






## RESEARCH ARTICLE

10.1029/2022JA031183

# Predicting Swarm Equatorial Plasma Bubbles via Machine Learning and Shapley Values

**Special Section:**  
Machine Learning in  
Heliophysics

S. A. Reddy<sup>1</sup> , C. Forsyth<sup>1</sup> , A. Aruliah<sup>2</sup> , A. Smith<sup>3</sup> , J. Bortnik<sup>4</sup> , E. Aa<sup>5</sup> , D. O. Kataria<sup>6</sup>,  
and G. Lewis<sup>1</sup>

<sup>1</sup>Mullard Space Science Laboratory, University College London, London, UK, <sup>2</sup>Department of Physics and Astronomy, University College London, London, UK, <sup>3</sup>Department of Mathematics, Physics and Electrical Engineering, Northumbria University, London, UK, <sup>4</sup>Department of Atmospheric and Oceanic Sciences, University of California at Los Angeles (UCLA), Los Angeles, CA, USA, <sup>5</sup>Haystack Observatory, Massachusetts Institute of Technology, Cambridge, MA, USA, <sup>6</sup>Southwest Research Institute, San Antonio, TX, USA

### Key Points:

- AI Prediction of EPBs (APE) can accurately predict the Swarm Ionospheric Bubble Index
- APE is an XGBoost regressor that outperforms similarly trained linear and random forest models
- Game theory techniques reveal the influence of solar and geomagnetic activity as well as geo-location, time, and season

### Correspondence to:

S. A. Reddy,  
sachin.reddy.18@ucl.ac.uk

### Citation:

Reddy, S. A., Forsyth, C., Aruliah, A., Smith, A., Bortnik, J., Aa, E., et al. (2023). Predicting swarm equatorial plasma bubbles via machine learning and Shapley values. *Journal of Geophysical Research: Space Physics*, 128, e2022JA031183. <https://doi.org/10.1029/2022JA031183>

Received 23 NOV 2022

Accepted 16 MAY 2023

### Author Contributions:

**Conceptualization:** S. A. Reddy, A. Smith, J. Bortnik  
**Formal analysis:** S. A. Reddy, A. Smith, J. Bortnik, E. Aa, G. Lewis  
**Funding acquisition:** S. A. Reddy, C. Forsyth  
**Investigation:** S. A. Reddy  
**Methodology:** S. A. Reddy, A. Smith, J. Bortnik  
**Project Administration:** S. A. Reddy  
**Software:** S. A. Reddy  
**Supervision:** C. Forsyth, A. Aruliah, D. O. Kataria, G. Lewis  
**Validation:** S. A. Reddy, E. Aa  
**Visualization:** S. A. Reddy  
**Writing – original draft:** S. A. Reddy

**Abstract** In this study we present AI Prediction of Equatorial Plasma Bubbles (APE), a machine learning model that can accurately predict the Ionospheric Bubble Index (IBI) on the Swarm spacecraft. IBI is a correlation ( $R^2$ ) between perturbations in plasma density and the magnetic field, whose source can be Equatorial Plasma Bubbles (EPBs). EPBs have been studied for a number of years, but their day-to-day variability has made predicting them a considerable challenge. We build an ensemble machine learning model to predict IBI. We use data from 2014 to 2022 at a resolution of 1s, and transform it from a time-series into a 6-dimensional space with a corresponding EPB  $R^2$  (0–1) acting as the label. APE performs well across all metrics, exhibiting a skill, association and root mean squared error score of 0.96, 0.98 and 0.08 respectively. The model performs best post-sunset, in the American/Atlantic sector, around the equinoxes, and when solar activity is high. This is promising because EPBs are most likely to occur during these periods. Shapley values reveal that F10.7 is the most important feature in driving the predictions, whereas latitude is the least. The analysis also examines the relationship between the features, which reveals new insights into EPB climatology. Finally, the selection of the features means that APE could be expanded to forecasting EPBs following additional investigations into their onset.

## 1. Introduction

In the post sunset  $F$  region of the ionosphere, plumes of low density plasma, known as *Equatorial Plasma Bubbles* (EPBs) are prone to form. These bubbles were first observed in ionosonde traces, and have subsequently been captured by radar, air glow images, and in-situ detectors (Argo & Kelley, 1986; Retterer & Roddy, 2014; Woodman & La Hoz, 1976). EPBs can cause fluctuations in the amplitude and phase of radio waves that traverse through them (Kintner et al., 2007). These *scintillations* adversely affect Global Navigation Satellite System (GNSS) and other communication systems which rely on quiet ionospheric conditions. Their morphology, onset, and development is complex and has been the subject of numerous studies over the years.

In the sunlit hemisphere, the neutral wind generally travels in an easterly direction toward the day-night terminator (Heelis et al., 2012), forcing the plasma in an upwards zenith direction under the action of the Lorentz force. Once in the nightside, ionization ceases and recombination dominates. This leads to a large density gradient between the  $E$  and  $F$  regions. When the interface between these layers is perturbed, the rarified lower  $F$  layer is forced vertically upward into the higher density plasma, which itself is being pulled down under the action gravity (Kelley, 2009). This mechanism is known as a *Generalized Rayleigh-Taylor instability*,  $\gamma$ , and its growth rate is described by Sultan (1996). The growth rate of the RTI,  $\gamma$ , was formulated by Sultan in 1996

$$\gamma = \frac{\sum_P^F}{\sum_P^E + \sum_P^F} \left( V_p - U_L^P - \frac{g_e}{v_{eff}} \right) K^F - R, \quad (1)$$

where  $\sum_P$  is the flux integrated Pederson conductivity for the E and F layers,  $V_p$  is the vertical plasma drift,  $U_L^P$  is the Pederson conductivity weighted neutral meridional wind,  $g_e$  is the altitude corrected acceleration due to gravity,  $v_{eff}$  is the ion-neutral collision frequency,  $K^F$  is the total  $F$  region flux electron tube content, and  $R$  is the ion-electron recombination rate (Sultan, 1996). Because of the conductivity ratio of  $\sum_P^F / \sum_P^E + \sum_P^F$  the onset of an EPB can only occur at night when the  $E$  region conductivity is very weak. High values for  $\sum_P^E$ ,  $U_L^P$ , and  $R$  will

©2023. 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.

**Writing – review & editing:** C. Forsyth, A. Aruliah, A. Smith, J. Bortnik, E. Aa, G. Lewis

act to suppress an EPB, whereas high values of  $V_p$ ,  $K^F$  and  $g_e/v_{eff}$  will destabilize the plasma and enhance the likelihood of an EPB (Carter et al., 2020; Sultan, 1996).

