

Space Weather



RESEARCH ARTICLE

10.1029/2020SW002639

Key Points:

- Data driven advanced machine learning methods applied to forecast the ionospheric total electron content 5 h ahead
- The random forest and Long-Short Term Memory methods are employed, the data sources are space measurements characterizing the solar-terrestrial environment
- Variable importance ranking showed that F10.7, Lyman alpha are top predictors agreeing well with the physics of ionospheric formation

Correspondence to:

G. K. Zewdie,
gzewdie3@gatech.edu;
phygbki21@gmail.com

Citation:

Zewdie, G. K., Valladares, C., Cohen, M. B., Lary, D. J., Ramani, D., & Tsidu, G. M. (2021). Data-driven forecasting of low-latitude ionospheric total electron content using the random forest and LSTM machine learning methods. *Space Weather*, 19, e2020SW002639. <https://doi.org/10.1029/2020SW002639>

Received 3 OCT 2020

Accepted 26 MAY 2021

Data-Driven Forecasting of Low-Latitude Ionospheric Total Electron Content Using the Random Forest and LSTM Machine Learning Methods

Gebreab K. Zewdie¹ , Cesar Valladares², Morris B. Cohen¹ , David J. Lary², Dhanya Ramani², and Gizaw M. Tsidu³

¹School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, USA, ²William B. Hanson Center for Space Sciences, University of Texas at Dallas, Richardson, TX, USA, ³Department of Earth and Environmental Sciences, Botswana International University of Science and Technology, Palapye, Botswana

Abstract In this research, we present data-driven forecasting of ionospheric total electron content (TEC) using the Long-Short Term Memory (LSTM) deep recurrent neural network method. The random forest machine learning method was used to perform a regression analysis and estimate the variable importance of the input parameters. The input data are obtained from satellite and ground based measurements characterizing the solar-terrestrial environment. We estimate the relative importance of 34 different parameters, including the solar flux, solar wind density, and speed the three components of interplanetary magnetic field, Lyman-alpha, the Kp, Dst, and Polar Cap (PC) indices. The TEC measurements are taken with 15-s cadence from an equatorial GPS station located at Bogota, Columbia (4.7110° N, 74.0721° W). The 2008–2017 data set, including the top five parameters estimated using the random forest, is used for training the machine learning models, and the 2018 data set is used for independent testing of the LSTM forecasting. The LSTM method as applied to forecast the TEC up to 5 h ahead, with 30-min cadence. The results indicate that very good forecasts with low root mean square (RMS) error (high correlation) can be made in the near future and the RMS errors increase as we forecast further into the future. The data sources are satellite and ground based measurements characterizing the solar-terrestrial environment.

Plain Language Summary Space weather affects satellite communications, precise military operations and can interfere with power grids on the ground. Physics-based space weather forecasting is extremely challenging due to the complicated nature of the physical drivers which can come from the Sun, the magnetosphere, the ionosphere, and the lower atmosphere. In this research, we used data-driven machine learning methods to forecast the ionospheric total electron content which provides the amount of ionization in the upper atmosphere and hence helps as a proxy to forecast other space weather phenomenon.

1. Introduction

The behavior of the Earth's ionosphere is the result of the interplay of several processes that include production from photo-ionization, loss by chemical reactions, the coupling of the plasma component with the thermosphere and the magnetosphere and transport processes due to electric fields and neutral wind dynamics (Zettergren & Semeter, 2012). These complex interconnected factors contribute to the ionosphere's high variability due to the influence of factors from solar, geomagnetic, and atmospheric sources (Davies, 1990; Kelly, 2012). These variabilities are often characterized by the spatio-temporal fluctuations of ionospheric total electron content (TEC), scintillation indices, electron density and various ionospheric chemical constituents.

The presence of large TEC in the ionosphere and perturbations and irregularities in the electron density impose adverse effect on communication and navigation signals (Basu et al., 2001; Kaplan & Hegarty, 2005; Kintner et al., 2001). The adverse effects of the ionosphere on GNSS and other communication systems are more pronounced during a geomagnetic storm as TEC exhibits strong perturbations (Ngwira et al., 2019). Hence, the total electron content can be seen as a good indicator of ionospheric activity in particular and adverse space weather events in general.

© 2021. The Authors.

This is an open access article under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

Consequently, estimating the present and future TEC in the ionosphere would contribute to understanding and mitigating these adverse space weather effects. Previous and current physics-based estimation of the TEC and other space weather parameters are usually complicated by the fact that the relative roles of the physical factors are different across different geographic regions and altitude ranges. Physical models are also not purely physical since some of the processes are parameterized based on simplifying assumptions and may not work globally and the empirical data may not be available for all geographic regions. Moreover, physics-based models for forecasting need observational data to be used as initial and/or boundary conditions at the model grids. Furthermore, the ionosphere is shaped by inputs from the Sun, the solar wind, the magnetosphere, the lower atmosphere and transport processes in the ionosphere itself. Often, we do not have a precise functional relationship between these parameters and TEC measurements or other space weather parameters. Most studies depend heavily on data assimilation techniques in modeling the ionospheric TEC (Bilitza, 2018; Hajj et al., 2004; Mandrake et al., 2005; Scherliess et al., 2009).

Machine learning methods are promising solutions for this kind of problems in which we do not either know the functional relationship between the input variables and the output parameter we want to estimate or the computational cost of the physical model is high. These methods are famous in learning from the data to extract the necessary information. They are particularly suitable for problems involving a suite of variables especially when linear techniques such as least squares regression is inadequate to describe a nonlinear system. Machine learning and deep learning methods are now quite popular in many industries and have achieved some impressive results (Camporeale, 2019).

Some machine learning methods have been applied successfully to ionospheric, magnetospheric and other space weather studies. Forecasting of geomagnetic indices such as the Kp and Dst (Tan et al., 2018; Wu & Lundstedt, 1996), coronal mass ejection propagation time (Bobra & Ilonidis, 2016), solar wind speed (Yang et al., 2018), relativistic electrons at geosynchronous orbits (Ling et al., 2010) have been achieved by applying machine learning methods. Other recent works by Bortnik et al. (2018), Z. Chen, et al. (2019), McGranaghan et al. (2018), and Gross and Cohen (2020) applied different machine learning methods for different space weather studies over the ionosphere, magnetosphere and the radiation belt. A non-linear regression analysis was used by Villalobos and Valladares (2020) to model TEC values over South and Central America. These authors found a non-linear Kp, solar flux, and day of year dependency for each pixel ($0.5^\circ \times 0.5^\circ$) and each 30-min TEC map obtained between 2008 and 2010. However, their numerical model presented differences as large as 30% of the TEC measurements. A comprehensive review of the application of machine learning for space weather nowcasting and forecasting is given by (Camporeale, 2019).

