling of quiet-time equatorial ionospheric vertical drifts. Here, we opted for a technique that is significantly less computationally expensive for training and optimization than neural networks. This choice is encouraged by the potential of improving the model as new observations and data sources become available.

To develop this model we use measurements made by the Jicamarca ISR between 1996 and 2018. Advances in Jicamarca's radar capabilities since the development of the SF99 model allowed us to take into consideration the height variability of the drifts in our model. It has been shown that the height variability can be significant near sunrise and sunset (e.g., Pingree and Fejer 1987; Fejer et al. 2014; Shidler et al. 2019; Shidler and Rodrigues 2019). Therefore, the model takes a step further and describes the quiet-time behavior of vertical drifts as a function of universal time, day of the year and solar flux as well as altitude.

In the following section, we provide a description of the radar measurements used in this study, and our selection of usable observations. We also present a brief description of the random forest technique. In "Results and discussion" section, we present and discuss the results of the random forest technique to modeling of ionospheric drift data, including overall model performance, comparisons with the SF99 model, case studies, examining the predicted height variation of the vertical drifts, and potential adjustments in model parameters. The conclusion summarizes our main results and provides concluding remarks.

Measurements and model

We now describe the dataset used in our development of an empirical model of vertical drifts using machine learning. We also describe how higher quality data were filtered, and how measurements considered to have been made under geomagnetically quiet-time conditions were selected. Finally, we describe the random forest technique, which was used to develop the model.

Measurements: Jicamarca radar drifts

The dataset used for the development of our model consists of measurements of *F*-region vertical plasma drifts made by the incoherent scatter radar (ISR) of the Jicamarca Radio Observatory—JRO (11.95°S, 76.67°W, $\sim 1°$ dip angle) between 1996 and 2018. The observations are publicly available in the Madrigal database.

The Jicamarca ISR is capable of providing semi-routine (10–45 days/year) measurements of the vertical drifts as a function of local time and height. The range of altitudes covered by the Jicamarca radar goes from about 200 km and can extend beyond 600 km during high solar flux conditions. Typical height resolution is about 15 km

and time resolution is approximately 5 min (Kudeki et al. 1999).

The main objective of this study is to develop a model of the vertical drifts during geomagnetically quiet conditions. To create this model, we devoted efforts to select only adequate measurements of the highest quality. First, we selected only measurements that were made under geomagnetically quiet conditions. The selection of quiet-time observations follows the approach described by Fejer and Scherliess (2001) where observations considered quiet were only those preceded by 6 h of AE indices below 300 nT. Given the availability of data and our objective to ensure, to the best of our ability, that only quiet-time observations were used, we extended the requirement of the AE index below 300 nT to 12 h. That is, for each measurement we found the hourly AE index at the time of the measurement plus the previous 11 hourly AE values. Observations where none of the 12 AE values exceeded 300 nT were assumed to be made under geomagnetically quiet conditions and were used in our development of the quiet-time drifts. Requiring more than 12 h of low AE index would be impractical for most types of datasets including the Jicamarca drifts. It would severely reduce the number of observations. In addition, we have also removed measurements occurring during both minor and major sudden stratospheric warming (SSW) events, which occur primarily in January and February (e.g., Chau et al. 2009, 2012).

Next, we inspected the measurements and removed potential outliers, that is, drift measurements that were not derived from purely incoherent scatter echoes. The outliers are caused, in most part, by irregularities associated with equatorial spread *F* (ESF), interference from artificial radio sources, and echoes from low Earth orbit (LEO) satellites. These echoes cause abnormal increases in the signal-to-noise ratio (SNR) height profiles and abrupt variations in the vertical drift height profiles.

ESF irregularities typically occur in the post-sunset sector (Sultan 1996). Recent studies show, however, high occurrences rates of ESF irregularities in the post-midnight sector during low solar flux conditions (e.g., Ajith et al. 2015; Zhan et al. 2018). Measurements that were contaminated by the presence of ESF (or satellite echoes) were detected and removed based on large signal-to-noise ratios (SNR). More specifically, measurements with SNR values greater than 1 dB were removed from the dataset.

In addition, we run the data through a final filter to remove low quality or unrealistic measurements. In the equatorial *F*-region ionosphere, the vertical plasma drifts vary with height by a few meters per second per 100 km (Shidler and Rodrigues 2019). Measurements where the drifts varied by more than 5 m/s from one range gate to the next (typically 15 km) were considered unrealistic and were removed from the dataset.

Finally, extreme outliers that were not detected using the above filtering process were removed using the interquartile range ($1.5 \times IQR$) rule (Baron 2013).

Model: random forest

The random forest is an ensemble machine learning technique that consists of several individual decision trees, each of which contributes to the final model prediction, and can be used for either supervised regression or classification problems (Breiman 1996). In this study, we used the random forest, which is part of the Scikit-learn Python package (https://scikit-learn.org), to perform supervised regression on Jicamarca ISR measurements of the vertical plasma drifts.

Scikit-learn's Random Forest Regressor machine learning algorithm uses binary decision trees that are constructed using the Classification and Regression Trees (CART) algorithm. Each binary decision tree consists of several internal nodes with the data in each node being split into one of two child nodes based on a splitting criterion. In our model, splits are chosen so as to best minimize the variance (or the mean square error—MSE) of the data in the two child nodes. More specifically, the weighted average of the MSE for the child nodes is calculated for all available split points, and the split point that minimizes this value is chosen. This process continues iteratively until a stopping criterion is met at which point the nodes are referred to as "leaves". The average of the data in each leaf is then a possible output of the decision tree.

