Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Forecasting the ionospheric propagation environment over lightning-active regions: a non-circular machine-learning formulation using NASA ISS-LIS, GNSS TEC and OMNI data

MD Firoz Alam
Indian Institute of Technology Roorkee
Mayank Goswami
mayankfph@iitr.ac.in

Indian Institute of Technology Roorkee

Research Article

Keywords:

Posted Date: November 10th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11671797/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Forecasting the ionospheric propagation environment over lightning-active regions: a non-circular machine-learning formulation using NASA ISS-LIS, GNSS TEC and OMNI data

MD Firoz Alam[1] and Mayank Goswami,[*1,2]

[1]Department of Physics, IIT Roorkee
[2]MF School of AI and DS, IIT Roorkee
mayank.goswami@ph.iitr.ac.in

November 9, 2026

Abstract

Lightning radiates electromagnetic energy over a very wide frequency range, and the low-frequency component travels to global distances inside the Earth–ionosphere waveguide. How far such a signal reaches, and the highest frequency a given path will carry, are set by the state of the ionosphere at the time of the discharge, which varies by the hour. We forecast that propagation environment over lightning-active regions and convert the forecast into the quantities a radio link actually needs: the F2-layer critical frequency f_oF_2, the maximum usable frequency (MUF) and the one-hop ground range. A central methodological point is addressed explicitly. Deriving f_oF_2 and MUF from Total Electron Content (TEC) and then training a model to predict them *from* TEC is circular, because the label is a closed-form function of an input feature; any model then recovers $R^2 \approx 1.0$ while learning nothing. We therefore pose the problem as a forecast of TEC at $t + \Delta t$, which is genuinely unobservable at t, and apply the physics to the forecast output rather than to the label. Using 381,858 ISS-LIS flashes, 17,665 fifteen-minute global ionosphere maps and 8,760 hourly solar and geomagnetic records spanning March–August 2018, five models are compared against a persistence benchmark at 1-, 2- and 6-hour lead times on a strictly chronological split. All five beat persistence at every lead time; histogram gradient boosting is best, reaching RMSE 3.15 TECU, R^2 0.68 and skill +0.68 at six hours, where persistence attains $R^2 = -2.11$. Skill *increases* with lead time because persistence decays across the diurnal cycle while solar geometry does not. Feature importance shows the model relying on current ionospheric state at short lead and switching to solar zenith angle at the target epoch at long lead, recovering the dominant physical driver without supervision. A case-by-case analysis qualifies the headline figure: at one-hour lead the model wins only 51.9% of individual cases and acts mainly as a variance-reducer. Lightning activity was found *not* to predict TEC, and we report that negative result rather than omit it.

1 Introduction

A lightning stroke is a broadband electromagnetic source. Its very-low-frequency (VLF) and extremely-low-frequency (ELF) components are guided between the conducting ground and the lower ionosphere and can be detected many thousands of kilometres from the parent storm [Rakov & Uman 2003]. The upper wall of that waveguide is a plasma whose electron density varies strongly with local time, season, latitude and geomagnetic activity.

Solar extreme-ultraviolet radiation ionises the upper atmosphere, producing free electrons distributed through the D, E and F layers. A plasma reflects electromagnetic waves below its characteristic plasma frequency and transmits those above it, so the ionosphere behaves as a frequency-dependent mirror [Davies 1990]. Total Electron Content, the integrated electron density in a column

*Corresponding author: mayank.goswami@ph.iitr.ac.in

of unit cross-section measured in TECU where 1 TECU = 10^{16} electrons m^{-2}, is the standard scalar description of that medium.

The propagation conditions for lightning-generated radio signals are therefore not fixed properties of geography but time-varying functions of the upper atmosphere. This paper asks whether that environment can be forecast several hours ahead from public satellite and ground-based observations, and whether lightning activity itself carries predictive information about it.