The problem of forecasting TEC is largely a time domain problem, in that the future evolution is dependent not only on the present state, but on the past history of various solar-terrestrial parameters. To address this need, we have chosen to investigate the Long-Short Term Memory (LSTM) deep recurrent neural network. LSTM method is a powerful and well-known branch of artificial neural networks famously known for solving time sequence data (Goodfellow et al., 2016). LSTMs have been recently applied for forecasting different space weather parameters. For example, Tan et al. (2018) applied the LSTM method to forecast the Geomagnetic Kp index using historical solar wind, interplanetary magnetic field and the Kp index itself as input. Wei et al. (2018) were able to achieve one day lead time forecasting of high-energy electron integral flux at geostationary orbit using the LSTM deep neural network machine learning method and the Kp, Ap, Dst, solar wind speed, magnetopause subsolar distance and the 2-MeV electron integral flux itself as input. LSTMs are particularly popular in speech recognition, solar power forecasting, traffic prediction and others (Gensler et al., 2016; Graves et al., 2013; Zhao et al., 2017). A brief introduction about the LSTM deep recurrent neural network is presented §2.2.

Other advanced machine learning methods have been applied to forecast ionospheric TEC. For example, Huang and Yuan (2014) employed radial basis function neural network improved by Gaussian mixture model to forecast 30 min TEC. Huang and Yuan (2014) used day of year, local time, previous TEC, its temporal and differential variation as input variables and found results with root-mean-square error less than 5 TECU ($1 \text{ TECU} = 10^{16} \text{ el/m}^2$). Habarulema et al. (2009) used solar geomagnetic indices, Day of year and hour number features in a simple neural network to model TEC. R. Chen, Wang, et al. (2019) employed deep neural networks to forecast global TEC maps. However, the application of machine learning and deep learning methods to study the ionosphere can be considered at its infancy state, especially at high and mid-latitudes,

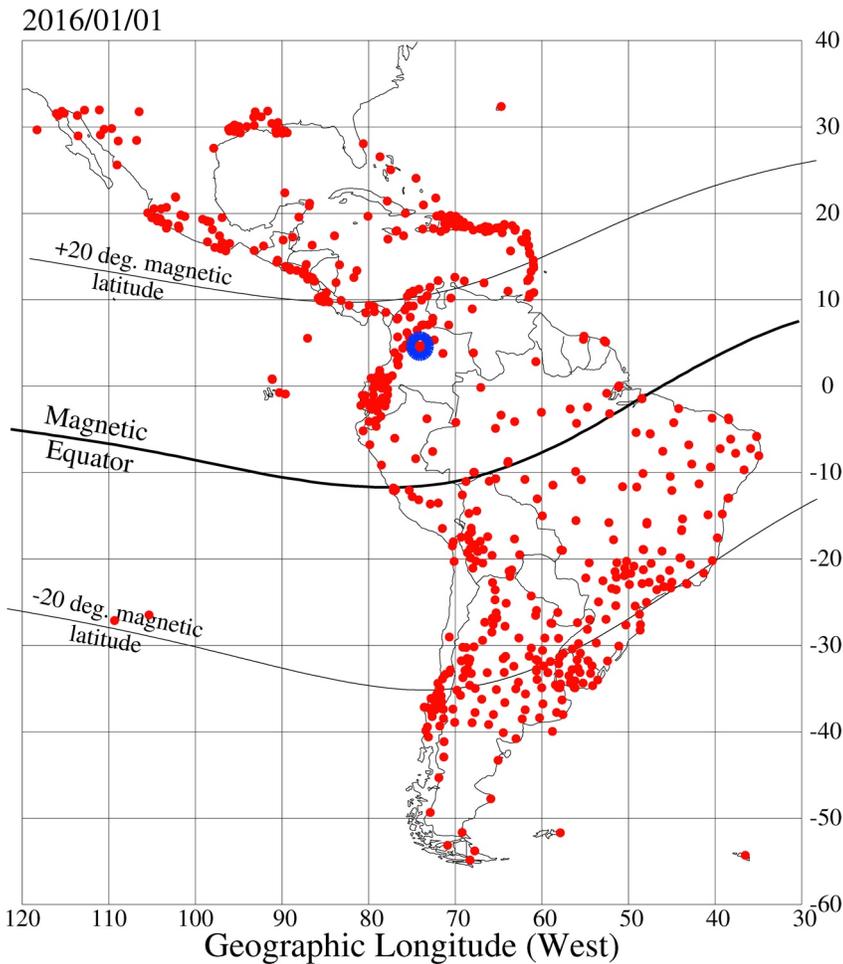


Figure 1. Showing the location of the Bogota, Columbia GPS station (blue circle) at the northern crest of the equatorial anomaly (4.7110° N, 74.0721° W). Red circles show other Low-latitude Ionospheric Sensor Network GPS stations as of 2016.

The TEC at this GPS station is calculated with 15-s cadence, collected for this study over one full cycle period from 2008 to 2018. We resample the TEC to a cadence of 30 min for efficient computational purposes.

The input features comprising the solar flux (F10.7), solar wind density and speed, the three components of interplanetary magnetic field, Lyman-alpha, the Kp, Dst and Polar Cap (PC) indices, the interplanetary magnetic field, a total of 34 parameters are extracted from NASA's OMNI (Operating Missions as a Node on the Internet) Space Physics Data Facility (SPDF). SPDF-OMNI contains multi-spacecraft measurements, account for the propagation estimated time from spacecraft to the magnetopause (McGranaghan et al., 2018). Table 1 presents the training features used in our machine learning methods. First, the random forest machine learning method is used to perform a preliminary regression analysis and variable importance estimation. Then the top 5 variables are selected and used into the LSTM deep recurrent neural network machine learning method to forecast TEC every 30 min for 5 h lead time.

Simple interpolation technique is applied to some of the parameters to match all the features to a 1 min time resolution before resampling the total data set to a 30 min cadence. All the data sets including the vertical TEC and the OMNI input features are cleaned for missing values. The presence of outliers in some features have been identified and removed as they can affect the machine learning output. For example, the solar flux has outlier measurements way greater than 140 in solar flux unit, other features such as the PC-index, Sigma Lat, Sigma Lon and other have outlier measurements which are removed before training our machine learning methods. Each feature and the v TEC has been scaled from 0 to 1 using the sklearn MIN-

given the ubiquitously available space and ground based data sets and the recent popularity of machine learning methods (McGranaghan et al., 2018).

In this research we use advanced machine learning methods and various space and ground based parameters and machine learning to forecast the ionospheric total electron content at a temporal resolution of half an hour and a lead time of up to 5 h.

The paper is organized in the following manner. Section 2 gives a succinct description of the machine learning methodology and the solar-terrestrial parameters that are associated with the variability of the TEC over the Bogota, Columbia equatorial station. Also in Section 2 the data processing and procedures of the experiment are described. Section 3 describes the results of the machine learning algorithm and the implication of these results. Sections 4 provides the discussion and conclusion part.

2. Data Analysis and Methodology

2.1. Data Analysis

We use TEC values collected by one of the Low-latitude Ionospheric Sensor Network (LISN) stations that has been operated since 2001 in Bogota, Columbia (location: 4.7110° N, 74.0721° W). This station is of prime importance due to its location under the northern crest of the equatorial anomaly (see Figure 1). Here, the plasma density is a strong function of the equatorial zonal electric field, the meridional component of the thermospheric wind, and several processes acting in other layers (e.g., stratosphere, mesosphere, and thermosphere) that influence or modify these quantities.