A drawback of using individual trees is that they exhibit high variance. That is, the output of a decision tree is sensitive to changes in the training dataset used. Random forests attempt to minimize this variance by averaging the output over several individual trees that are randomly constructed (Biau 2012). There are two sources of randomness associated with a random forest: (1) each decision tree is trained using a bootstrap re-sample of the data, and (2) constructing the trees relies on selecting split points from a random sub-sample of the input features.

A number of factors led us to choose this machine learning technique. First, the ensemble random forest is computationally inexpensive to train and can be done in parallel. Second, the large number of decision trees used in the model helps to improve performance and minimize over-fitting (Breiman 2001). Another factor considered was the relative ease of optimizing the hyperparameters for model performance. The hyperparameters we chose to tune included (a) the number of trees in the ensemble, (b) the maximum

number of input features to consider when looking for the best node split, (c) the minimum number of samples required for a node split and (d) the minimum number of samples required in the leaves after a node split. Scikit-learn's default values were used for all other parameters.

The first parameter we optimized was the number of decision trees used in the random forest. Theoretically, the root-mean-square error (RMSE) is a monotonically decreasing function of the number of trees in the random forest (Probst and Boulesteix 2018), and models will only improve by adding more trees. Finding the optimal number of trees is therefore done by analyzing the trade-offs between improvements in model performance with additional trees and the amount of computational resources available. In this study, we use 250 trees in our random forest. This number was large enough that adding additional trees resulted in negligible improvements in model performance using the RMSE as the metric, but did not make the computational run-times for training and cross-validation prohibitively long. We then tuned the parameter that determines the number of input features to consider when looking for the best split. In the ensemble package used, we can specify whether to use all input features or a randomly selected sub-sample of the input features when considering node splits. We found the best performance when a maximum of half of the input features were considered. The optimal numbers for parameters (c) and (d) were optimized by doing an exhaustive grid-search using Scikit-learn's GridSearchCV method using five fold cross-validation. For our final settings we used 250 decision trees in the random forest, a minimum of 29 samples required for a node to split, and a minimum of 10 samples required in each leaf of the decision tree.

The model takes the day of the year (DOY), altitude, universal time (UT), and solar flux ($F_{10.7}$ index) as input features with the equatorial vertical plasma drifts as the single output. These features were chosen as to allow comparisons between our machine learning model of the vertical drifts and the well-tested, widely used SF99 model. We point out that the two models were developed using different datasets. It also should be noted that the SF99 model only used observations from altitudes where SNR of the ISR echoes were the highest, which was typically between 300 and 400 km. Therefore, the SF99 does not include altitude as an input feature.

Results and discussion

A total of 402 days of measurements made between 1996 and 2018 were available to this study. We must point out, however, that only a few hours of observations were made in some of these days. For our analyses, we randomly selected 20 days of measurements from the original set of observations to serve as a testing subset for model validation. The remaining 382 days of measurements were shuffled and further subdivided into 20 unique subsets. We then created 20 random forest models with each model using one of the unique subsets as a validation set and the remaining 19 subsets as a training subset. The random forest model that minimized the RMSE of the validation subset was used in our analysis. The number of measured drift values in the testing subset corresponds to about 9% of the total number of measurements available.

Overall model performance

The overall performance of our model is presented in Fig. 1. It shows the distribution of model predicted values versus measurements, for the training (green) and testing (red) subsets. We also used the RMSE and the coefficient of determination (R^2) to evaluate the performance of our model results. As expected, the performance of the model for the training subset exceeds the performance for the testing subset. The model produced results with an RMSE of 2.85 m/s and $R^2 = 0.97$ for the training subset and results with an RMSE of 8.60 m/s and $R^2 = 0.71$ for the testing subset.

One of the sources for the large RMSE of the testing subset is the intrinsic day-to-day variability present in the ionosphere (Fejer et al. 1989). While the model can capture the overall behavior the drifts, it does not capture the short-term variations in the drifts associated with,

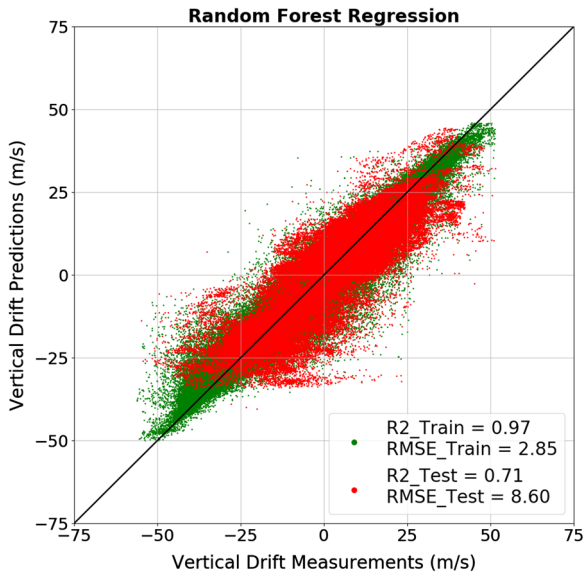

Fig. 1 Random Forest Regression; Overall performance of our random forest model for drifts in the training subset (green) and testing subset (red)

for instance, small-scale density (conductivity) structures, atmospheric gravity waves, and changes in tidal and planetary wave forcing (Fejer and Scherliess 1997). In addition, it is possible that our 12-h AE index filter did not completely eliminate all the drifts' measurements affected by disturbance electric fields of magnetospheric origin. Disturbed dynamo electric fields can occur more than 12 h after the observed high latitude disturbances (Fejer et al. 1991).