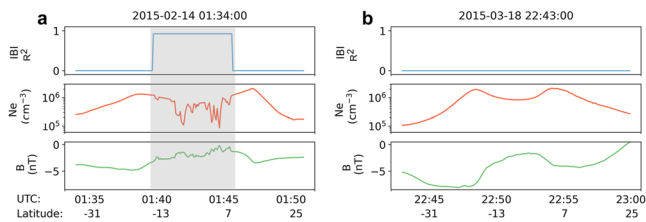
The spatiotemporal prediction of EPB occurrence has remained an on-going challenge for a number of years. Whilst the growth rate is described by Equation 1, the terms themselves are influenced by local time, geolocation, season, and solar and geomagnetic activity (Burke et al., 2004; Carter, Yizengaw, et al., 2014; S. Kumar et al., 2016; J. Smith & Heelis, 2017; Aa et al., 2020; Carter et al., 2020). To complicate matters, these climatological markers can often contradict themselves and the relationship between them is nuanced. Geomagnetic activity can both enhance and suppress the onset of an EPB via modified equatorial electrodynamics due to different perturbation electric fields (e.g., Aa et al., 2019; Abdu, 2012; Carter et al., 2016; S. Kumar et al., 2016). The under-shielding prompt penetration electric field (PPEF) tends to be dominant during the storm main phase due to suddenly varying magnetosphere convection, which has an eastward polarity in the dayside through local dusk but westward polarity in the nighttime. This typically enhances equatorial upward plasma drift in the dusk sector and thus facilitates the development of postsunset EPBs, but may disrupt post-midnight EPBs via downward plasma drift. On the other hand, the disturbance dynamo electric field (DDEF)—due to changes in global thermosphere circulation—usually dominates during the storm recovery phase, which has an opposite polarity with PPEF and so tends to suppress postsunset EPBs, but enhances postmidnight EPBs. In addition, the over-shielding penetration electric field due to substorm activity has an opposite polarity with that of PPEF, thereby suppressing the postsunset EPBs, but enhancing postmidnight EPBs. The combination and interaction of these perturbation electric fields leads to complicated occurrence patterns and spatio-temporal variations of EPBs.

Interest in machine learning (ML) within the heliophysics community has grown enormously in recent years (Camporeale, 2019), but its direct application to EPBs remains more limited. A *random forest regressor* has been employed to predict the vertical plasma drifts, or  $V_p$  in Equation 1 (Shidler & Rodrigues, 2020). This is a significant term in the overall onset of an EPB (Tsunoda et al., 2018). Others have used an all-sky imager to train a *convolution neural network* to detect EPBs, although the results seem more preliminary (Srisamoodkham et al., 2022). EPBs are also known as *Spread F*, which is a broader class of irregularities or wave-like structures within the ionosphere (Lan et al., 2018). Here ensemble and *deep learning* methods have been employed to classify and automatically detect Spread F in ionograms (Lan et al., 2018; Luwanga et al., 2022). EPBs are a known cause of radio wave scintillations (Kintner et al., 2007), and ML has been used to predict when and where scintillations may occur (Jiao et al., 2017; Linty et al., 2018; McGranaghan et al., 2018). Lastly, deep learning has also been applied to predict storm-driven irregularities within the ionosphere (Liu et al., 2021).

In this study we present AI Prediction of EPBs (APE), an ML model that predicts the Ionospheric Bubble Index (IBI) index on Swarm. First, we introduce Swarm and the IBI product. Then, we analyze the  $R^2$  value which is created by IBI and contains plasma bubbles. Third, we describe the ML models and their performance. Finally, we use Shapley values to interpret and explain the complex interactions within APE, all of which highlights the scientific benefits of using such an approach.

## 2. Instrumentation, Data and Observations

Swarm is a three-spacecraft Earth exploration constellation that launched on 22 November 2013. Two spacecraft, *Alpha* and *Charlie*, were at an initial altitude of roughly 470 km, whereas *Bravo* was at 520 km (Friis-Christensen et al., 2008). Alpha and Charlie operate side-by-side, separated by about  $1.4^\circ$  in longitude. All three have a circular near-polar orbit of  $87^\circ$ . Swarm automatically detects EPB's via its Ionospheric Bubble Index (IBI) product, which we use to train our machine learning models. EPBs can be characterized by prolonged and simultaneous changes in B and Ne (Stolle et al., 2006). Swarm has an on-board magnetometer and Langmuir probe to measure these quantities respectively. IBI correlates the strength of  $\Delta Ne$  and  $\Delta B$  (where residual B fluctuations in the range 0.04–0.5 Hz exceed 0.2 nT) using the Pearson correlation co-efficient ( $R$ ). An  $R^2 > 0.5$  is tagged as a “confirmed bubble” and  $<0.5$  is an “unconfirmed bubble.” In addition to a strong  $R^2$  score, bubbles are only confirmed if: detected at night, at latitude  $<45^\circ$ , there are no gaps in the data, and no non-physical measurements from the Langmuir probe or magnetometer. This reduces the risk of contamination from non-EPB events, but it does not stop some plasma blobs from being erroneously labeled as EPBs (Park et al., 2013). These will be more pronounced during solar minimum (Choi et al., 2012).



**Figure 1.** Two examples of Swarm passing over the equator. (a) Swarm detects an EPB as indicated by the gray-box. (b) Quiet time conditions with no bubbles present.

An example IBI EPB is shown inside the gray box of Figure 1a. Here a  $\Delta B$  occurs simultaneously with a  $\Delta N_e$  between the period 0140–0147, which in turn triggers an IBI  $R^2$  of 0.97. This value equates to a very high chance of EPB detection. A quiet bubble-free ionosphere is shown in Figure 1b.

IBI data was accessed via ESA's virtual research environment for Swarm (<https://vires.services>) and the Python package *virescient* (A. Smith et al., 2022). We also use *virescient* to map F10.7 and Kp values to the IBI data set. We use data from 2014 to 2022 at a resolution of 1s across all three spacecraft where  $R^2 > 0$ . The date range covers the declining phase of solar cycle 24 and the start of solar cycle 25. We transform the data from a time-series into a 6-dimensional space consisting of MLT, latitude, longitude, day-of-the-year, Kp, and F10.7, with each dimension having a corresponding  $R^2$  value (0–1) provided by IBI. This allows us to make a prediction of IBI based on the climatology of EPBs which are dependent on time, geolocation, season, and geomagnetic and solar activity (Aa et al., 2020; Burke et al., 2004; Carter et al., 2016, 2020). It also ensures that the model can be expanded to *forecasting*, as Kp and F10.7 are readily available via NOAA (<https://www.swpc.noaa.gov/products>). After re-binning and cleaning, we have  $\sim 42k$  samples for the machine learning models. Figure 2 shows the distribution of  $R^2$  across the 9-year period. As seen the majority of values cluster around  $R^2 = 0$  and  $R^2 = 0.9$ . We are mainly interested in  $R^2 > 0.7$ .

Next, we examine the distribution of the 42k samples across the 6 features. Figure 3 shows that “confirmed” and “unconfirmed” bubbles are not uniform across the climate markers.

Most confirmed bubbles are in the post-sunset time frame (19–24 MLT), with a small increase at 4 MLT (Figure 3a). The distribution of confirmed bubbles is centered around the geographic equator with only a few instances beyond  $25^\circ$  glat (Figure 3b). Next, we see that most bubbles occur in the American/Atlantic sector, but that instances exist at all longitudes (Figure 3c). The majority of EPBs occur around the equinox months and winter solstice, with little activity in July and August (Figure 3d). Figure 3e shows that the number of confirmed EPBs declines with Kp, and there are no bubbles detected at  $Kp > 7$ . Lastly, we see that EPB activity peaks around  $F10.7 = 125$ , but an additional population exists at  $F10.7 = 220$  (Figure 3f). This panel also reveals that EPBs are generally less likely to occur at  $F10.7 < 90$ . Overall, these results align with the existing literature on EPB climatology (e.g., Aa et al., 2020; Abdu, 2012; Burke et al., 2004; Carter et al., 2016; Park et al., 2013). Figure 3 also provides some insight into magnetic-only fluctuations ( $R^2 < 0.5$ ) in the ionosphere, with F10.7 and Kp showing some interesting distributions (Figures 3e and 3f).