Table 1
Solar and Solar-Terrestrial Features Used to Train the Random Forest and LSTM Recurrent Neural Network Methods.

Feature name	Unit	Feature name	Unit
F10.7 solar flux	sfu ($10^{-22}\text{W m}^{-2}\text{ Hz}^{-1}$)	Sigma alpha/Prot.	ratio
Lyman alpha	$10^{-15}\text{ erg}^{-1}\text{ cm}^{-2}\text{ A}^{-1}$	Sigma_T	K
Polar Cap (PC) index	mV/m	SYM_H, ASY_D	nT
Dst	nT	Sigma IMF v_r	nT
Log. Angle of B	GSE	Sigma _{Np}	N/cm ³
R Sun spot	Number	V_x	m/s
AU index	nT	flow speed	m/s
SYM_D	nT	AP index	nT
IMF B magnitude	nT	sigma IMF Magnitude	nT
Lat. Angle of B	GSE	ASY_H	nT
AL index	nT	K_p	0–9
Sigma V	m/s	Sigma Lon	Degree
Sigma B_z	nT (GSE)	SW proton density	N/cm ³
Sigma B_y	nT (GSE)	Flow pressure	Pa
Sigma B_x	nT (GSE)	V_y	m/s
SW plasma temperature	K	V_z	m/s
Sigma IMF vector ave	nT	AE index	nT

Most of the features are obtained from NASA's Space Physics Data Facility (SPDF) OMNI (Operating Missions as a Node on the Internet) data.

MaxScaler (Pedregosa et al., 2011) before training the machine learning methods to minimize the impact of large dynamical range.

2.2. Machine Learning

Machine learning is a mathematical approach in which computer systems “learn by example” and extract useful information from a large set of historical data, often very large amounts of data spanning as large number of parameters as possible. Recently, machine learning has been applied to various fields in geosciences and remote sensing, agriculture, banking, etc (Camps-Valls, 2009; Lary et al., 2016; Zhang et al., 2016), and prediction of atmospheric gases such as CO₂ (Gardner & Dorling, 1998) and ozone (Prybutok et al., 2000; Yi & Prybutok, 1996). Beyond geosciences it is used very widely for applications such as for spam filtering (Guzella & Caminhas, 2009), credit scores, fraud detection, image processing, etc. The availability of large space and ground based data sets makes machine learning suitable for ionospheric study. The need for adequate computational facility can be satisfied with the advent of the now commonplace resources such as high performance computing (McGranaghan et al., 2018).

Machine learning methods can learn the behavior of the system and retrieve the necessary information if they are provided with data spanning as many parameters as possible in the training. It can “learn” the behavior of the system even in the case the relation between the information and the parameters is non-linear and multivariate (Lary et al., 2016).

Some commonly used machine learning approaches include Neural Networks, Support Vector Machines, decision trees, and Random Forests (an ensemble of decision trees). The other powerful method for time series forecasting is recurrent neural networks. Although there are different types of machine learning algorithms currently used, there is no single method that always will perform better than the rest for all problems. The best machine learning method to apply depends on the problem we solve and the available training data (Kotsiantis, 2007). The following subsections briefly describe the various machine learning approaches that we have employed in this research.

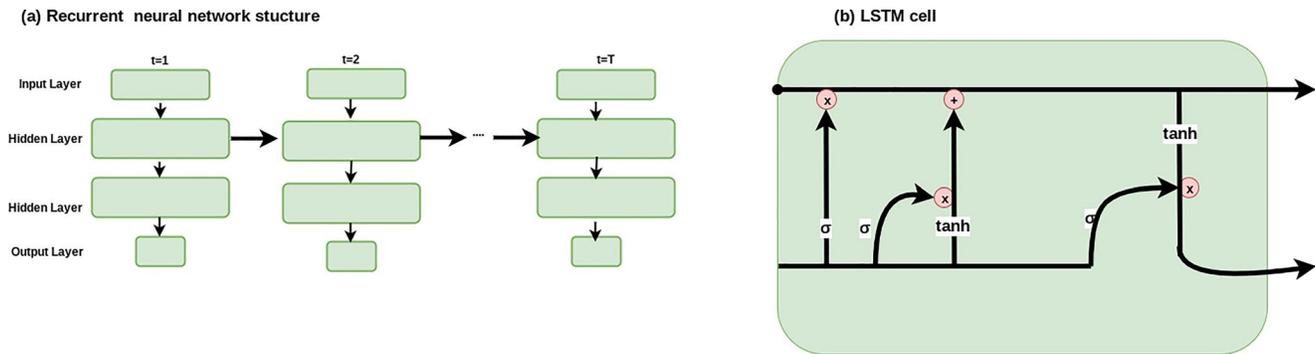


Figure 2. Showing network structure of recurrent neural networks and a single Long-Short Term Memory cell.

2.2.1. Random Forest

One of the popular machine learning methods known for its robust performance on regression and classification is the random forest method introduced by (Breiman, 2001). The random forest machine learning algorithm works based on random sampling of data to form ensemble of decision trees for both regression and classification problems. Each tree will provide its “vote” to make a decision in classification and a model functional estimate for regression. After a number of regression trees are grown using a randomly selected subset of training samples and variables, prediction will be made by aggregating (majority vote for classification or averaging for regression) the predictions of the ensemble. The basic idea here is ensemble learners from randomly resampled data set produce better model performance than developing a single regression tree from the total data set. In that way random forests decreases the variance of the model without increasing bias (Breiman, 2001; Friedman et al., 2001; Verikas et al., 2011).

The main advantage of random forests for the purpose of this study is that it provides a useful facility to rank the relative importance of the input variables (Genuer et al., 2010), allowing us to isolate the variables that are helpful for forecasting. The random forest estimates the variable importance of a feature by estimating the Gini impurity for classification and variance for regression when that feature is used as a splitting node for the regression tree (Han et al., 2016; Strobl et al., 2007). In this case, we are computing how much each feature contributes to decreasing the weighted impurity or variance. While a machine learning model such as LSTM can in principle learn which variables are important, the more refined is the data set, the less training data will be required to achieve good results. As such, narrowing down the variables to the important ones can shorten the training process for the LSTM.

2.2.2. Recurrent Neural Networks

Recurrent neural networks are a special type of neural networks known for time series forecasting and sequential analysis (Yu et al., 2019). Recurrent neural networks work on the principle of the cyclical connectivity and information flow of neurons in the human brain (Tan et al., 2018). They allow information to persist because they have loops across the hidden layers that connect the previous information to the present task as shown in Figure 2a. These cyclic connections present in recurrent neural networks make them more powerful than ordinary feed forward neural networks. Recurrent neural networks perform the same task for each component of the sequence with the output dependent on the previous computation of the sequence in the sense that the recurrent neural network has “memory” that encodes information for the previous computation.