A benefit of the ensemble random forest technique is that it provides information on the importance of the input parameters. Figure 2 summarizes the importance of each input (feature) for model predictions. The feature importance is a normalized factor indicating how much a feature contributes to model output. Scikit-Learn quantifies feature importance by examining the layer depth of a feature used at an internal node. Features used at nodes near the top layers of the tree will contribute to the final prediction of a larger fraction of the input samples than nodes near the bottom of the tree, and will therefore have a greater relative importance. Additional details about the feature importance are provided by Louppe (2015). The results indicate that universal time (UT) is the dominant feature for predicting the drifts. This is expected given that vertical drifts have a strong diurnal variation. The results also show that solar flux ($F_{10.7}$) and day of year (DOY) have similar levels of importance. The overall diurnal behavior of the drifts does not vary much with solar flux and season. However, some features of the drifts such as the pre-reversal enhancement (PRE) near sunset are strongly controlled by season (Abdu et. al 1981; Tsunoda 1985) and solar flux conditions (Scherliess and Fejer 1999; Smith et al. 2015; Shidler et al. 2019).

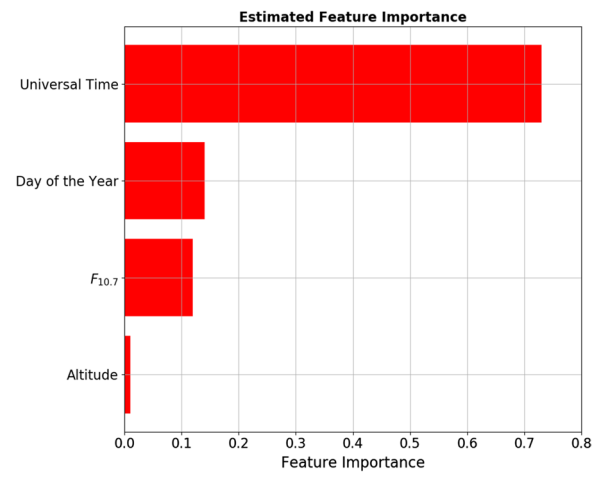

Fig. 2 Estimated Feature Importance; Estimated importance of each input feature

The least important feature in the model is altitude. This is most likely a result of the weak height variation of the drifts within main *F*-region heights during most times. Previous studies, however, have pointed out that significant height variations occur near the terminators (e.g., Pingree and Fejer 1987; Shidler and Rodrigues 2019). The random forest model, nevertheless, is still able to detect and model the behavior of the drifts with height. We will return to the height variability of the drifts later in this discussion.

Case studies

We now turn our attention to a more focused inspection of our model results with respect to the training and testing subset, and a comparison of our model results with those of the SF99 model.

Training subset

Figure 3 shows drift values predicted by our machine learning model (red markers) versus measurements (black markers) for 20 days chosen randomly from our training subset. The year, DOY and F10.7 for each day are indicated on the top of each panel. These drift values correspond to an altitude of 360 km. The SF99 model prediction is also shown for comparison (solid green line).

These example cases serve to show that our model captures well the overall behavior of the drifts used in the training. The results indicate, in particular, that our model does not overfit on days with increased local time variability. See, for instance, day 31 and day 188 of 2010. There is a relatively large fluctuation in the drift values observed between 0400 and 0600 LT. These fluctuations are caused by very low plasma densities over Jicamarca and that are probed by the ISR in the late night and pre-sunrise sector at times. The reduced plasma density causes low SNR echoes, large uncertainties in the measurements and the observed variability in the drifts. The relatively large residuals between predictions and actual measurements during these hours contribute to the spread of the training subset in Fig. 1.

We can also see from Fig. 3 that the SF99 captures important features of the drifts. As expected the SF99 does not perfectly match the observations since it is a climatological model and cannot (is not intended to) capture the quiet-time day-to-day variability of the drifts.

Testing subset

We now compare our model predictions to the subset of testing measurements and to SF99 model predictions.