Most confirmed bubbles are in the post-sunset time frame (19–24 MLT), with a small increase at 4 MLT (Figure 3a). The distribution of confirmed bubbles is centered around the geographic equator with only a few instances beyond  $25^\circ$  glat (Figure 3b). Next, we see that most bubbles occur in the American/Atlantic sector, but that instances exist at all longitudes (Figure 3c). The majority of EPBs occur around the equinox months and winter solstice, with little activity in July and August (Figure 3d). Figure 3e shows that the number of confirmed EPBs declines with Kp, and there are no bubbles detected at  $Kp > 7$ . Lastly, we see that EPB activity peaks around  $F10.7 = 125$ , but an additional population exists at  $F10.7 = 220$  (Figure 3f). This panel also reveals that EPBs are generally less likely to occur at  $F10.7 < 90$ . Overall, these results align with the existing literature on EPB climatology (e.g., Aa et al., 2020; Abdu, 2012; Burke et al., 2004; Carter et al., 2016; Park et al., 2013). Figure 3 also provides some insight into magnetic-only fluctuations ( $R^2 < 0.5$ ) in the ionosphere, with F10.7 and Kp showing some interesting distributions (Figures 3e and 3f).

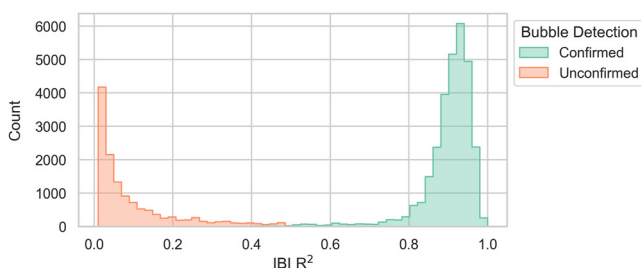
### 3. Machine Learning

We use supervised machine learning (ML) algorithms to predict the IBI value provided by Swarm. Supervised methods require *labels*,  $y_i$ , which we assign to  $R^2$ . We use regression specific architectures as the labels are considered a continuous value. ML has a unique ability to identify complex relationships in data that contains rare events. It can also handle heterogeneity in space-time and large amounts of noise (Camporeale, 2019; Karpatne et al., 2018). Because of this, we believe it is well suited to the task of predicting IBI and the EPBs contained within it.

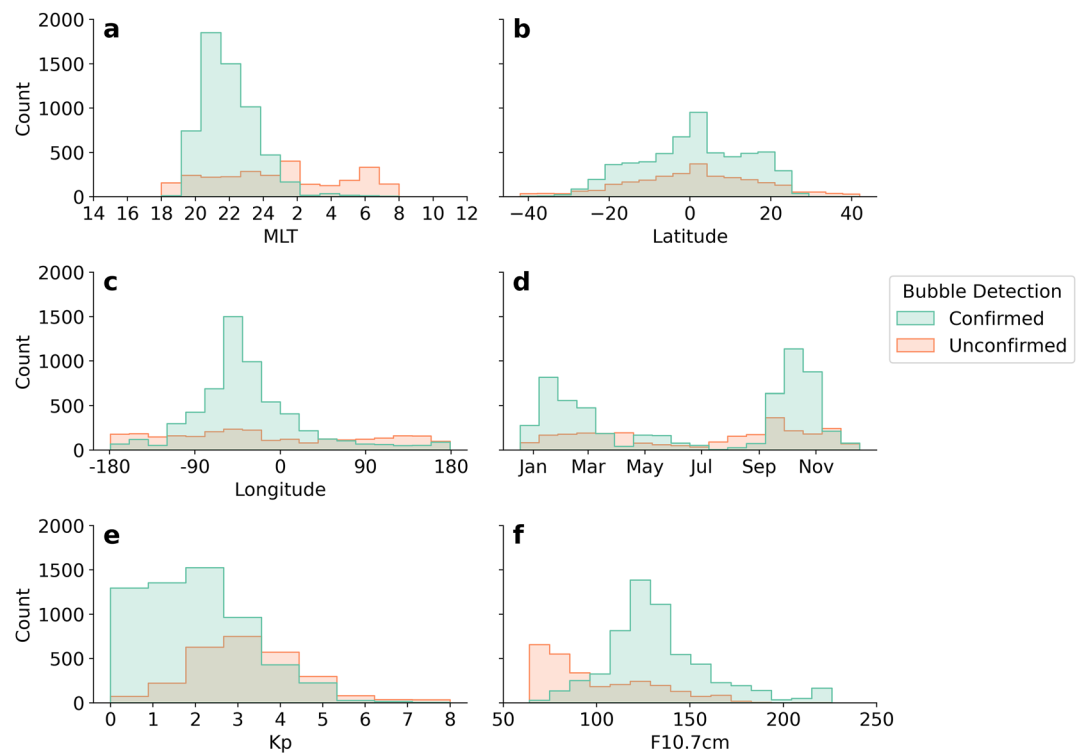
Our main algorithm is the *eXtreme Gradient Boosting* (XGBoost) method which is a tree-based ensemble learner. XGBoost has good control over bias and variance, whilst remaining computationally inexpensive to train and enabling *explainability* (Chen & Guestrin, 2016; Lundberg et al., 2020). The model's prediction ability is expressed by

$$\hat{y}_i = \sum_{k=1}^K f_k(x_i), f_k \in \mathcal{F}, \quad (2)$$

where  $\hat{y}_i$  is the prediction value,  $K$  is the number of trees,  $x_i$  is the input data,  $f_k$  is a function in the functional space  $\mathcal{F}$ , and  $\mathcal{F}$  is the set of all the possible regression trees (Chen & Guestrin, 2016). To evaluate the model's performance we need an *objective function* (Géron, 2019)



**Figure 2.** Distribution of  $R^2$  detected by IBI across 2014–2022, where  $R^2 > 0.5$  = “confirmed,” and  $R^2 < 0.5$  = “unconfirmed” (Park et al., 2013).



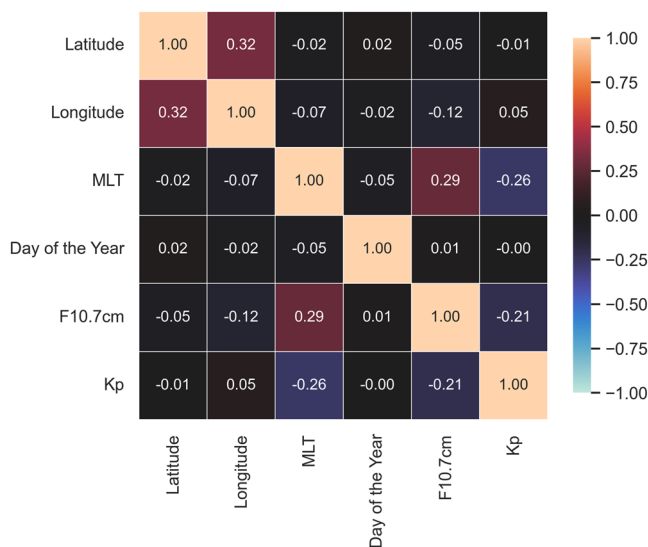
**Figure 3.** Bubble detection across the six climate features. *Confirmed* bubbles ( $R^2 > 0.5$ ) exhibit different spatio-temporal characteristics than *unconfirmed* magnetic-only fluctuations ( $R^2 < 0.5$ ).

$$obj = \sum_{i=1}^n l(y_i, \hat{y}_i^{(t)}) + \sum_{k=1}^K \omega(f_k), \quad (3)$$