2 Problem, objectives and motivation

2.1 Problem

- **No labels exist.** The lightning dataset used here contains 22 flash-level variables, all optical or geometric. There is no field strength, no frequency, no direction and no ray path. The prediction target must be constructed, not read from a file.
- **The obvious construction is circular.** Computing f_oF_2 from TEC and then predicting it from TEC makes the label a deterministic function of an input. The resulting $R^2 \approx 1.0$ is an artefact of the experimental design.
- **Persistence is a strong baseline.** For ionospheric quantities, assuming nothing changes is hard to beat at short lead times, so a model must be measured against it rather than against zero.

2.2 Objectives

1. Build a machine-learning model that forecasts the ionospheric propagation environment over regions where lightning is occurring, posed so that the result is not circular.
2. Convert the forecast into physically meaningful propagation quantities: f_oF_2, MUF and one-hop ground range.
3. Benchmark honestly against persistence and report failure modes alongside successes.
4. Test whether lightning activity contributes predictive information about ionospheric state.
5. Use only public NASA data, with a reproducible pipeline.

2.3 Why it matters

Long-range and submarine communication links depend on the Earth–ionosphere waveguide, and the usable frequency band shifts with ionospheric state. VLF sferic networks locate lightning strokes by time-of-arrival, so errors in assumed propagation conditions translate directly into position errors. Ionospheric delay is the dominant error source for single-frequency GNSS receivers, which makes TEC forecasting an operational space-weather product. A geomagnetic storm alters TEC, and therefore all of the above, within hours.

3 Theory

3.1 From TEC to critical frequency

Peak electron density is obtained from TEC using an equivalent slab thickness τ:

$$N_mF_2 \approx \frac{\mathrm{TEC}}{\tau} \tag{1}$$

The plasma frequency follows from the standard relation, in which the coefficient 8.98 returns **hertz** when N_mF_2 is expressed in m^{-3}:

$$f_o F_2\,[\mathrm{Hz}] = 8.98\sqrt{N_m F_2} \tag{2}$$

This unit convention matters in practice. Writing the coefficient as 8.98×10^{-3} for an answer in MHz overstates $f_o F_2$ by a factor of 1000, giving critical frequencies of several thousand MHz instead of a few MHz. The error was made and caught during development, and is recorded here because it is silent: nothing fails, the numbers are merely wrong.

3.2 Oblique propagation

For a path with angle of incidence φ at the reflecting layer, the maximum usable frequency is

$$\mathrm{MUF} = \frac{f_o F_2}{\cos\varphi} \tag{3}$$

and the one-hop ground range for take-off elevation θ, reflection height h and Earth radius R is

$$D = 2R\left[\frac{\pi}{2} - \theta - \arcsin\!\left(\frac{R\cos\theta}{R+h}\right)\right] \tag{4}$$

Evaluated across the observed TEC range with $\tau = 300\,\mathrm{km}$, these give $f_o F_2$ between 3.67 and 13.12 MHz and MUF between 11.90 and 42.59 MHz at 5° take-off, with a one-hop ground range of 2,877 km. The implied M-factor $\mathrm{MUF}/f_o F_2 = 3.25$ is consistent with standard values for low take-off angles, and the $f_o F_2$ range matches observed mid-latitude values. Both serve as independent checks that the physics layer is correctly implemented.

4 Previous work

Two bodies of work bear on this problem.

Lightning and its radio signature. Lightning mapping arrays locate discharge sources by time-of-arrival of VLF/VHF radiation across multiple receivers, producing three-dimensional reconstructions of individual flashes; the techniques, their limitations and the prospects for deep learning are reviewed in [arXiv:2006.00563]. More directly relevant, Konan, Mishra & Lotz [2020] applied machine learning, including YOLO-based object detection, to the automatic detection and characterisation of *whistlers*, VLF waves generated by lightning that propagate between hemispheres along geomagnetic field lines, reporting misdetection and false-alarm rates below 15%. Together these establish that the radio signature of lightning is routinely measurable and that its propagation is governed by the ionosphere.