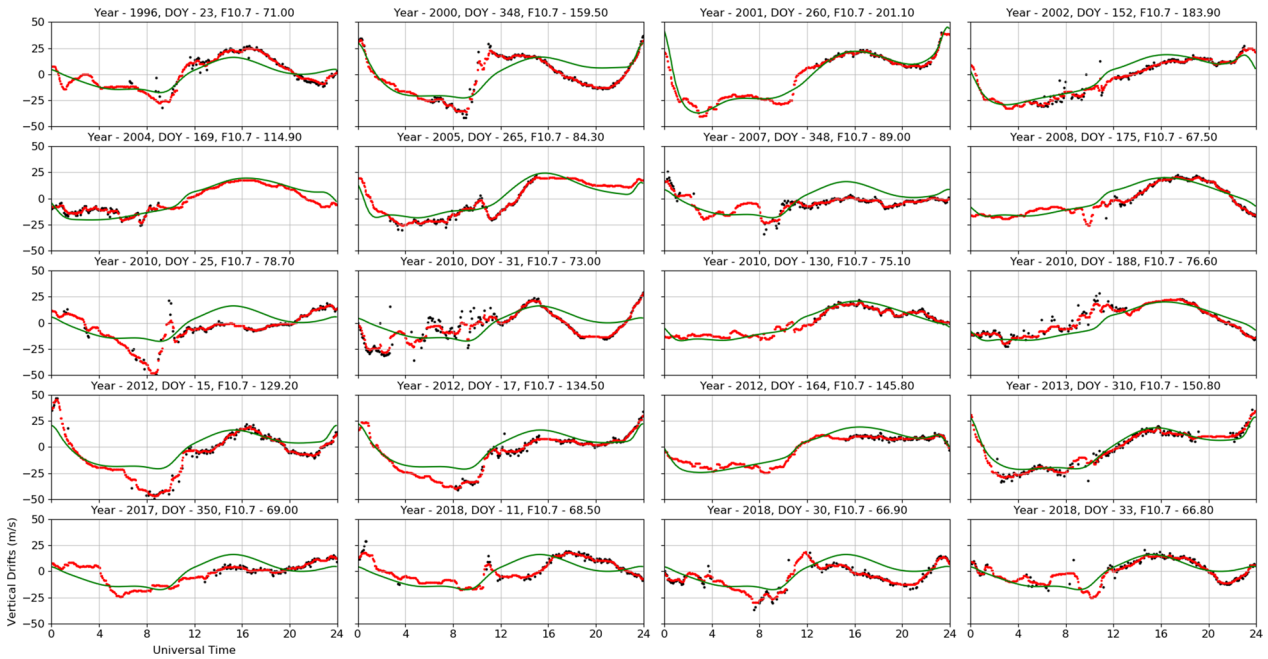

Fig. 3 Local Time Plots of Measured and Predicted Vertical Drifts for Training Subset; Scatter plots of measured vertical drifts (black) and predicted vertical drifts (red) corresponding to an altitude of 360 km for 20 days chosen randomly from our training dataset. The green line shows the SF99 model predictions

Figure 4 shows the comparison for 20 days that were randomly selected from the original set of Jicamarca measurements and were not used in the training of our model. Our model results and the measurements are for an altitude of 360 km. The results show that our model is able to capture the expected diurnal variation of the drifts,

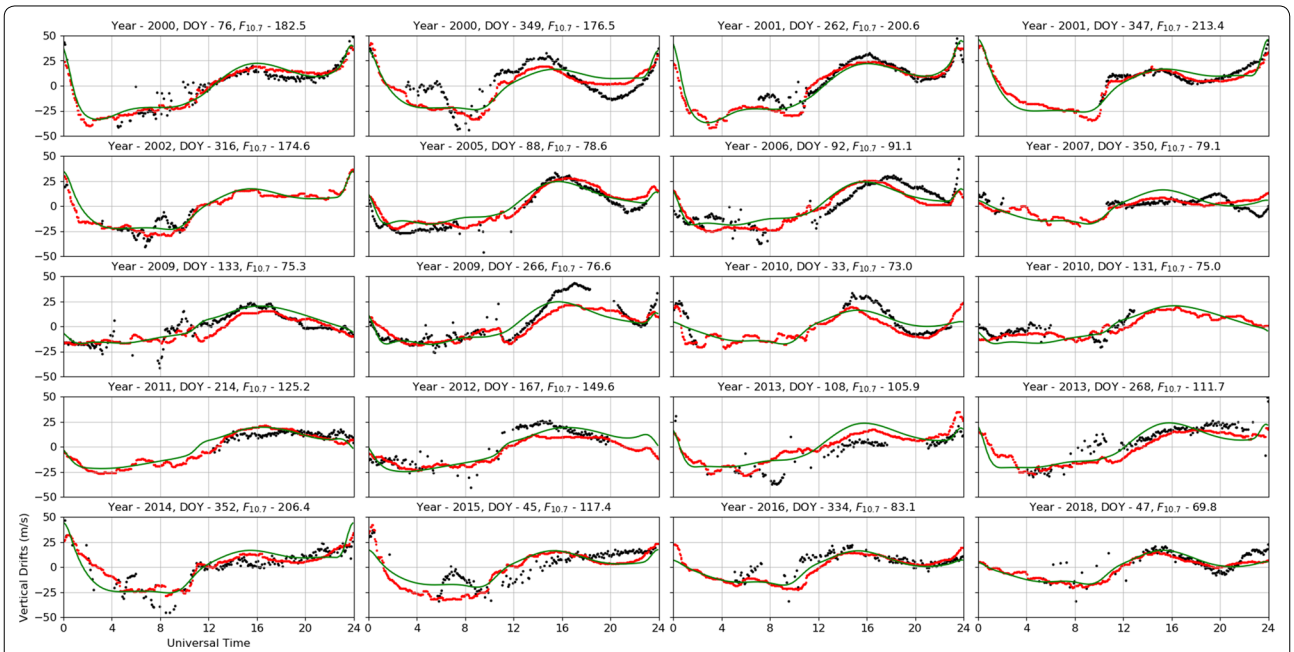

Fig. 4 Local Time Plots of Measured and Predicted Vertical Drifts for Testing Subset; Scatter plots of measured vertical drifts (black) and predicted verticals (red) corresponding to an altitude of 360 km for the 20 days of testing data. The green line shows the SF99 model predictions

with upward drifts during the day and downward drifts at night. Our model also predicts the occurrence of the PRE peak fairly well. We point out that the horizontal axis in Fig. 4 is in universal time, and the PRE typically occurs around 1900 LT, which corresponds to 2400 UT.

The comparison of our model with measurements and with the SF99 model shows that, in general, our model predicts the same overall behavior. In some cases, the behavior predicted by both models is virtually the same. For instance, see curves for DOY 76 of 2000, DOY 334 of 2016 and DOY 47 of 2018. There are a few cases, however, where the machine learning model seems to predict an unusual behavior of the drifts that is not predicted by the SF99. For instance, our model is capable of predicting the abrupt change in the drifts observed near sunrise (1000 UT) on DOY 347 of 2001. It is also able to predict the weak drifts observed between 1100 and 1700 UT on DOY 350 of 2007. As expected, we find larger errors in the testing predictions during pre-sunrise hours due to the reduced accuracy of the ISR drift measurements during this time.

To better evaluate the performance of our model with respect to the SF99, we computed the RMSE for both model predictions with respect to the testing subset. Figure 5 shows the results of our evaluation. It shows scatter plots of the measurements (actual) versus modeled predictions for the SF99 model (green) and for our machine learning model (red) for three different altitudes: 250, 360, and 500 km. We found that the machine learning model outperformed the SF99 model at each altitude but only slightly. We must point out that this comparison takes into consideration measurements made at all local times. While the drifts do not vary much with height during most local times, the variation might be important near sunrise and sunset. This is discussed further in the following section.