The recently popular and robust deep recurrent neural network applied for solving scientific and commercial problems is the Long-Short Term Memory (LSTM) network. LSTMs, developed by Hochreiter and Schmidhuber (1997) and then refined by Graves (2012), are an improvement to recurrent neural networks in that LSTMs are known for removing the vanishing gradient problem (Hochreiter, 1998) that inhibits recurrent neural networks from learning. The vanishing gradient arises when the gradient of the loss function with respect to the weights is highly reduced during backpropagation time in a long sequence recurrent neural network (Bengio et al., 1994; Mikolov et al., 2014). LSTMs avoid the vanishing gradient problem of recurrent neural networks by remembering information for a long period of time through their special

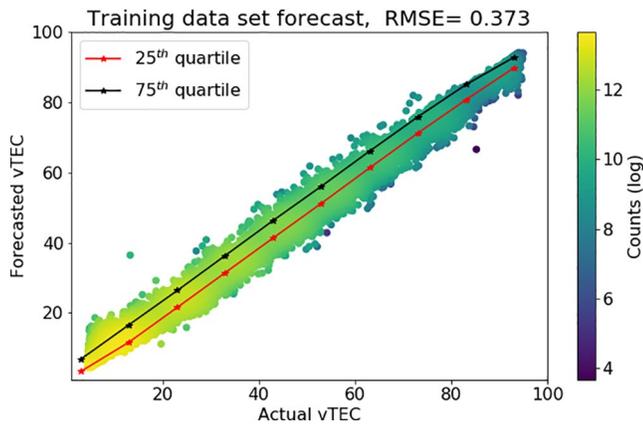


Figure 3. Scatter diagram showing the actual vTEC and estimated vTEC using the random forest machine learning method applied to the training data set.

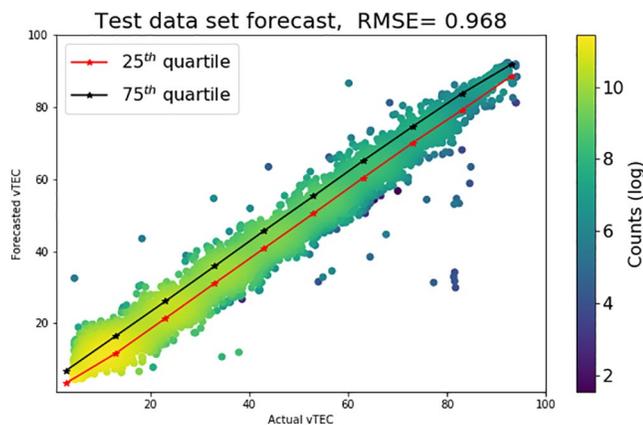


Figure 4. Scatter diagram showing the actual vTEC and estimated vTEC using the random forest machine learning method applied to the test data set.

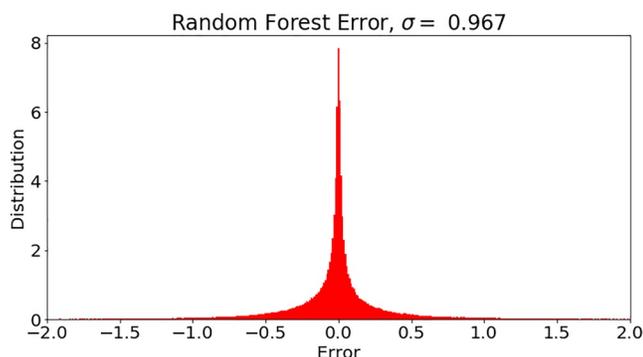


Figure 5. The error distribution of the vertical total electron content estimated using the random forest regression for the test data set.

four interacting structures in each repeating unit as shown in Figure 2b. These four structures are the cell state (the top horizontal line in Figure 2b) which is the memory part of the LSTM unit and the other three structures are input, output and forget gates that control the flow of information through. Explicit explanation about LSTM structures can be found in Yu et al. (2019).

3. Results

We first employed the random forest machine learning method to estimate the functional relationship between the various space weather parameters shown in Table 1 to the ionospheric vTEC measured at Bogota Columbia using the 2008 to 2013 data. The random forest machine learning method is also used to estimate and rank the contribution of each parameter (Figure 6) for the vTEC regression. The top five parameters namely: the F10.7 solar flux, Lyman alpha, ASY_D, PC index, and Dst as well as the vTEC itself are used to train the LSTM model.

The LSTM recurrent neural network is then employed to forecast the vTEC every 30 min for up to 5 h lead time. The LSTM model is initialized with random weights and the efficient ADAM (Adaptive Moment Estimation) optimizer (Kingma & Ba, 2014) is used. The loss function used is the root mean squared error. The LSTM model is run on top of tensorflow (Abadi et al., 2016) using the keras (Gulli & Pal, 2017) neural network library. The number of training epochs are set to 200; initial hyperparameter selection showed that after about the 200th epoch the root mean square error doesn't decrease significantly.

Figures 3 and 4 show scatter plots of the result of the random forest method applied to estimate the vertical TEC. The color scale in both figures represent the bin counts in logarithmic scale and shows the distribution of the actual and forecasted vTEC. To depict the distribution of the vTEC for each bin, the 25th and 75th quartile plots are shown by, respectively, the red and black lines. The data is split into 20% test set and 80% training set. The random forest model is developed using the training data and forecasts are made for the training set (Figure 3) and test set (Figure 4). Figure 3 shows scatter plots of forecasts made using the training data and the target vTEC used to supervise the model. Similarly, Figure 4 presents scatter plots of the actual vTEC withheld for testing and forecasts made using the test set. It is no surprise that in the Figure 3 the scatter plots are less spread away from the diagonal (RMSE = 0.373) than the test data set as they are result of forecasts using the same training data set that is used to develop the model. The scatter plots in Figure 4 are more spread away from the diagonal (RMSE = 0.968). The bin color scale shows that most of the TEC fall below about 25 TEC units.

Error distribution of estimates of the random forest forecast for the independent test data is depicted in Figure 5. In this case the error is the difference between the forecasted vTEC based on the independent test data and the actual vTEC withheld before training the random forest method. We clearly observe that the error is distributed with mean centered near zero and standard deviation close to 1 TEC unit showing robust performance of the random forest method.

The random forest can be applied to estimate the contribution of each feature for the model. This method contributes significantly by compli-

menting model interpretation and explanation issues that are common in other methods such as the black-box nature of the neural network (Tzeng & Ma, 2005) method. The random forest calculates the feature importance based on the calculation of the Gini impurity (Han et al., 2016) described in §2.2.1. Other methods such as estimation of the mean squared error in the out-of-bag sample for all number of trees before and after the values of that parameter are permuted (Genuer et al., 2010) can also be applied to estimate the variable importance in the random forest regression and classification problems.

Figure 6 presents the variable importance estimated and ordered in their respective rank for our TEC forecasting. The names and units of the features are listed in Table 1. Clearly we observe that solar flux (F10.7) and Lyman alpha are the top parameters. Other features such as the ASY_D, Polar Cap index and D_{st}, Log Log Angle of B, R Sun spot, and AU_index are among the top parameters. The features such as the solar wind plasma speed, temperature and AE_index are the lowest important features. We can see that a whole chunk of features do not contribute significantly to the model. However, the solar parameters such as, the solar flux (F10.7), Lyman alpha, R Sun spot and D_{st} are among the top predictors consistent with the physics of ionospheric formation.

