# Supplementary Material

## 1 SUPPLEMENTARY TEXT

### 1.1 The process of random forest training

Random Forest is an ensemble of $D$ trees $\{T_1(X), \cdots, T_D(X)\}$, where $X = \{x_1, \cdots, x_p\}$ is a $p$-dimensional vector of properties associated with a scintillation index. The ensemble produces $D$ output $\{\hat{Y}_1 = T_1(X), \cdots, \hat{Y}_D = T_D(X)\}$ where $\hat{Y}_d, d = 1, \cdots, D$, is the prediction for a scintillation index by the $d$th tree. Outputs of all trees are aggregated to produce one final prediction, $\hat{Y}$. For regression problem in this study, $\hat{Y}$ is the average of the individual tree predictions.

Given data on a set of $n$ radio occulation events for training, $S = \{(X_1, Y_1), \cdots, (X_n, Y_n)\}$, where $X_i, i = 1, \cdots, n$, is a vector of descriptors and $Y_i$ is either the corresponding desired value (e.g., the S4max intensity), the training algorithm proceeds as follows. (i) From the training data of $n$ events, draw a bootstrap sample (i.e., randomly sample, with replacement, $n$ samples). (ii) For each bootstrap sample, grow a tree with the following modification: at each node, choose the best split among a randomly selected subset of $m$ ($m < n$) descriptors. Here $m$ is essentially the only tuning parameter in the algorithm. The tree is grown to the maximum size (i.e., until no further splits are possible) and not pruned back. (iii) Repeat the above steps until (a sufficiently large number) $D$ such trees are grown.

### 1.2 The method to calculate the Altman-Bland plot

The steps for construction of V-shaped limits for the regression of differences on averages in Altman-Bland plots. (i) Calculate the average A (the average of observed and predicted values) and difference D (the difference of observed and predicted values). (ii) Construct the OLS regression of differences on averages. That is: $D = a + b(A)$. (iii) Extract the residuals as the differences between observed and predicted values. (iv) Convert the residuals into absolute values AR (i.e. remove negative signs). Construct the OLS regression of absolute residuals on averages. That is: $AR = a + b(A)$. (v) Adjust the coefficients for regression in above equation by multiplying them by $\sqrt{\pi/2} = 1.2533$. The sample standard deviation for the differences is: $SD = 1.2533a + 1.2533b(A)$. (vi) Thus, the limits for difference D for a given value of average A are: $PD \pm 1.96SD$, where the PD (predicted difference) is obtained from $D = a + b(A)$ and 1.96 is the standardized normal deviate corresponding to two-sided $P = 0.05$.

### 1.3 The preprocessing workflow for training SELF-ANN

This paragraph presents a comprehensive overview of the processing flow for training SELF-ANN. (i) The model initially acquires valid input variables, including year, month, day, hour, latitude, longitude, and altitude. (ii) According to the time and space parameters, the program will match other parameters corresponding to them, including geopotential, temperature, the u component of wind, Dst index, and F10.7 index in this work, as depicted by the variables with bold font in Table S1. They will be linearly interpolated if they are not consistent with the data set format. (iii) Each parameter will be normalized by the formula: $X_i = (X_i - \text{mean})/\text{std}$. (iv) The backward propagation algorithm is applied in the training process. (v) After the training, the forward propagation is performed in the trained SELF-ANN model, thereby generating the predicted S4max intensity.

## 2 SUPPLEMENTARY TABLES AND FIGURES

**Table S1.** The original input variables used in the random forest algorithm. Variables in bold font are used for SELF-ANN training.

| Variable | Description | Time | Reslution | Pressure Level |
|---|---|---|---|---|
| **local_time** | Local time of tangent point position at S4max | | | |
| **latitude** | Latitude of tangent point position at S4max | N/A | N/A | N/A |
| **longitude** | Longitude of tangent point position at S4max | | | |
| **altitude** | Altitude of tangent point position at S4max | | | |
| **f10.7** | The solar radio flux at 10.7 cm | 2008~2014 | Hour | |
| **dst** | The geomagnetic Dst index | 2008~2014 | Hour | N/A |
| **day_of_year** | The day of the year | N/A | N/A | |
| divergence | The horizontal divergence of velocity | | | |
| **geopotential** | The gravitational potential energy | | | |
| potential_vorticity | The capacity for air to ratate in the atmosphere | | | 5 hPa |
| relative_humidity | The water vapour pressure as a percentage | | | 10 hPa |
| **temperature** | The temperature in the atmosphere | 2008~2014 | $1° \times 1°$ | 50 hPa |
| **u_component_of_wind** | The eastward component of the wind | | | 200 hPa |
| v_component_of_wind | The northward component of the wind | | | 500 hPa |
| vertical_velocity | The speed of air motion in the vertical direction | | | |
| vorticity | The rotation of air in the horizontal | | | |