Height variation

Our machine learning modeling approach takes into consideration the height variation of the vertical drifts. While this variation is weak during most hours, especially if average values are considered, it can be significant near sunrise and sunset (Shidler and Rodrigues 2019).

To examine the height variation predicted by the machine learning model, we ran predictions for a full year with the solar flux set to 150 SFU. Then, for each universal (or local) time we estimated the height variation by fitting a linear model to the drift profiles for altitudes between 200 and 500 km. The results are presented in Fig. 6. The black curve represents the mean height gradient of the drifts for each local time. The error bars represent the variability (standard variation) of the drift values for each local time. Our model is able to predict the expected behavior of the height variation of the vertical drifts throughout the day. It shows that the gradients are mostly positive before ~1200 LT, and mostly negative after that time which were also found in previous studies and observations (Pingree and Fejer 1987; Chau and Woodman 2004; Rodrigues et al. 2015; Shidler et al. 2019; Shidler and Rodrigues 2019). In addition, the model predicts the enhancements in height gradients for the vertical drifts near sunrise and sunset that have been found in previous analyses of Jicamarca drifts carried out by Shidler and Rodrigues (2019).

For comparison, the red curve shows the results of Shidler and Rodrigues (2019). The curve represents the local time variation of the mean height gradients

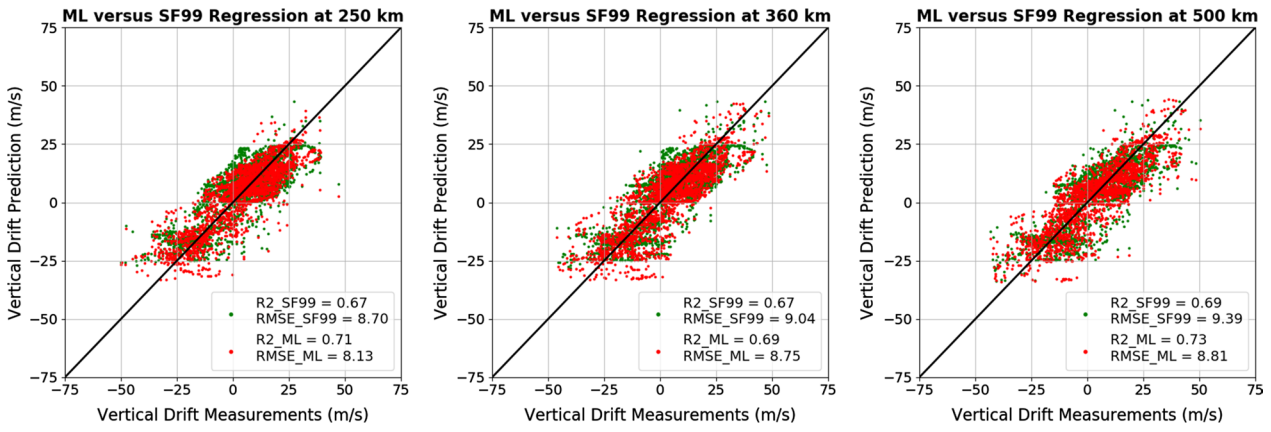

Fig. 5 Regression Plots Comparing SF99 Versus Our Machine Learning Model; Scatter plots for actual measurements versus predictions using the SF99 model (green) and machine learning model (red) at 250 km (left), 360 km (middle), and 500 km (right). Each panel shows the coefficient of determination (R^2) and the RMSE values for the machine learning and SF99 models

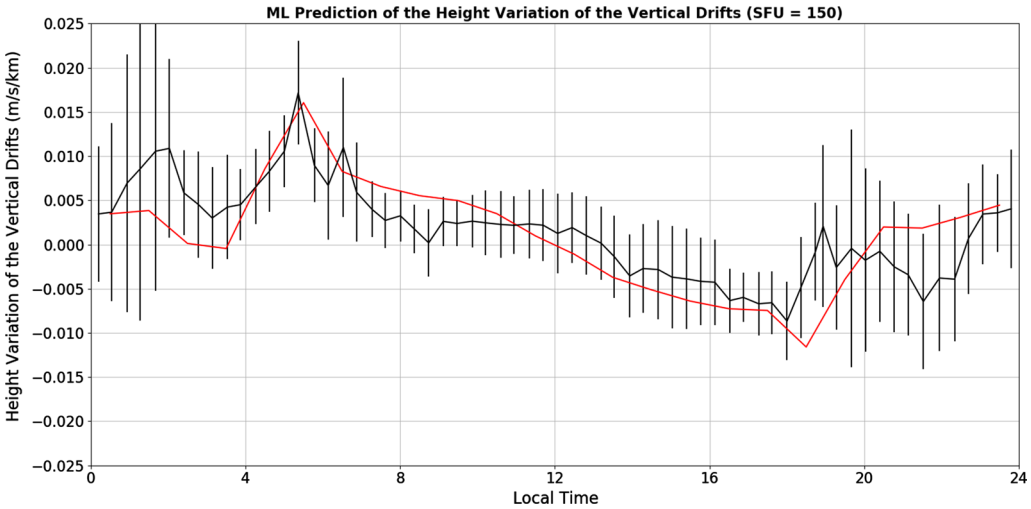

Fig. 6 Predicted Height Gradients of Vertical Drifts for 150 SFU; Mean and standard deviation of the predicted height variation of the vertical drifts over the period of one year for $F_{10.7} = 150$ SFU

of vertical drifts. These mean height gradients were obtained averaging individual height gradients estimated from drift profiles measured by the Jicamarca radar. One can see the positive (negative) gradients before (after) local noon, and the enhancements near the terminators. We must point out that differences between the observed mean gradients and the gradients derived from our model are most likely caused by differences in how the quantities were computed. For instance, the observed mean gradients (red curve) were estimated from observed drift profiles made over a wide range of solar flux conditions (mean F10.7 = 150 SFU and values varying from 105.4 to 277.4 SFU) and non-uniform distribution of days throughout the year. The gradients derived from our model were for a fixed F10.7 of 150 SFU and for every day of an entire year. In addition, the observed mean gradients were computed for 1-h wide bins. The model gradients were computed for specific times (i.e. no local time binning).