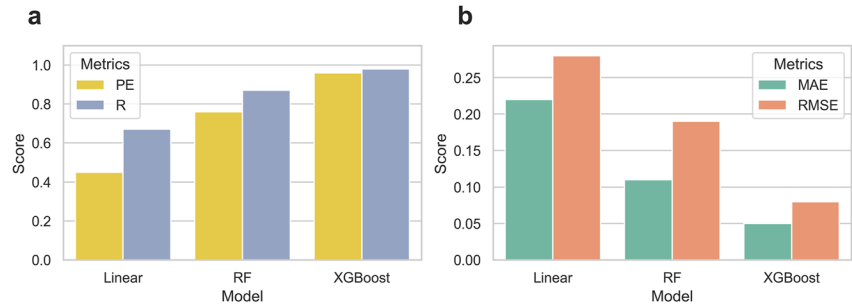
where  $y_i$  is the target value ( $R^2$ ),  $\hat{y}_i^{(t)}$  is the prediction of the  $i$ th instance at the  $t$ th iteration, and  $\omega$  is the complexity of the model (Chen & Guestrin, 2016). The term on the left is the *loss function*, and the term on the right is the *regularization* term. Regularization controls the magnitude of the parameters, and thus reduces the model's complexity (Géron, 2019). We use the XGBoost package for python ([xgboost.readthedocs.io](https://xgboost.readthedocs.io)) and Sci-kit learn ([scikit-learn.org](https://scikit-learn.org)) to perform the modeling and analysis. *GridSearchCV* was used to identify the optimal hyperparameters, which are as follows: *nestimators* = 300, *alpha* = 0.1, *subsample* = 0.5, and *eta* = 0.2. The last three parameters are used to prevent overfitting. We divide the samples into train and test datasets with a 80%–20% split. This is randomised initially and then fixed to prevent data leakage across the training runs.

We also tested a Random Forest method (Breiman, 2001) and a standard linear regression approach as part of our study. These will feature as a basis for global performance comparison, but are not subject to extensive analysis.

The model's input features and the linear correlation between them is shown in Figure 4. It reveals that there is no strong *linear* correlation between any of the features, which provides further justification for using an ML approach.



**Figure 4.** A correlation plot showing the relationship between the features. No strong linear correlation exists between any of the features.



**Figure 5.** The skill (PE), association (R), and performance (MAE, RMSE) of three learning models on the 20% test set. XGBoost outperforms the random forest and linear method across all four metrics.

### 3.1. Assessment Metrics

Several metrics are used to assess the performance, skill, and association of the model. *Root Mean Squared Error* (RMSE) and *Mean Absolute Error* (MAE) are typical performance tests for regression problems (Chai & Draxler, 2014),

$$MAE = \frac{1}{n} \sum_{i=1}^n |(\hat{y}_i - y_i)|, \quad (4)$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i)^2}, \quad (5)$$

where  $n$  is the number of samples. Accuracy metrics tell us how close the prediction is to the true value, but they do not tell us how well the model captures the up-and-down trends of the data set. *Association* can be represented by the Pearson correlation coefficient  $R$

$$R = \frac{\sum (y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum (y_i - \bar{y})^2 \sum (\hat{y}_i - \bar{\hat{y}})^2}}, \quad (6)$$

This tells us if the predictions are close to the target in some part of the data range, but not in others. An ideal value is  $R = 1$ . Finally, we examine the skill of the model by looking at its *Prediction Efficiency* which is based on its mean square error (Murphy, 1988)

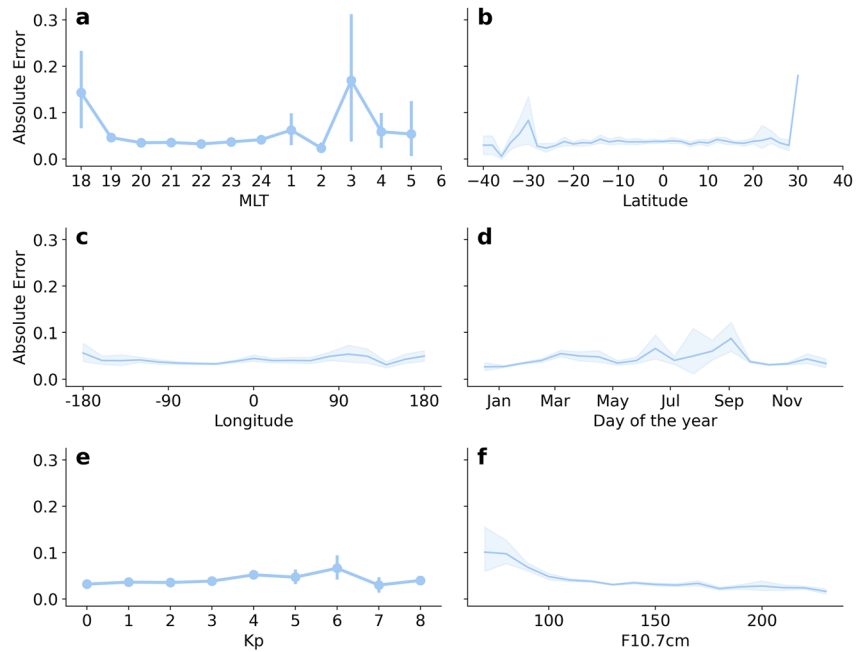
$$PE = 1 - \frac{\sum (\hat{y}_i - y_i)^2}{\sum (y_i - \bar{y})^2}, \quad (7)$$

A model with perfect skill is  $PE = 1$ , while  $PE < 0$  shows that the model is no better at making predictions than the average of the target values  $\langle y \rangle$ .

## 4. Results

The following section presents the performance of the machine learning models in terms of error, association, and skill. It goes on to interpret the behavior of the XGBoost model via Shapley values, determining the importance of the features and the relationships between them.

Figure 5a shows the association (Equation 6) and skill (Equation 7) of the three modeling techniques. As shown, the machine learning techniques outperform the standard linear model, particularly with respect to prediction efficiency (0.45 vs. 0.96), which justifies their use. The same trend continues with RMSE (Equation 5) and MAE (Equation 4), with the RF and XGBoost architecture outperforming the linear regression method across both metrics. The ensemble learners offer a considerable leap across the four metrics, but XGBoost comfortably outperforms the RF in all areas. It achieves a PE, R, MAE, and RMSE of 0.96, 0.98, 0.05, and 0.08 respectively, all of which are excellent scores. XGBoost also trains 3.8X faster than the RF, because it sub-samples and



**Figure 6.** The *absolute error*  $|(\hat{y}_i - y_i)|$  of APE across the 6 climate features. 0 is an ideal score. The uncertainties are calculated with the *bootstrapping* method (Efron & Tibshirani, 1994), and are represented by vertical bars (a) and (e) and the shaded areas (b–d, f).

approximates the split points amongst the trees (Chen & Guestrin, 2016). We now select the XGBoost model for further analysis and name it *AI Prediction of EPBs*, or APE.

#### 4.1. APE

Figure 5 tells us how APE is performing at a global level, but it does not tell us how it performs across the feature space. For example, does the model perform better at certain local times or during specific levels of geomagnetic activity? Figure 6 looks at the *absolute error* between the prediction and target,  $|(\hat{y}_i - y_i)|$ , across the features. Error bars are calculated using the *bootstrapping* method (Efron & Tibshirani, 1994).

Generally speaking, APE performs very well across the entire feature space (Figure 6). It performs poorer at 18 and 3 MLT (Figure 6a), outside the equatorial region (Figure 6b), and during low F10.7 (Figure 6f). These are periods when EPB activity is expected to be lower and is therefore not of concern. The performance also tracks directly to the availability of the data (Figure 3). That is, when there are more confirmed EPB events to learn from, model performance increases.