**Table S2.** The detailed architectures for the proposed model, SELF-ANN. fc means fully connected layers; relu means rectified linear unit function; softmax means $\mathrm{Softmax}(x_i) = (e^{x_i})/\sum_j e^{x_j}$.

| layer name | output size | SELF-ANN | stack |
|---|---|---|---|
| layer1 | 64×64 | fc layer, relu | 1 |
| layer2 | 112×112 | 7×7, 64, stride 2 | 1 |
| layer3 | 56×56 | 3×3 max pool, stride 2 | 1 |
| | | [1×1,64]<br>[3×3,64]<br>[1×1,256] | 3 |
| layer4 | 28×28 | [1×1,128]<br>[3×3,128]<br>[1×1,512] | 4 |
| layer5 | 14×14 | [1×1,256]<br>[3×3,256]<br>[1×1,1024] | 6 |
| layer6 | 7×7 | [1×1,512]<br>[3×3,512]<br>[1×1,2048] | 3 |
| layer7 | 1×1 | average pool, fc layer, softmax | 1 |
| layer8 | 1 | fc layer | 1 |
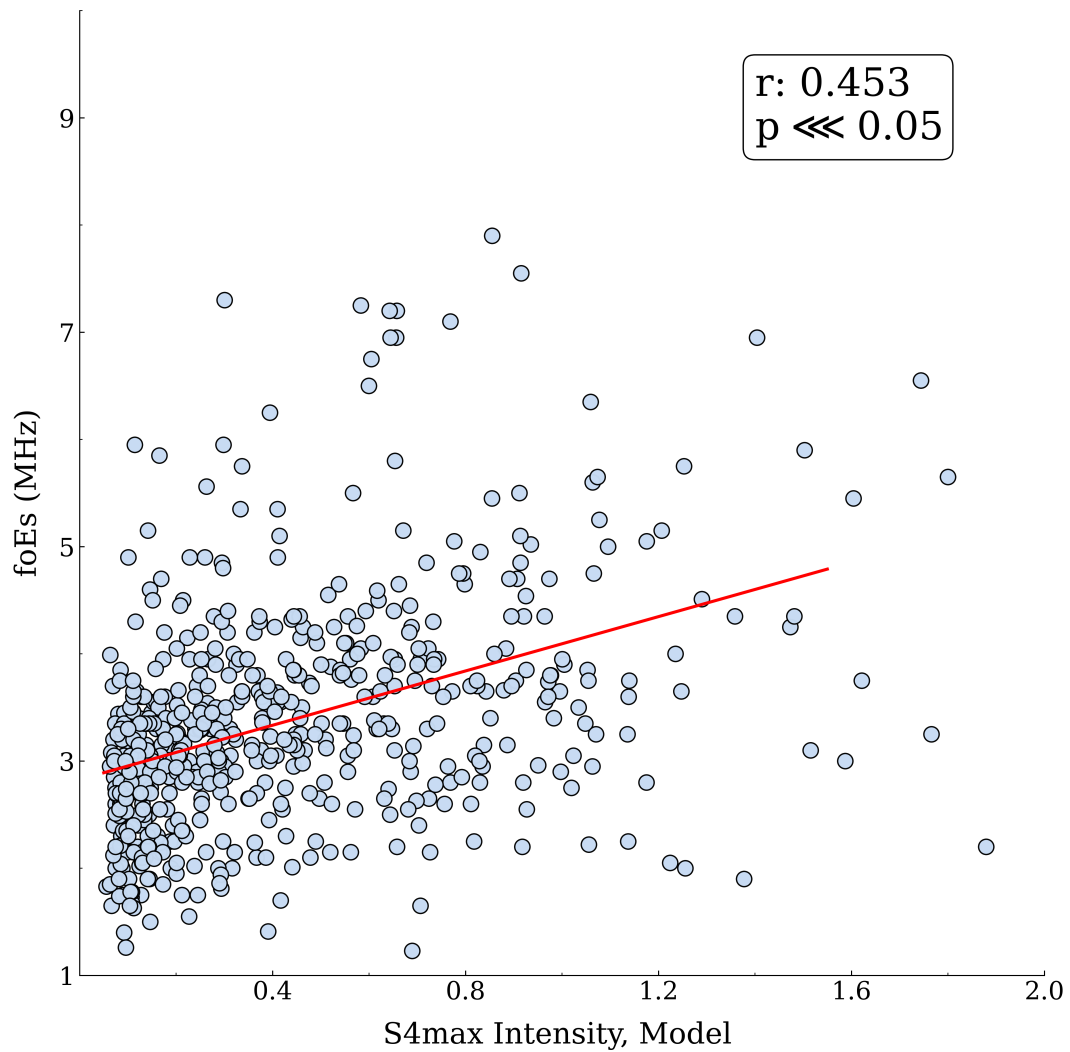
**Figure S1.** Scatter plot of the hourly manually scaled foEs measured by an ionosonde at Mohe (MH453) (52.0°N, 122.5°E) versus the hourly S4max scintillation index from SELF-ANN outputs in the period 2010–2014. The red line represents ordinary least square line of best fit.