On the minimum samples per leaf

As mentioned earlier, some of the hyperparameters were tuned using Scikit-learn's GridSearchCV method. The best performing hyperparameters were determined by the lowest average RMSE using fivefold cross-validation. However, when examining the model output we found some excessive variations in the model drift curves, particularly in the pre-sunrise sector. These variations are caused by, in most part, by the larger variability in the measured drifts around that time. This larger variability is a result of the low SNR of the observed echoes and large uncertainties in the measured drifts used to train the model.

We anticipate that excessive variations in the drift curves might not be adequate when using them to drive physics-based numerical models of the ionosphere such as SAMI2 (Huba et al. 2000). We found that it is possible to smooth out the local time prediction of the vertical drifts by increasing the minimum number of samples required in each leaf of a decision tree.

Figure 7 shows the machine learning predictions for the days in the testing subset using a model trained with a minimum of 10 samples required per leaf (red) and a model trained with a minimum of 150 samples required per leaf (blue). The actual measurements for the testing subset are shown in black. Increasing the minimum samples per leaf has little impact on the daytime drifts but can substantially smooth out model predictions in the early morning.

We must note that this is to illustrate that smooth predictions can be obtained for certain applications. The choice for the minimum samples per leaf to be used is a result of trial-and-error, and increasing this parameter resulted in a larger RMSE for the testing subset.

Conclusions

We present the results of an effort to model quiet-time vertical plasma drifts in the low-latitude *F*-region ionosphere using the random forest machine learning technique. The model is capable of describing the climatological variation of the drifts as function of universal time, day of the year, solar flux, and altitude (200–600 km).

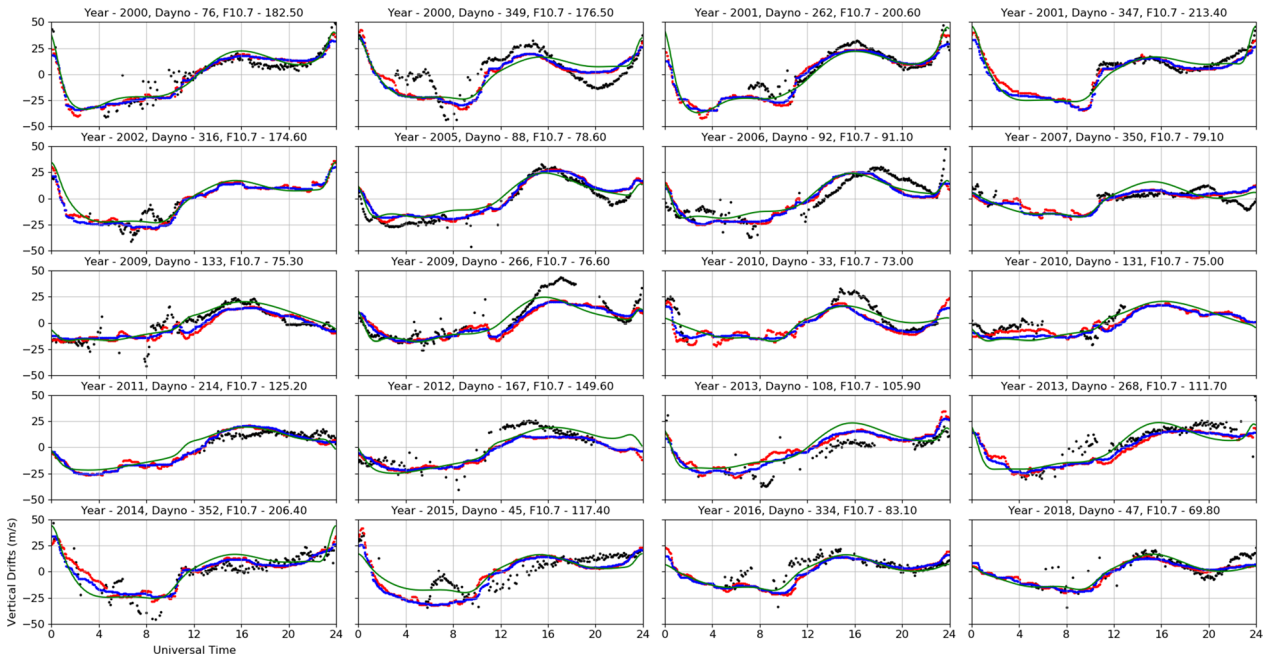

Fig. 7 Effects of Modifying Min Samples Per Leaf Hyperparameter on Model Predictions; Comparison of machine learning predictions using a minimum of 10 samples per leaf (Red) and a minimum of 150 samples per leaf (blue). The actual measurements are shown in black with SF99 model predictions in green. Increasing the minimum number of samples per leaf has the effect of smoothing

The model has been trained using 382 days of measurements of the vertical plasma drifts made by the incoherent scatter radar of the Jicamarca Radio Observatory (11.95°S, 76.87° W, ∼ 1° dip lat) between 1996 and 2018. In our analysis, we compare our machine learning model results with the Scherliess and Fejer (1999) model (SF99 model), a widely used empirical model of the vertical drifts developed using a different set of Jicamarca measurements. Both models were tested on a dataset comprising 20 days of observations of the vertical drifts from Jicamarca that were not used in the training of either model.