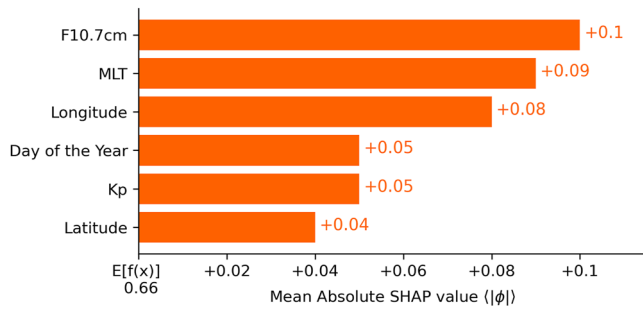
#### 4.2. Explainability

A key tenet of the study is to understand the factors that influence predictions, as well as the connections between them. To do this we use *Shapley Values*, which allow us to approximate feature contribution via cooperate game theory (Shapley, 1953). The *SHapley Additive exPlanations* (SHAP) package for Python ([shap-lrjball.readthedocs.io/](https://shap.lrjball.readthedocs.io/)) treats the features as players, and prediction of  $R^2$  as the pay-off (Lundberg et al., 2020). The predictions and SHAP contributions are calculated with

$$f(x) = E[f(x)] + \sum_n \phi_n, \tag{8}$$

where  $f(x)$  is the prediction of  $R^2$ ,  $E[f(x)]$  is the *expected value* which is  $\approx \langle R^2 \rangle$  and is equal to 0.66, and  $\phi_n$  is the SHAP value for each of the features  $n$ .  $\phi$  represents the contribution to the pay-off, weighted and summed over all possible feature value combinations. Shapley values have the properties of efficiency, symmetry, and additivity, which ensures the pay-off is *fair* (Lundberg et al., 2020; Shapley, 1953).  $E[f(x)]$  can be thought of as the climatology





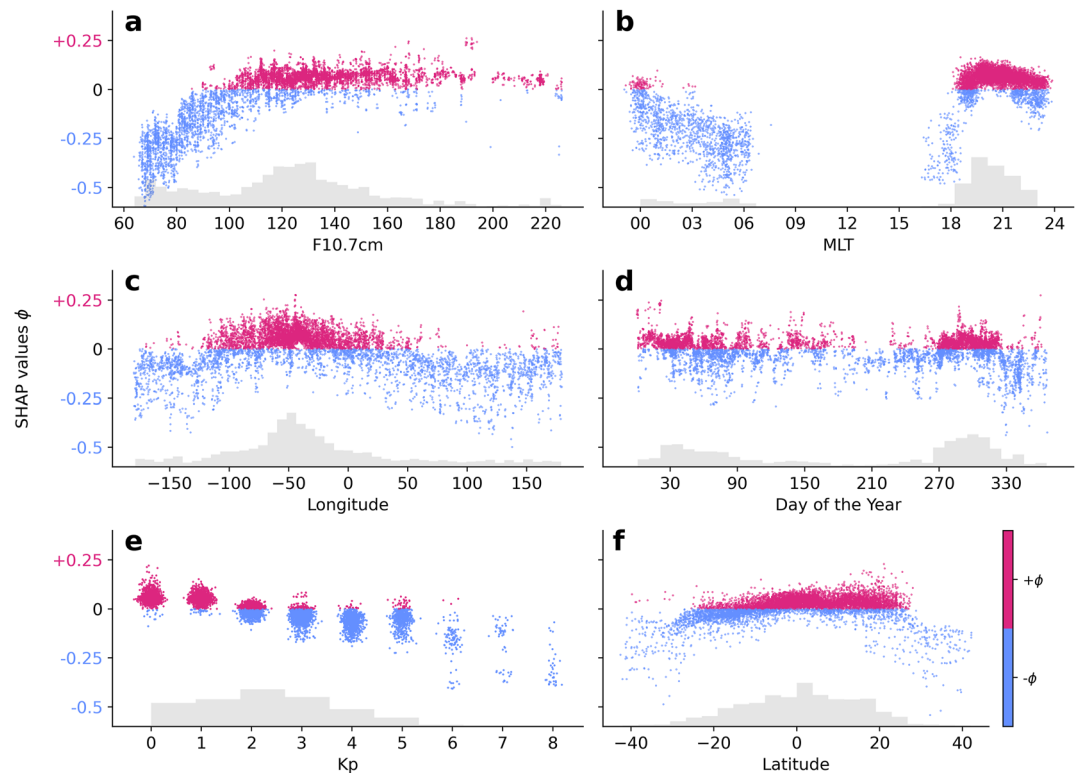
**Figure 7.** The mean absolute SHAP value across the six features. F10.7 contributes an absolute average of 0.1 to the 0.66 baseline and is considered the most important feature. Latitude contributes 0.04 to  $E[f(x)]$  is considered the least.

of  $R^2$ , and each of the feature values can contribute to this in a positive ( $\phi > 0$ ) or negative ( $\phi < 0$ ) way. Shapley values are emerging as the de facto method for explaining the output of ML models (Merrick & Taly, 2020), but their interpretation requires caution and expertise (I. E. Kumar et al., 2020).

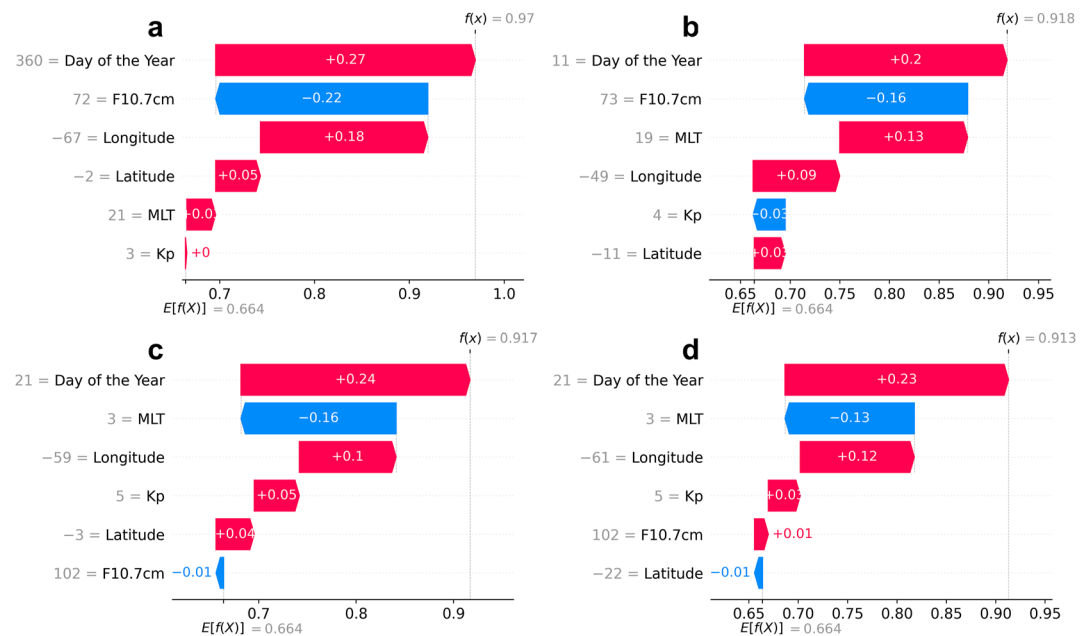
Figure 7 shows the mean absolute SHAP value across the six features. It shows that, on average, an F10.7 value will influence the prediction by 0.1, which is sufficient enough to consider a prediction a “confirmed bubble” (Park et al., 2013). Latitude contributes the least with  $\phi = 0.04$ . Figure 7 also shows that F10.7 is the most influential feature, whilst Latitude is the least.