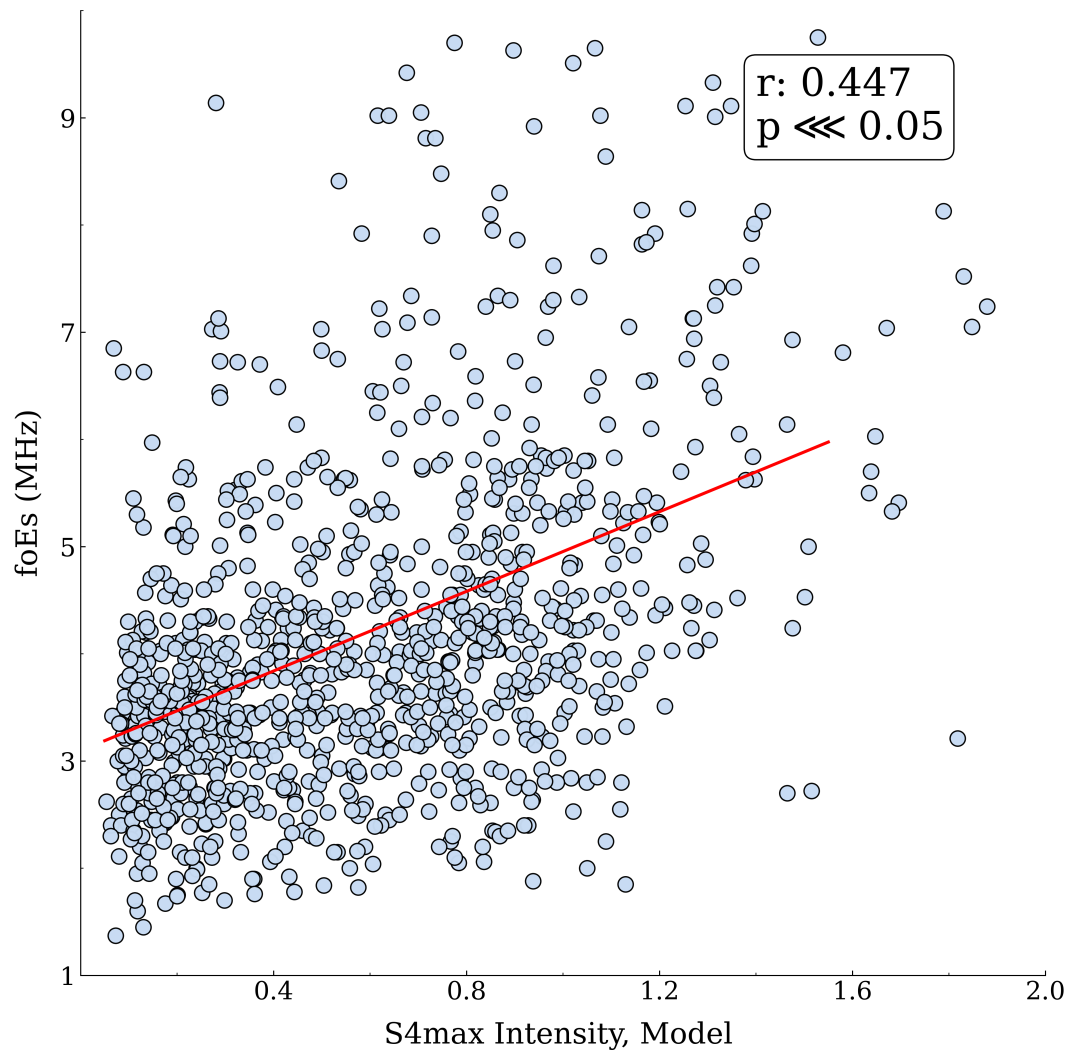
**Figure S2.** Scatter plot of the hourly manually scaled foEs measured by an ionosonde at Sanya (SA418) (18.3°N, 109.4°E) versus the hourly S4max scintillation index from SELF-ANN outputs in the period 2008–2014. The red line represents ordinary least square line of best fit.
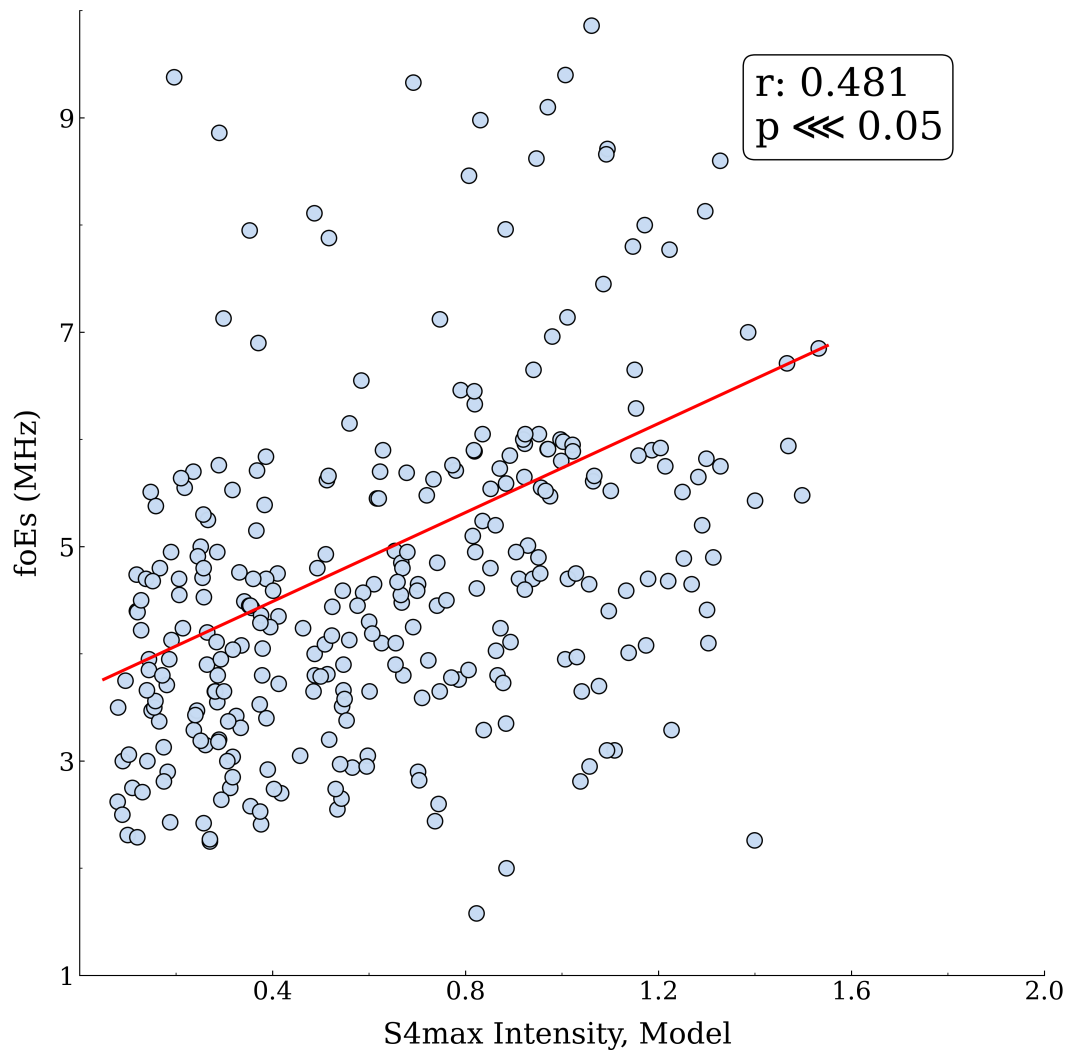
**Figure S3.** Scatter plot of the hourly manually scaled foEs measured by an ionosonde at Shaoyang (SH427) (27.1°N, 111.3°E) versus the hourly S4max scintillation index from SELF-ANN outputs in the period 2012–2014. The red line represents ordinary least square line of best fit.
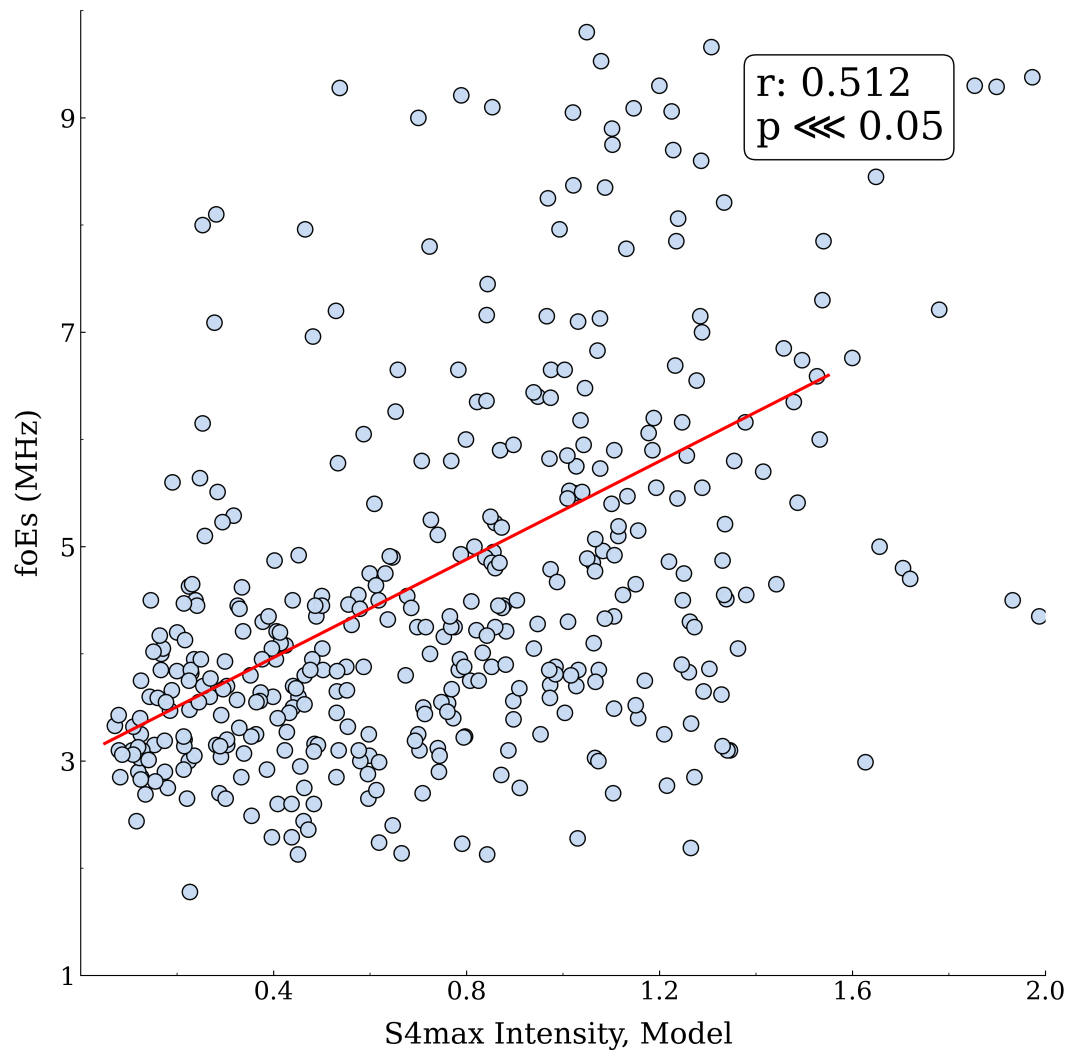
**Figure S4.** Scatter plot of the hourly manually scaled foEs measured by an ionosonde at Wuhan (WU430) (30.5°N, 114.4°E) versus the hourly S4max scintillation index from SELF-ANN outputs in the period 2010–2014. The red line represents ordinary least square line of best fit.