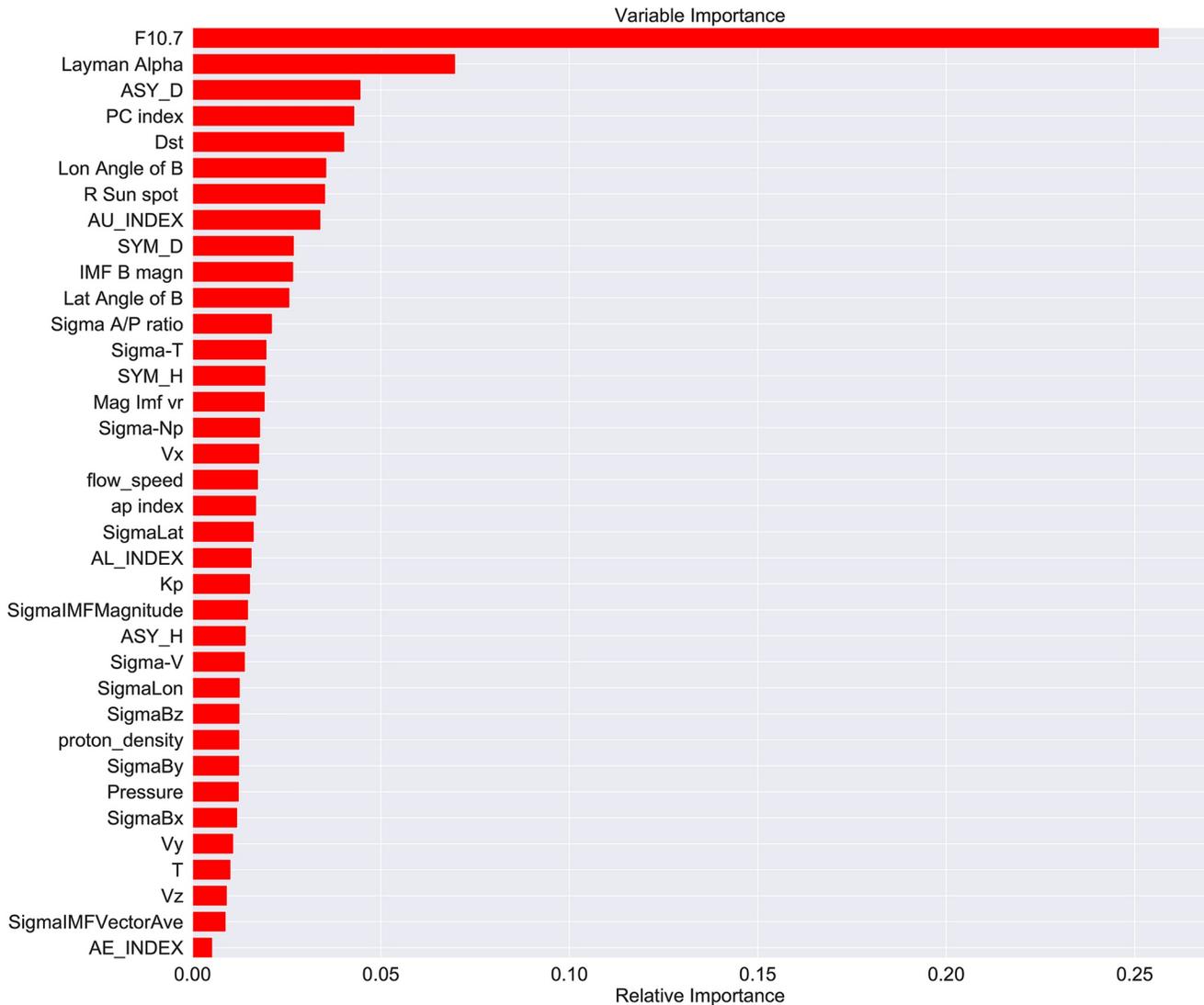


Figure 6. Showing the variable importance estimated using the random forest method and their ranking order.

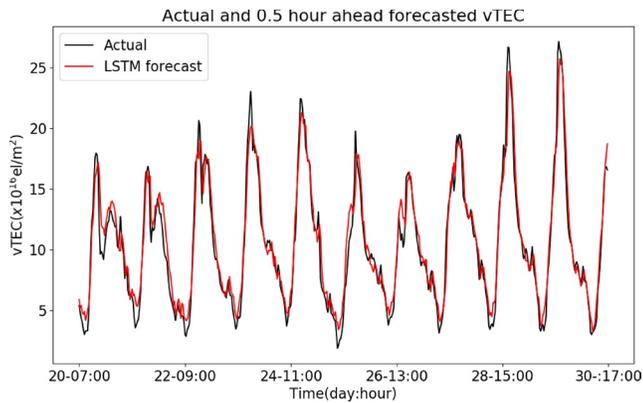


Figure 7. 30-min ahead forecast of the vTEC using the Long-Short Term Memory method.

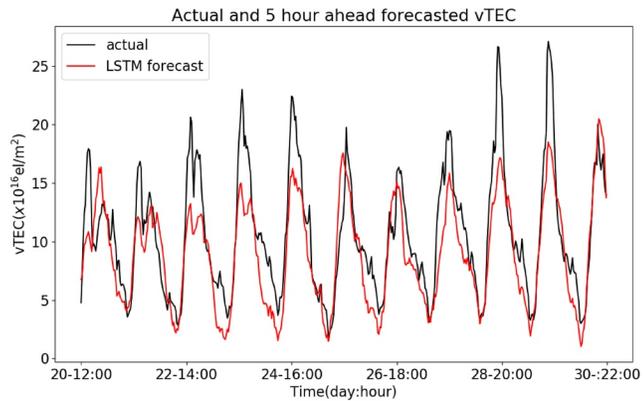


Figure 8. 5 h ahead forecast of the vTEC using the Long-Short Term Memory method.

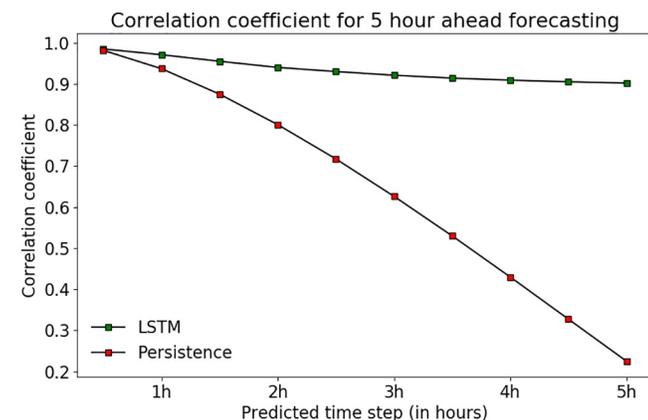


Figure 9. Correlation coefficient between real and forecasted total electron content, as a function of forecast horizon.