We found that the machine learning model can describe the overall behavior of the drifts with a slightly smaller root mean square error (RMSE) than the SF99 model. In addition, the model is capable of capturing the diurnal variation of the gradients including the gradient enhancements near sunrise and sunset, which is in good agreement with physical expectations and with previous studies and observations. Finally, the model can easily be expanded and improved as more drift measurements are made and become available for training. One can envision, for instance, expanding the model to include disturbance electric fields with the time history of the AE index as an additional input.

Abbreviations

AE: The Auroral Electrojet index is used to quantify geomagnetic activity; CART : Classification and Regression Trees algorithm used to train the decision trees; DOY: Day of the year; EIA: Equatorial ionization anomaly; ESF: Equatorial spread F; F_{107}: The F_{107} index is used to quantify solar activity; GRT: Generalized Rayleigh–Taylor instability; ISR: Incoherent scatter radar; MSE: Mean squared error; PRE: Pre-reversal enhancement of the vertical plasma drifts; RMSE: Root mean squared error; SF99: Scherliess and Fejer (1999) empirical model of the equatorial vertical plasma drifts; SFU: Solar flux unit ($1\ SFU = 10^{-22}\ Wm^{-2}Hz^{-1}$); SNR: Signal-to-noise ratio of the incoherent scatter radar measurement; SSW: Sudden stratospheric warming; UT: Universal time.

Acknowledgements

The authors would like to thank Dr. David Lary from UT Dallas for useful comments related to the random forest algorithm.

Authors' contributions

SAS analyzed the data, produced the graphs, and made contributions to the interpretation of results and writing of this manuscript. FSR provided directions for this study, interpreted results, and contributed to the writing of this manuscript. Both authors read and approved the final manuscript.

Funding

Work at UT Dallas was supported by the NSF (AGS-1554926 and AGS-1916055). The Jicamarca Radio Observatory is a facility of the Instituto Geofisico del Peru operated with support from the NSF (AGS-1433968) through Cornell University.

Availability of data and materials

The datasets used in this study are available on the Madrigal Database, [http://jro.igp.gob.pe/madrigal/]. A copy of the random forest model is available for download at zenodo.org/record/3836595 (doi:10.5281/zenodo.3836596).

Competing interests

The author's declare they have no competing interests.

Received: 19 March 2020 Accepted: 30 June 2020
Published online: 14 July 2020

References

Abdu MA, Bittencourt JA, Batista IS (1981) Magnetic declination control of the equatorial F region dynamo electric field development and spread F. J Geophys Res. 86(A13):11443–11446. https://doi.org/10.1029/JA086iA13p11443

Abdu MA (2001) Outstanding problems in th equatorial ionosphere-thermosphere electrodynamics relevant to spread F. J Atmos Sol Terr Phy 63:869–884. https://doi.org/10.1016/S1364-6826(00)00201-7

Ajith KK, Tulasi Ram S, Yamamoto M, Yokoyama T, Sai Gowtam V, Otsuka Y, Tsugawa T, Niranjan K (2015) Explicit Characteristics of evolutionary-type plasma bubbles observed from equatorial atmosphere radar during the low to moderate solar activity years 2010–2012. J Geophys Res. 120:1371–1382. https://doi.org/10.1002/2014JA020878

Anderson D, Anghel A, Chau J, Veliz O (2004) Daytime vertical EXB drift velocities inferred from ground-based magnetometer observatoin at low latitudes. Space Weather 2:S11001. https://doi.org/10.1029/2004SW000095

Anghel A, Anderson D, Maruyama N, Chau J, Yumoto K, Bhattacharyya A, Alex S (2007) Interplanetary electric fields and their relationship to low-latitude electric fields under disturbed conditions. J Atmo Solar-Terr Phys. 69:1147–1159. https://doi.org/10.1016/j.jastp.2006.08.018

Baron M (2013) Probability and Statistics for Computer Scientists. CRC Press, Boca Raton

Basu S, MacKenzie E, Basu S (1988) Ionospheric constraints on VHF/UHF communications links during solar maximum and minimum periods. Radio Sci. 23(3):363–378. https://doi.org/10.1029/RS023i003p00363

Biau G (2012) Analysis of a random forests model. J Mach Learn Res. 13(1):1063–1095. https://doi.org/10.5555/2188385.2343682

Breiman L (1996) Bagging predictors. Mach Learn 24(2):123–140. https://doi.org/10.1007/BF00058655

Breiman L (2001) Random forests. Mach Learn 45:5–32. https://doi.org/10.1023/A:1010933404324

Carrano CS, Groves KM, Caton RG (2012) Simulating the impacts of ionospheric scintillation on L band SAR image formation. Radio Sci. RS0L20:20. https://doi.org/10.1029/2011RS004956

Chaitanya PP, Patra AK (2020) A neural network-based model for daytime vertical ExB drift in the Indian sector. J. Geophys. Res. https://doi.org/10.1029/2020JA027832

Chau JL, Woodman RF (2004) Daytime vertical and zonal velocities from 150-km echoes: their relevance to F-region dynamics. Geophys Res Lett 31:L17801. https://doi.org/10.1029/2004GL020800

Chau JL, Fejer BG, Goncharenko LP (2009) Quite variability of equatorial ExB drifts during a sudden stratospheric warming event. Geophys Res Lett 36:L05101. https://doi.org/10.1029/2008GL036785