We now turn our attention to the feature inputs and corresponding SHAP values. Figure 8 shows that  $\phi$  can be positive and negative, but Equation 8 means that we can only interpret the contribution to  $R^2$  when we take the sum of all the SHAP values.  $\phi > 0$  equates to increasing EPB likelihood, whereas  $\phi < 0$  is decreasing.

In the F10.7 panel (Figure 8a), we see that low solar activity corresponds to extremely negative SHAP values. This suggests that IBI is primarily detecting magnetic-only fluctuations and that EPBs require  $F10.7 > 90$ . Second, post-sunset values of MLT equate to the highest values of  $\phi$ , with the contribution peaking at 21 MLT (Figure 8b). It also shows a largely negative contribution after midnight, meaning that most EPBs occur after sunset. Longitude generally follows the known pattern of increased EPB formation over the American/Atlantic sector (Figure 8c), but there are positive contributions across the longitudinal space. Unlike the previous features, Day of the Year values generally contribute in a positive and negative way across the entire feature space (Figure 8d). We see high values of  $\phi > 0$  around the equinoxes and winter solstice, which is to be expected as EPB formation is generally highest during this period. That said, we also see a high positive cluster around the Earth-Sun perihelion, with the highest value of  $\phi$  on Day of the Year = 19. Kp provides perhaps the most intriguing



**Figure 8.** SHAP  $\phi$  contributions across the feature space.  $\phi > 0$  increases the predicted value of  $R^2$ , whereas  $\phi < 0$  decreases it. Predictions of  $R^2 > 0.7$  are considered to be EPBs, so large values of  $\phi > 0$  are more likely to be associated with plasma bubbles. Generally the SHAP values follow the climatology outlined in Figure 3.



**Figure 9.** A “waterfall” plot showing 4 predictions around the Earth-Sun perihelion and  $Kp > 2$ . The final prediction value is denoted by  $f(x)$ , and the values represent the contribution to this from the baseline  $E[f(x)] \approx \langle R^2 \rangle = 0.66$ . SHAP ensures that the sum of the contributions always enables a prediction between 0 and 1.

insight into EPB climatology (Figure 8e). It clearly shows that increasing  $Kp$  equates to negative SHAP values, which reduce the likelihood of an EPB. Beyond  $Kp > 6$  we only see  $\phi < 0$  which increases the likelihood of a  $B$ -only fluctuation. Lastly, we see that positive SHAP values are mainly centered around Latitude =  $0^\circ$ – $20^\circ$  which is expected given EPBs known formation and our use of geodetic coordinates (Figure 8f).

Next we examine some of the  $\phi > 0$  values at  $Kp = 4$ – $5$  and Day of the Year = 360 to 21. These are intriguing because the former are the *only* positive contributions to EPB prediction during a moderate storm, and the latter exhibits the highest  $\phi > 0$  contribution for that feature. Figure 9 illustrates the values for  $Kp$  and Day of the Year, as well as the other features that contribute to  $R^2$ . In all cases we see that the IBI value is  $> 0.9$ , and is therefore almost certainly an EPB (Park et al., 2013). It's also evident that Day of the Year is the dominant “player,” with contributions as high as  $\phi = +0.27$  (Figure 9a). More importantly, Figures 9c and 9d show the only examples of high  $Kp$  equating to positive SHAP values, which also coincide with the Earth-Sun perihelion. Examining this as a whole, Figure 9 shows that a combination of winter solstice/Earth-Sun perihelion,  $Kp > 2$ , and low F10.7 equates to a high chance of detecting an EPB.

## 5. Discussion

APE can reliably predict the IBI  $R^2$  index on Swarm. If it predicts an  $R^2 > 0.7$  it can be considered an EPB. The model has a high accuracy of  $RMSE = 0.08$  and exhibits excellent *skill* and *association*. SHAP values reveal the most important features, how features contribute to predictions, and the interrelation between them. We now expand on the IBI observations and SHAP values with respect to geomagnetic activity and seasonal effects. Generally speaking the IBI climate feature observations (Figure 3) and SHAP (Figures 8 and 9) values align with the existing literature: EPBs mainly occur in post-sunset, in the American/Atlantic sector, around the equinox months, and when solar activity is high (Aa et al., 2020; Abdu, 2012; Burke et al., 2004; Park et al., 2013). They also show that magnetic-only fluctuations ( $R^2 < 0.5$ ) are more likely post-midnight, during low F10.7, and high  $Kp$ .

The above suggests that geomagnetic activity *suppresses* EPB onset. This is supported by the test-set histograms in Figure 8e, which shows that less data *still* results in more positive SHAP contributions when  $Kp$  is low. However, geomagnetic activity is both able to *suppress* and *enhance* EPB formation, via DDEF and/or over and under-shielding electric fields (Aa et al., 2019; Abdu, 2012). Unfortunately, this cannot be fully captured



by concurrent Kp owing to DDEF's time-delay effects to the equator. It's possible that indices such as DST or AE are better suited to capturing this, but neither are currently available as forecast products, and thus were excluded from the feature space. To fully capture the influence of geomagnetic activity on EPBs, bespoke indices may be required. For now, this exceeds the remit of this study, especially when the model accuracy is as high as  $RMSE = 0.08$ . Kp has been shown to capture day-to-day variability of EPBs during "EPB Season" (Carter, Retterer, et al., 2014), but additional work is required to capture them during "off-season." If we assume that  $F10.7 < 90$  to be off-season, then Figures 9c and 9d, shows Kp could be useful at all times, particularly around the Earth-Sun perihelion. That said, Figures 9c–9d also shows that identical values of F10.7 can have different contributions for different predictions, which shows that interpreting Shapley values requires caution (I. E. Kumar et al., 2020). Another interesting feature is that  $Kp = 5$  also coincides with the only  $\phi > 0$  at 3 MLT (Figures 8b and 9c–d). EPBs are suppressed after sunset, but enhanced after midnight during large  $\Delta DDEF$  (Abdu, 2012), so these points could be direct evidence of over-shielding effects. That said, the MLT values contribute to the pay-off in a negative way ( $-\phi$ ) and others have reported that over-shielding is more impactful than under-shielding on vertical drifts (Hui & Vichare, 2019), and so more evidence is required to support this.

Turning to the cluster of positive SHAP values around the Earth-Sun perihelion (Figure 8d). We would not expect these  $\phi$  values to be higher than the vernal and autumnal equinoxes or December solstice when EPB onset is most probable (Burke et al., 2004). One possible explanation is an increase in the F region density around the Gregorian new year, potentially arising from the Earth-Sun perihelion (Rishbeth & Uller-Wodarg, 2006). The exact cause of this semi-annual variation remains unknown, but we do know that an increased  $\sum_p^F$  in Equation 1 would increase the growth rate of an EPB (Carter et al., 2020; Sultan, 1996). This F region asymmetry has also been linked to increased atmospheric gravity waves, which are a known seeding mechanism for EPBs (Abdu et al., 2009; Singh et al., 1997). That said, the asymmetry happens every year, yet we do not see a large number of points around this period. Although further investigation is required into both the seasonal and geomagnetic influences on EPB formation, this discussion highlights the potential of Shapley values to improve our understanding of bubble climatology and predictability.