The result of the application of the LSTM method to forecast vTEC 30 min ahead into the future is shown in Figure 7. In Figure 7, the black and red curves respectively show the actual and forecasted vTEC 30 min into the future using the LSTM method. Similarly the forecasted and actual GPS TEC 5 h into the future is shown in Figure 8. Visual inspection of the Figures 7 and 8, clearly illustrates the 30 min ahead forecast a more accurate result than the forecast 5 h ahead of time. The 5 h ahead forecast in Figure 8 fails to get the various peaks and fluctuations in the vTEC.

Figures 7 and 8 show that we can effectively forecast the vTEC in the immediate future. And when we try to forecast further into the future, the model produces a less accurate forecast of the TEC. Further more, Figure 8 illustrates that the LSTM has relatively a hard time in forecasting the day time VTEC than the night time TEC. The LSTM method, especially, fails to capture fluctuations in the TEC during the day time compared to the night time. The night time vTEC doesn't fluctuate as much as the day time vTEC and therefore it is easy to forecast using the LSTM. In some situations the LSTM has a tendency to over estimate the day time peak vTEC while effectively capturing the night time trough.

The forecasting of GPS vTEC at the Bogota, Columbia station has been done in 30-min segments for up to 5 h into the future. Root mean square (RMS) error and correlation coefficient between the LSTM forecasted and actual vTEC for every half hour forecast has been calculated and shown, respectively, in Figures 9 and 10. For baseline performance comparison the persistence forecast is also included. For both LSTM and persistence forecast, the correlation coefficient decreases as we forecast further into the future (Figure 9). Similarly, the RMS error increases as we progress from a half hour forecast to the 5 h lead time forecast. However, even for a five-hour lead time forecast, the correlation coefficient remains fairly high, 0.88 and the persistence forecast falls significantly as we forecast further into the future.

4. Summary and Discussion

Forecasting the ionospheric space weather is important to mitigate its effect on global navigation and communication systems and power grids. However, developing a comprehensive and accurate ionospheric forecasting model is a challenge as the physical drivers are complex across the different regions and seasons (Mallika et al., 2018). Ionospheric TEC is a key parameter in describing the state of the ionosphere and accurate forecasting of it is crucial for ionospheric space weather characterization and mitigation.

The complex nature of the global ionosphere is shaped by solar and interplanetary activities as wells inputs from stratosphere, troposphere and mesosphere. Various long and short term physics and data assimilation based forecasts have been developed in the past (Amerian et al., 2013; Jakowski et al., 2011). However, physics based models hardly capture the complex structure of the ionosphere as the mathematical relationship between the solar, geomagnetic and lower atmospheric parameters across the various ionospheric geographic regions and different altitudes is not comprehensively and precisely known. Empirical models such as the IRI (Bilitza, 2001) and Nequick models (Nava et al., 2008) which have

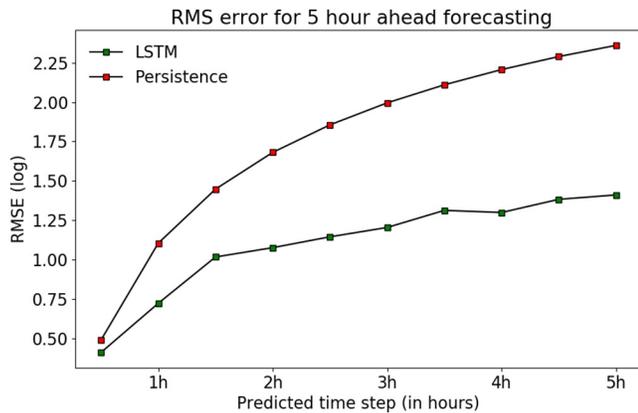


Figure 10. Root mean square error between real and forecasted total electron content, as a function of forecast horizon.

been commonly applied for ionospheric TEC forecasting are biased and produced inaccurate predictions. For example comparison of direct TEC derived from GPS and IRI model have shown large temporal discrepancies especially over the low latitude ionosphere (Karia et al., 2015).

Therefore, data driven forecasting based on advanced machine learning methods is an ideal candidate for effectively forecasting not only just the TEC but also other space weather events. Although the application of machine learning for forecasting is not necessarily new, it is at its golden time due to the availability of large data sets and the increase in computational power based on specialized processors (Camporeale, 2019). The near Earth space environment is particularly suitable to apply advanced machine learning methods due to the ubiquitously availability of data from different satellites, balloons, ground based magnetometers and GPS receivers, radars and imagers for about half a century.

In this paper, we used the random forest and LSTM machine learning methods to forecast $vTEC$ using an equatorial GPS station at Bogota, Columbia. The random forest method was used to estimate the variable importance and rank them in order of their influence to the model. The LSTM deep recurrent neural network is then used to forecast the $vTEC$ 5 h ahead every half hour. Only the top 5 parameters in the random ranking as well as the past history of the TEC are used as input for the LSTM forecasting. Using a selected number of the top predictors highly improves the computational necessity and training time. The method is able to forecast the TEC even in the drastic behavior of the equatorial ionosphere. And undoubtedly, it will perform much better at the mid-latitude ionosphere where drastic changes in the ionosphere are uncommon. Subsequent papers will present the application of the method at the polar region ionosphere. The method can also be applied to forecast other space weather parameters at all regions and altitudes.

A few important points should be indicated regarding the variables used in the forecasting and the machine learning methods. It has been widely known that the equatorial ionosphere is driven by solar radiation coming from the sun and the parameters which are known to influence the TEC are solar and magnetic indices, geographic position of the receiver and line of sight of the satellite (Habarulema et al., 2009; Hofmann-Wellenhof et al., 2012). As we clearly see in the variable importance ranking using the random forest machine learning (see Figure 6), the solar parameters: F10.7 solar flux, layman alpha, R sun spot and the magnetic indices: ASY_D, Dst and PC indices stood in the top 6. This is a testimony that the machine learning results agree well with the physics of ionospheric formation.

A few other recent researchers used the LSTM methods to forecast different space weather parameters and its application is on the rise. For example, Yang et al. (2018) used the LSTM method to forecast the occurrence of geomagnetic storms ($kp > 5$) using the solar wind and planetary magnetic field as well the kp index itself to the LSTM method. Yang et al. (2018) found that the LSTM method improved the kp forecasting, compared to other methods, despite a lower mean absolute and mean squared errors. Our application of the LSTM method to forecast $vTEC$ produced results with higher correlation coefficient (≈ 0.98) at short lead time than at long lead time. A recent work by Z. Chen et al. (2019) used a hybrid LSTM and CNN (Convolutional Neural Network) for forecast extreme weather events bringing superior results than traditional numerical and statistical methods. Application of the hybrid LSTM-CNN method will help for ionospheric 2D and 3D forecasting and identifying electron density perturbations.

Data Availability Statement

All the TEC values corresponding to Bogota in Colombia can be found at the LISN server at the following address: lisn.igp.gob.pe. Other solar wind, magnetosphere, and ionosphere parameters are from the NASA Space Physics Data Facility (SPDF) web page (omniweb.gsfc.nasa.gov).