Chau JL, Fejer BG, Goncharenko LP, Han-Li Liu (2012) Equatorial and low latitude ionospheric effects during sudden stratospheric warming events. Space Sci Rev. 168:385–417. https://doi.org/10.1007/s11214-011-9797-5

de La Beaujardiére O et al (2004) C/NOFS: a mission to forecast scintillations. J Atmos Sol Terr Phys. 66(17):1573. https://doi.org/10.1016/j.jastp.2004.07.030

Dubazana MB, Habarulema JB (2018) An empirical model of vertical plasma drift over the African sector. Space Weather 16:619–635. https://doi.org/10.1029/2018SW001820

Eccles JV, St Maurice JP, Schunk RW (2015) Mechanisms underlying the prereversal enhancement of the vertical plasma drift in the low-latitude ionosphere. J Geophys Res Space Phys 120:4950–4970. https://doi.org/10.1002/2014JA020664

Fejer BG, de Paula ER, Batista JS, Bonelli E, Woodman RF (1989) Equatorial F region vertical plasma drifts during solar maxima. J Geophys Res. 94(A9):12049–12054. https://doi.org/10.1029/JA094iA09p12049

Fejer BG, de Paula ER, Gonzalez SA, Woodman RF (1991) Average vertical and zonal F region plasma drifts over Jicamarca. J Geophys Res. 96(A8):13901–13906. https://doi.org/10.1029/91JA01171

Fejer BG, Scherliess L (1997) Empirical models of storm time equatorial zonal electric fields. J. Geophys. Res. 102(A11):24047–24056. https://doi.org/10.1029/97JA02164

Fejer BG, de Paula ER, Scherliess L (1999) Effects of the vertical plasma drift velocity on the generation and evolution of equatorial spread F. J Geophys Res. 104(A9):19859–19869. https://doi.org/10.1029/1999JA900271

Fejer BG, Scherliess L (2001) On the variability of equatorial F-region vertical plasma drifts. J Atmos Sol-Terr Phys. 63:893–897. https://doi.org/10.1016/S1364-6826(00)00198-X

Fejer BG, Hui D, Chau JL, Kudeki E (2014) Altitudinal dependence of evening equatorial F region vertical plasma drifts. J Geophys Res 119:5877–5890. https://doi.org/10.1002/2014JA019949

Huba JD, Joyce G, Fedder JA (2000) The formation of an electron hole in the topside equatorial ionosphere. Geophys Res Lett. 27(2):181–184. https://doi.org/10.1029/1999GL010735

Kudeki E, Fawcett CD (1993) High resolution observations of 150 km echoes at Jicamarca. Geophys Res Lett. 20:1987–1990. https://doi.org/10.1029/93GL01256

Klobuchar JA, Anderson DN, Doherty PH (1991) Model studies of the latitudinal extent of the equatorial anomaly during equinoctial conditions. Radio Sci. 26(4):1025–1047. https://doi.org/10.1029/91RS00799

Kintner PM, Ledvina BM, de Paula ER (2007) GPS and ionospheric scintillations. Space Weather 5:S09003. https://doi.org/10.1029/2006SW000260

Kudeki E, Bhattacharyya S, Woodman RF (1999) A new approach in incoherent scatter F region ExB drift measurements at Jicamarca. J Geophys Res 104(A12):28145–28162. https://doi.org/10.1029/1998JA900110

Louppe G (2015) Understanding Random Forests From Theory to Practice, arXiv:1407.7502v3

Pingree JE, Fejer BG (1987) On the height variation of equatorial F region vertical plasma drifts. J Geophys Res. 92(A5):4763–4766. https://doi.org/10.1029/JA092iA05p04763

Probst P, Boulesteix A (2018) To tune or not to tune the number of trees in random forest. J Mach Learn Res. https://doi.org/10.5555/3122009.3242038

Rodrigues FS, Smith JM, Milla M, Stoneback RA (2015) Daytime ionospheric equatorial vertical drifts during the 2008–2009 extreme solar minimum. J Geophys Res. 120:1452–1459. https://doi.org/10.1002/2014JA020478

Scherliess L, Fejer BG (1999) Radar and satellite global equatorial F region vertical drift model. J Geophys Res. 104:6829–6842. https://doi.org/10.1029/1999JA900025

Schunk RW, Nagy AF (2009) Ionospheres: Physics, Plasma Physics, and Chemistry, Cambridge Atmospheric and Space Science Series. Cambridge University Press, Cambridge

Smith JM, Rodrigues FS, Paula ER (2015) Radar and satellite investigations of equatorial evening vertical drifts and spread F. Ann Geophys 33:1403–1412. https://doi.org/10.5194/angeo-33-1403-2015

Shidler S, Rodrigues FS, Fejer BG, Milla MA (2019) Radar studies of height-dependent equatorial F region vertical and zonal plasma drifts. J Res Geophys. https://doi.org/10.1029/2019JA026476

Shidler S, Rodrigues FS (2019) On the magnitude and variability of height gradients in the equatorial F region vertical plasma drifts. J Res Geophys. https://doi.org/10.1029/2019JA026661

Sultan PJ (1996) Linear theory and modeling of the Rayleigh–Taylor instability leading to the occurrence of equatorial spread F. J Geophys Res. 101(A12):26875–26891. https://doi.org/10.1029/96JA00682

Tsunoda RT (1985) Control of the seasonal and longitudinal occurrence of equatorial scintillations by the longitudinal gradient in integrated E region Pedersen conductivity. J Geophys Res. 90(A1):447–456. https://doi.org/10.1029/JA090iA01p00447

Zhan W, Rodrigues FS, Milla MA (2018) On the genesis of postmidnight equatorial spread F: results for the American/Peruvian sector. Geophys Res Lett. 45:7354–7361. https://doi.org/10.1029/2018GL078822

Publisher's Note