## 6. Conclusions

In this paper have shown that machine learning can successfully predict the Ionospheric Bubble Index (IBI) on-board the Swarm spacecraft. IBI detects equatorial plasma bubbles in the ionosphere by assessing changes in the plasma density and magnetic field. AI Predictions of EPBs (APE) is able to accurately predict IBI across a range of spatio-temporal conditions. The main findings of our study are summarized below:

1. APE fully captures the climatology of EPBs detected by Swarm. This is made possible with the size and resolution of the data set (9 years @ 1s), feature selection, and regression-specific model architecture. APE could also be expanded to forecasting as Kp and F10.7 are currently available via NOAA.
2. The XGBoost approach outperforms the other methods (linear regression and random forest) across all metrics. It performs extremely well; presenting a skill, association, and root mean square error score of 0.96, 0.98, 0.08 respectively.
3. APE performs well across the entire feature space, especially post-sunset, in the American/Atlantic sector, around the equinoxes, and when solar activity is high. This is encouraging as most EPBs occur during these periods and locations. Extra consideration may be required when using APE around 3 MLT.
4. SHAP values reveal that F10.7 is the most influential feature, whereas latitude is the least. SHAP values generally align with the existing climatology of IBI EPBs, which validates these results.
5. Additional metrics may be required to fully capture the effects of geomagnetic activity on EPB predictions, but this may compromise APE's ability to forecast them. There is some evidence of high Kp generally suppressing EPB activity, but further investigation into under and over-shielding is required.
6. The Shapley analysis also reveals that a combination low solar activity, active geomagnetic conditions, and the Earth-Sun perihelion all contribute to increased EPB likelihood. To the best of our knowledge, this is the first time this exact combination of features has been linked to bubble detection. Although its underlying mechanism needs additional investigation, it does showcase the ability of Shapley values to enable new insights into EPB climatology and predictability.

## Acronyms

EPB	equatorial plasma bubble
GNSS	global navigation satellite system

PPEF	prompt penetration electric field
DDEF	disturbance dynamo electric field
ML	machine learning
APE	AI prediction of EPBs
IBI	ionospheric bubble index
MLT	magnetic local time
XGBoost	eXtreme Gradient Boosting
RF	random forest
SHAP	SHapley Additive exPlanations

## Data Availability Statement

Swarm datasets can be accessed with the VirES Python client (<https://github.com/ESA-VirES/VirES-Python-Client>), or via the Virtual Research Environment (<https://vires.services>). Users need to create a free-account to access the data.

## References

- Aa, E., Zou, S., & Liu, S. (2020). Statistical analysis of equatorial plasma irregularities retrieved from swarm 2013–2019 observations. *Journal of Geophysical Research: Space Physics*, 125(4), e2019JA027022. <https://doi.org/10.1029/2019ja027022>
- Aa, E., Zou, S., Ridley, A., Zhang, S., Coster, A. J., Erickson, P. J., et al. (2019). Merging of storm time midlatitude traveling ionospheric disturbances and equatorial plasma bubbles. *Space Weather*, 17(2), 285–298. <https://doi.org/10.1029/2018SW002101>
- Abdu, M. (2012). Equatorial spread f/plasma bubble irregularities under storm time disturbance electric fields. *Journal of Atmospheric and Solar-Terrestrial Physics*, 75, 44–56. <https://doi.org/10.1016/j.jastp.2011.04.024>
- Abdu, M., Alam Kherani, E., Batista, I., De Paula, E., Fritts, D., & Sobral, J. (2009). Gravity wave initiation of equatorial spread f/plasma bubble irregularities based on observational data from the spreadfex campaign. *Annales Geophysicae*, 27(7), 2607–2622. <https://doi.org/10.5194/angeo-27-2607-2009>
- Argo, P., & Kelley, M. (1986). Digital ionosonde observations during equatorial spread f. *Journal of Geophysical Research*, 91(A5), 5539–5555. <https://doi.org/10.1029/ja091ia05p05539>
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32. <https://doi.org/10.1023/a:1010933404324>
- Burke, W. J., Huang, C. Y., Gentile, L. C., & Bauer, L. (2004). Seasonal-longitudinal variability of equatorial plasma bubbles. *Annales Geophysicae*, 22(9), 3089–3098. <https://doi.org/10.5194/angeo-22-3089-2004>
- 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>
- Carter, B. A., Currie, J. L., Dao, T., Yizengaw, E., Retterer, J., Terkildsen, M., et al. (2020). On the assessment of daily equatorial plasma bubble occurrence modeling and forecasting. *Space Weather*, 18(9), e2020SW002555. <https://doi.org/10.1029/2020SW002555>
- Carter, B. A., Retterer, J., Yizengaw, E., Groves, K., Caton, R., McNamara, L., et al. (2014). Geomagnetic control of equatorial plasma bubble activity modeled by the TIEGCM with KP. *Geophysical Research Letters*, 41(15), 5331–5339. <https://doi.org/10.1002/2014gl060953>
- Carter, B. A., Yizengaw, E., Pradipta, R., Retterer, J., Groves, K., Valladares, C., et al. (2016). Global equatorial plasma bubble occurrence during the 2015 St. Patrick's Day storm. *Journal of Geophysical Research: Space Physics*, 121(1), 894–905. <https://doi.org/10.1002/2015ja022194>
- Carter, B. A., Yizengaw, E., Retterer, J. M., Francis, M., Terkildsen, M., Marshall, R., et al. (2014). An analysis of the quiet time day-to-day variability in the formation of postsunset equatorial plasma bubbles in the Southeast Asian region. *Journal of Geophysical Research: Space Physics*, 119(4), 3206–3223. <https://doi.org/10.1002/2013JA019570>
- Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)?—arguments against avoiding RMSE in the literature. *Scientific Model Development*, 7(3), 1247–1250. <https://doi.org/10.5194/gmd-7-1247-2014>
- Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 785–794).
- Choi, H.-S., Kil, H., Kwak, Y.-S., Park, Y.-D., & Cho, K.-S. (2012). Comparison of the bubble and blob distributions during the solar minimum. *Journal of Geophysical Research*, 117(A4), A04314. <https://doi.org/10.1029/2011ja017292>
- Efron, B., & Tibshirani, R. J. (1994). *An introduction to the bootstrap*. CRC press.
- Friis-Christensen, E., Lühr, H., Knudsen, D., & Haugmans, R. (2008). Swarm—an Earth observation mission investigating geospace. *Advances in Space Research*, 41(1), 210–216. <https://doi.org/10.1016/j.asr.2006.10.008>
- Géron, A. (2019). *Hands-on machine learning with scikit-learn, keras, and tensorflow: Concepts, tools, and techniques to build intelligent systems*. O'Reilly Media, Inc."
- Heelis, R., Crowley, G., Rodrigues, F., Reynolds, A., Wilder, R., Azeem, I., & Maute, A. (2012). The role of zonal winds in the production of a pre-reversal enhancement in the vertical ion drift in the low latitude ionosphere. *Journal of Geophysical Research*, 117(A8), A08308. <https://doi.org/10.1029/2012ja017547>
- Hui, D., & Vichare, G. (2019). Variable responses of equatorial ionosphere during undershielding and overshielding conditions. *Journal of Geophysical Research: Space Physics*, 124(2), 1328–1342. <https://doi.org/10.1029/2018ja025999>
- Jiao, Y., Hall, J. J., & Morton, Y. T. (2017). Automatic equatorial GPS amplitude scintillation detection using a machine learning algorithm. *IEEE Transactions on Aerospace and Electronic Systems*, 53(1), 405–418. <https://doi.org/10.1109/taes.2017.2650758>
- Karpatne, A., Ebert-Uphoff, I., Ravela, S., Babaie, H. A., & Kumar, V. (2018). Machine learning for the geosciences: Challenges and opportunities. *IEEE Transactions on Knowledge and Data Engineering*, 31(8), 1544–1554. <https://doi.org/10.1109/tkde.2018.2861006>
- Kelley, M. C. (2009). *The earth's ionosphere: Plasma physics and electrodynamics*. Academic press.
- Kintner, P. M., Ledvina, B. M., & De Paula, E. (2007). GPS and ionospheric scintillations. *Space Weather*, 5(9). <https://doi.org/10.1029/2006sw000260>