Acknowledgments

This work was supported by the Defense Advanced Research Projects Agency (DARPA) under Department of the Interior Award D19AC00009 to Georgia Tech. We would like to thank the International GNSS Service (IGS), Geocentric Reference System for the Americas (SIRGAS) for providing GPS data. One of the authors Dr. Valladares was partially supported by Grants AGS-1552161, AGS-1563025, AGS-1933056 and ONR contract N-00014-17-1-2157. The Low Latitude Ionospheric Sensor Network (LISN) is a project led by the University of Texas at Dallas in collaboration with the Geophysical Institute of Peru and other institutions that provide information in benefit of the space weather scientific community. The authors would also like to thank NASA's Space Physics Data Facility (SPDF) for making freely available different space weather related data sets.

References

- Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., et al. (2016). Tensorflow: A system for large-scale machine learning. In *12th {USENIX} symposium on operating systems design and implementation ({OSDI}16)* (pp. 265–283).
- Amerian, Y., Hossainali, M. M., & Voosoghi, B. (2013). Regional improvement of IRI extracted ionospheric electron density by compactly supported base functions using GPS observations. *Journal of Atmospheric and Solar-Terrestrial Physics*, *92*, 23–30. <https://doi.org/10.1016/j.jastp.2012.09.011>
- Basu, S., Basu, S., Valladares, C., Yeh, H.-C., Su, S.-Y., MacKenzie, E., et al. (2001). Ionospheric effects of major magnetic storms during the International Space Weather Period of September and October 1999: GPS observations, VHF/UHF scintillations, and in situ density structures at middle and equatorial latitudes. *Journal of Geophysical Research: Space Physics*, *106*(A12), 30389–30413. <https://doi.org/10.1029/2001ja001116>
- Bengio, Y., Simard, P., & Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, *5*(2), 157–166. <https://doi.org/10.1109/72.279181>
- Bilitza, D. (2001). International reference ionosphere 2000. *Radio Science*, *36*(2), 261–275. <https://doi.org/10.1029/2000rs002432>
- Bilitza, D. (2018). Iri the International Standard for the Ionosphere. *Advances in Radio Science*, *16*, 1–11. <https://doi.org/10.5194/ars-16-1-2018>
- Bobra, M. G., & Ilonidis, S. (2016). Predicting coronal mass ejections using machine learning methods. *The Astrophysical Journal*, *821*(2), 127. <https://doi.org/10.3847/0004-637x/821/2/127>
- Bortnik, J., Chu, X., Ma, Q., Li, W., Zhang, X., Thorne, R. M., et al. (2018). Artificial neural networks for determining magnetospheric conditions. In *Machine learning techniques for space weather* (pp. 279–300). Elsevier. <https://doi.org/10.1016/b978-0-12-811788-0.00011-1>
- Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. <https://doi.org/10.1023/a:1010933404324>
- 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>
- Camps-Valls, G. (2009). Machine learning in remote sensing data processing. In *2009 IEEE international workshop on machine learning for signal processing* (pp. 1–6).
- Chen, R., Wang, X., Zhang, W., Zhu, X., Li, A., & Yang, C. (2019). A hybrid CNN-LSTM model for typhoon formation forecasting. *Geoinformatica*, *23*(3), 375–396. <https://doi.org/10.1007/s10707-019-00355-0>
- Chen, Z., Jin, M., Deng, Y., Wang, J.-S., Huang, H., Deng, X., & Huang, C.-M. (2019). Improvement of a deep learning algorithm for total electron content maps: Image completion. *Journal of Geophysical Research: Space Physics*, *124*(1), 790–800. <https://doi.org/10.1029/2018ja026167>
- Davies, K. (1990). *Ionospheric radio* (No. 31). IET.
- Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The Elements of Statistical Learning* (Vol. 1). Berlin: Springer series in statistics Springer.
- Gardner, M. W., & Dorling, S. (1998). Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. *Atmospheric Environment*, *32*(14), 2627–2636. [https://doi.org/10.1016/s1352-2310\(97\)00447-0](https://doi.org/10.1016/s1352-2310(97)00447-0)
- Gensler, A., Henze, J., Sick, B., & Raabe, N. (2016). Deep learning for solar power forecasting—an approach using AutoEncoder and LSTM neural networks. In *2016 IEEE international conference on systems, man, and cybernetics (SMC)* (pp. 002858–002865).
- Genuer, R., Poggi, J.-M., & Tuleau-Malot, C. (2010). Variable selection using random forests. *Pattern Recognition Letters*, *31*(14), 2225–2236. <https://doi.org/10.1016/j.patrec.2010.03.014>
- Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT press.
- Graves, A. (2012). Supervised sequence labeling. In *Supervised sequence labeling with recurrent neural networks* (pp. 5–13). Springer. https://doi.org/10.1007/978-3-642-24797-2_2
- Graves, A., Mohamed, A.-R., & Hinton, G. (2013). Speech recognition with deep recurrent neural networks. In *2013 IEEE international conference on acoustics, speech and signal processing* (pp. 6645–6649).
- Gross, N., & Cohen, M. (2020). VLF remote sensing of the D region ionosphere using neural networks. *Journal of Geophysical Research: Space Physics*, *125*(1), e2019JA027135. <https://doi.org/10.1029/2019ja027135>
- Gulli, A., & Pal, S. (2017). *Deep learning with keras*. Packt Publishing Ltd.
- Guzella, T. S., & Caminhas, W. M. (2009). A review of machine learning approaches to spam filtering. *Expert Systems with Applications*, *36*(7), 10206–10222. <https://doi.org/10.1016/j.eswa.2009.02.037>
- Habarulema, J. B., McKinnell, L.-A., Cilliers, P. J., & Opperman, B. D. (2009). Application of neural networks to South African GPS TEC modeling. *Advances in Space Research*, *43*(11), 1711–1720. <https://doi.org/10.1016/j.asr.2008.08.020>
- Hajj, G., Wilson, B., Wang, C., Pi, X., & Rosen, I. (2004). Data assimilation of ground GPS total electron content into a physics-based ionospheric model by use of the kalman filter. *Radio Science*, *39*(1). <https://doi.org/10.1029/2002rs002859>
- Han, H., Guo, X., & Yu, H. (2016). Variable selection using mean decrease accuracy and mean decrease Gini based on random forest. In *2016 7th IEEE international conference on software engineering and service science (icess)* (pp. 219–224).
- Hochreiter, S. (1998). The vanishing gradient problem during learning recurrent neural nets and problem solutions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, *6*(02), 107–116. <https://doi.org/10.1142/s0218488598000094>
- Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780. <https://doi.org/10.1162/neco.1997.9.8.1735>
- Hofmann-Wellenhof, B., Lichtenegger, H., & Collins, J. (2012). *Global positioning system: Theory and practice*. Springer Science & Business Media.
- Huang, Z., & Yuan, H. (2014). Ionospheric single-station TEC short-term forecast using RBF neural network. *Radio Science*, *49*(4), 283–292. <https://doi.org/10.1002/2013rs005247>
- Jakowski, N., Hoque, M., & Mayer, C. (2011). A new global TEC model for estimating transionospheric radio wave propagation errors. *Journal of Geodesy*, *85*(12), 965–974. <https://doi.org/10.1007/s00190-011-0455-1>
- Kaplan, E., & Hegarty, C. (2005). *Understanding GPS: Principles and applications*. Artech house.
- Karia, S., Patel, N., & Pathak, K. (2015). Comparison of GPS based TEC measurements with the IRI-2012 model for the period of low to moderate solar activity (2009–2012) at the crest of equatorial anomaly in Indian region. *Advances in Space Research*, *55*(8), 1965–1975. <https://doi.org/10.1016/j.asr.2014.10.026>
- Kelly, M. (2012). *The Earth's Ionosphere: Plasma Physics and Electrodynamics* (Vol. 43). Elsevier.
- Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv Preprint arXiv:1412.6980*.
- Kintner, P. M., Kil, H., Beach, T. L., & de Paula, E. R. (2001). Fading timescales associated with GPS signals and potential consequences. *Radio Science*, *36*(4), 731–743. <https://doi.org/10.1029/1999rs002310>
- Kotsiantis, S. (2007). Supervised machine learning: A review of classification techniques. *Informatica*, *31*, 249–268.