## Acknowledgments

SR designed the study, built the codes and ML models, analyzed the results, and wrote the manuscript. CF and AA provided ionospheric and space weather expertise. AWS and JB designed the ML pipeline and data transformation techniques. EA provided bubble and ionospheric irregularities expertise. DK and GL analyzed the results. All authors contributed to the editing of the manuscript. We thank Jaehung Park and Claudia Stolle for answering key questions related to plasma bubbles and the IBI product. We also give thanks to the machine learning in heliophysics community for their valuable feedback. SAR is supported by the Science & Technology Facilities Council under Grant ST/R505171/1. AA acknowledges NERC Grant Ref: NE/W003112/1. AWS is supported by STFC Consolidated Grant ST/S000240/1, and NERC Grants NE/P017150/1 and NE/V002724/1. JB gratefully acknowledges subgrant 1559841 to the University of California, Los Angeles, from the University of Colorado Boulder under NASA Prime Grant agreement 80NSSC20K1580. EA acknowledges NSF awards AGS-1952737 and AGS-2033787, as well as NASA support 80NSSC22K0171 and 80NSSC21K1310. And G.L. is supported by UKSA ST/X002152/1.

- Kumar, I. E., Venkatasubramanian, S., Scheidegger, C., & Friedler, S. (2020). Problems with Shapley-value-based explanations as feature importance measures. In *International conference on machine learning* (pp. 5491–5500).
- Kumar, S., Chen, W., Liu, Z., & Ji, S. (2016). Effects of solar and geomagnetic activity on the occurrence of equatorial plasma bubbles over Hong Kong. *Journal of Geophysical Research: Space Physics*, *121*(9), 9164–9178. <https://doi.org/10.1002/2016ja022873>
- Lan, T., Zhang, Y., Jiang, C., Yang, G., & Zhao, Z. (2018). Automatic identification of spread f using decision trees. *Journal of Atmospheric and Solar-Terrestrial Physics*, *179*, 389–395. <https://doi.org/10.1016/j.jastp.2018.09.007>
- Linty, N., Farasin, A., Favenza, A., & Dovis, F. (2018). Detection of GNSS ionospheric scintillations based on machine learning decision tree. *IEEE Transactions on Aerospace and Electronic Systems*, *55*(1), 303–317. <https://doi.org/10.1109/taes.2018.2850385>
- Liu, L., Morton, Y. J., & Liu, Y. (2021). Machine learning prediction of storm-time high-latitude ionospheric irregularities from GNSS-derived ROTI Maps. *Geophysical Research Letters*, *48*(20), 1–11. <https://doi.org/10.1029/2021gl095561>
- Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., et al. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, *2*(1), 56–67. <https://doi.org/10.1038/s42256-019-0138-9>
- Luwanga, C., Fang, T.-W., Chandran, A., & Lee, Y.-J. (2022). Automatic spread-F detection using deep learning. *Radio Science*, *57*(5), e2021RS007419. <https://doi.org/10.1029/2021rs007419>
- 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>
- Merrick, L., & Taly, A. (2020). The explanation game: Explaining machine learning models using Shapley values. In *International cross-domain conference for machine learning and knowledge extraction* (pp. 17–38).
- Murphy, A. H. (1988). Skill scores based on the mean square error and their relationships to the correlation coefficient. *Monthly Weather Review*, *116*(12), 2417–2424. [https://doi.org/10.1175/1520-0493\(1988\)116<2417:ssbotm>2.0.co;2](https://doi.org/10.1175/1520-0493(1988)116<2417:ssbotm>2.0.co;2)
- Park, J., Noja, M., Stolle, C., & Lühr, H. (2013). The ionospheric bubble index deduced from magnetic field and plasma observations onboard Swarm. *Earth Planets and Space*, *65*(11), 1333–1344. <https://doi.org/10.5047/eps.2013.08.005>
- Retterer, J. M., & Roddy, P. (2014). Faith in a seed: On the origins of equatorial plasma bubbles. *Annales Geophysicae*, *32*(5), 485–498. <https://doi.org/10.5194/angeo-32-485-2014>
- Rishbeth, H., & Uller-Wodarg, I. C. F. M. (2006). Why is there more ionosphere in January than in July? The annual asymmetry in the f2-layer (Vol. 24). Retrieved from [www.ann-geophys.net/24/3293/2006/](http://www.ann-geophys.net/24/3293/2006/)
- Shapley, L. S. (1953). Stochastic games. *Proceedings of the National Academy of Sciences of the United States of America*, *39*(10), 1095–1100. <https://doi.org/10.1073/pnas.39.10.1953>
- Shidler, S. A., & Rodrigues, F. S. (2020). Modeling equatorial ionospheric vertical plasma drifts using machine learning. *Earth Planets and Space*, *72*(1), 102. <https://doi.org/10.1186/s40623-020-01227-w>
- Singh, S., Johnson, F., & Power, R. (1997). Gravity wave seeding of equatorial plasma bubbles. *Journal of Geophysical Research*, *102*(A4), 7399–7410. <https://doi.org/10.1029/96ja03998>
- Smith, A., pacesm, & Santillan, D. (2022). *Esa-vires/vires-python-client: v0.10.2*. Zenodo. <https://doi.org/10.5281/zenodo.6515908>
- Smith, J., & Heelis, R. A. (2017). Equatorial plasma bubbles: Variations of occurrence and spatial scale in local time, longitude, season, and solar activity. *Journal of Geophysical Research: Space Physics*, *122*(5), 5743–5755. (Great intro. Good for writing up perhaps). <https://doi.org/10.1002/2017JA024128>
- Srisamoodkham, W., Shiokawa, K., Otsuka, Y., Ansari, K., & Jamjareegularn, P. (2022). Detecting equatorial plasma bubbles on all-sky imager images using convolutional neural network. In *Communication and intelligent systems* (pp. 481–487). Springer.
- Stolle, C., Lühr, H., Rother, M., & Balasis, G. (2006). Magnetic signatures of equatorial spread F as observed by the CHAMP satellite. *Journal of Geophysical Research*, *111*(2), 1–13. <https://doi.org/10.1029/2005JA011184>
- Sultan, P. (1996). Linear theory and modeling of the Rayleigh-Taylor instability leading to the occurrence of equatorial spread f. *Journal of Geophysical Research*, *101*(A12), 26875–26891. <https://doi.org/10.1029/96ja00682>
- Tsunoda, R. T., Saito, S., & Nguyen, T. T. (2018). Post-sunset rise of equatorial F layer—Or upwelling growth? *Progress in Earth and Planetary Science*, *5*(1), 22. <https://doi.org/10.1186/s40645-018-0179-4>
- Woodman, R. F., & La Hoz, C. (1976). Radar observations of F region equatorial irregularities. *Journal of Geophysical Research*, *81*(31), 5447–5466. <https://doi.org/10.1029/ja081i031p05447>