- Lary, D. J., Alavi, A. H., Gandomi, A. H., & Walker, A. L. (2016). Machine learning in geosciences and remote sensing. *Geoscience Frontiers*, 7(1), 3–10. <https://doi.org/10.1016/j.gsf.2015.07.003>
- Ling, A., Ginet, G., Hilmer, R., & Perry, K. (2010). A neural network-based geosynchronous relativistic electron flux forecasting model. *Space Weather*, 8(9), 1–14. <https://doi.org/10.1029/2010sw000576>
- Mallika, I. L., Ratnam, D. V., Ostuka, Y., Sivavaraprasad, G., & Raman, S. (2018). Implementation of hybrid ionospheric tec forecasting algorithm using PCA-NN method. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(1), 371–381.
- Mandrake, L., Wilson, B., Wang, C., Hajj, G., Mannucci, A., & Pi, X. (2005). A performance evaluation of the operational Jet Propulsion Laboratory/University of Southern California global assimilation ionospheric model (jpl/usc gain). *Journal of Geophysical Research: Space Physics*, 110(A12). <https://doi.org/10.1029/2005ja011170>
- McGranaghan, R. M., Mannucci, A. J., Wilson, B., Mattmann, C. A., & Chadwick, R. (2018). New capabilities for prediction of high-latitude ionospheric scintillation: A novel approach with machine learning. *Space Weather*, 16(11), 1817–1846. <https://doi.org/10.1029/2018sw002018>
- Mikolov, T., Joulin, A., Chopra, S., Mathieu, M., & Ranzato, M. (2014). Learning longer memory in recurrent neural networks. *arXiv Preprint arXiv:1412.7753*.
- Nava, B., Coisson, P., & Radicella, S. (2008). A new version of the NeQuick ionosphere electron density model. *Journal of Atmospheric and Solar-Terrestrial Physics*, 70(15), 1856–1862. <https://doi.org/10.1016/j.jastp.2008.01.015>
- Ngwira, C. M., Habarulema, J.-B., Astafyeva, E., Yizengaw, E., Jonah, O. F., Crowley, G., et al. (2019). Dynamic response of ionospheric plasma density to the geomagnetic storm of 22–23 June 2015. *Journal of Geophysical Research: Space Physics*, 124(8), 7123–7139. <https://doi.org/10.1029/2018ja026172>
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.
- Prybutok, V. R., Yi, J., & Mitchell, D. (2000). Comparison of neural network models with ARIMA and regression models for prediction of houston's daily maximum ozone concentrations. *European Journal of Operational Research*, 122(1), 31–40. [https://doi.org/10.1016/S0377-2217\(99\)00069-7](https://doi.org/10.1016/S0377-2217(99)00069-7)
- Scherliess, L., Thompson, D. C., & Schunk, R. W. (2009). Ionospheric dynamics and drivers obtained from a physics-based data assimilation model. *Radio Science*, 44(1). <https://doi.org/10.1029/2008rs004068>
- Strobl, C., Boulesteix, A.-L., & Augustin, T. (2007). Unbiased split selection for classification trees based on the Gini index. *Computational Statistics & Data Analysis*, 52(1), 483–501. <https://doi.org/10.1016/j.csda.2006.12.030>
- Tan, Y., Hu, Q., Wang, Z., & Zhong, Q. (2018). Geomagnetic index Kp forecasting with LSTM. *Space Weather*, 16(4), 406–416. <https://doi.org/10.1002/2017sw001764>
- Tzeng, F.-Y., & Ma, K.-L. (2005). *Opening the black box-data driven visualization of neural networks*. IEEE.
- Verikas, A., Gelzinis, A., & Bacauskiene, M. (2011). Mining data with random forests: A survey and results of new tests. *Pattern Recognition*, 44(2), 330–349. <https://doi.org/10.1016/j.patcog.2010.08.011>
- Villalobos, J., & Valladares, C. (2020). Statistical analysis of tec distributions observed over south and central america. *Radio Science*, 55(1), e2018RS006725. <https://doi.org/10.1029/2018rs006725>
- Wei, L., Zhong, Q., Lin, R., Wang, J., Liu, S., & Cao, Y. (2018). Quantitative prediction of high-energy electron integral flux at geostationary orbit based on deep learning. *Space Weather*, 16(7), 903–916. <https://doi.org/10.1029/2018sw001829>
- Wu, J.-G., & Lundstedt, H. (1996). Prediction of geomagnetic storms from solar wind data using Elman recurrent neural networks. *Geophysical Research Letters*, 23(4), 319–322. <https://doi.org/10.1029/96gl00259>
- Yang, Y., Shen, F., Yang, Z., & Feng, X. (2018). Prediction of solar wind speed at 1 AU using an artificial neural network. *Space Weather*, 16(9), 1227–1244. <https://doi.org/10.1029/2018sw001955>
- Yi, J., & Prybutok, V. R. (1996). A neural network model forecasting for prediction of daily maximum ozone concentration in an industrialized urban area. *Environmental Pollution*, 92(3), 349–357. [https://doi.org/10.1016/0269-7491\(95\)00078-x](https://doi.org/10.1016/0269-7491(95)00078-x)
- Yu, Y., Si, X., Hu, C., & Zhang, J. (2019). A review of recurrent neural networks: LSTM cells and network architectures. *Neural Computation*, 31(7), 1235–1270. https://doi.org/10.1162/neco_a_01199
- Zettergren, M., & Semeter, J. (2012). Ionospheric plasma transport and loss in auroral downward current regions. *Journal of Geophysical Research: Space Physics*, 117(A6), A06306. <https://doi.org/10.1029/2012ja017637>
- Zhang, L., Zhang, L., & Du, B. (2016). Deep learning for remote sensing data: A technical tutorial on the state of the art. *IEEE Geoscience and Remote Sensing Magazine*, 4(2), 22–40. <https://doi.org/10.1109/mgrs.2016.2540798>
- Zhao, Z., Chen, W., Wu, X., Chen, P. C., & Liu, J. (2017). LSTM network: A deep learning approach for short-term traffic forecast. *IET Intelligent Transport Systems*, 11(2), 68–75. <https://doi.org/10.1049/iet-its.2016.0208>