Machine-learning forecasting of the ionosphere. Cesaroni et al. [2020] developed a neural-network model for global TEC forecasting, and Liu et al. [2020] applied deep learning to the same problem. Subsequent work has adopted convolutional LSTM and transformer architectures to forecast global TEC maps at lead times from one hour to several days. These studies share a common framing: they forecast TEC globally and uniformly, driven by TEC history and space-weather indices.

Position of this work. The present study differs in three respects. It samples specifically at lightning locations rather than on a uniform global grid; it carries the forecast through to propagation quantities rather than stopping at TEC; and it explicitly tests whether lightning activity contributes predictive information about ionospheric state, a question the TEC-forecasting literature does not pose.

5 Method

5.1 Avoiding a circular formulation

The direct reading of the problem suggests: take TEC at the flash location, compute $f_o F_2$, compute MUF and one-hop range, and train a model to predict the latter from features including TEC. This is circular, because the label is a deterministic closed-form function of an input feature. A sufficiently

flexible model recovers the algebra and reports $R^2 \approx 1.0$, a score that carries no information about the physical system.

The problem is therefore recast as a **forecast**, which breaks the circularity because the target at $t + \Delta t$ is not observable at t. One sample is one (TEC grid cell, 15-minute epoch) in which lightning was observed. The inputs are lightning features, TEC history to t, solar and geomagnetic drivers, and solar geometry. The target is vertical TEC in that cell at $t + \Delta t$ for $\Delta t \in \{1\,\mathrm{h}, 2\,\mathrm{h}, 6\,\mathrm{h}\}$. The benchmark is persistence. The physics of Section 3 is applied to the forecast output and never used to construct the label.

Solar zenith angle at the *target* epoch is included as a feature. This is not leakage: it is deterministic astronomy, computable arbitrarily far in advance. No quantity derived from the target epoch's TEC enters the input set.

5.2 Data

Dataset	Role	Volume
NASA ISS-LIS V3 Quality Controlled	lightning flashes	381,858 flashes
UPC-IonSAT UQRG maps (NASA CDDIS)	TEC history and target	17,665 maps
NASA OMNI2 hourly	Kp, ap, Dst, AE, F10.7, sunspot	8,760 records

Table 1: Data sources. Period: 1 March – 31 August 2018, on a $2.5^\circ \times 5^\circ$ grid at 450 km.

Two details are worth stating correctly. First, ISS-LIS contains no propagation information whatsoever; all 22 of its flash-level variables are optical or geometric. Second, the IGS final combined ionosphere product is *2-hourly*; the 15-minute product used here is UPC-IonSAT's rapid map (UQRG), at 97 maps per day.

5.3 Feature construction

One sample is a (grid cell, 15-minute epoch) pair containing lightning. Aggregation is essential: a single storm produces thousands of flashes within one cell in one epoch, and retaining each flash as a separate row would place near-identical samples in both training and test partitions. The 381,858 flashes aggregate to 26,883 samples.

Forty features in five groups: lightning activity (flash count, radiance statistics, footprint, duration, group and event counts, oblongness); TEC history (at t and at $-15\,\mathrm{m}$, $-30\,\mathrm{m}$, $-1\,\mathrm{h}$, $-2\,\mathrm{h}$, $-3\,\mathrm{h}$, plus tendencies); spatial context (mean TEC of the eight neighbouring cells and the local gradient); geometry (latitude, magnetic latitude, solar zenith angle at t and $t + \Delta t$, local-time and seasonal harmonics); and space weather (Kp, ap, Dst, AE, F10.7, sunspot number).

5.4 Evaluation protocol

The partition is strictly chronological: earliest 70% for training, the next 15% for validation, and a final 15% test period untouched during model selection. Ionospheric fields are strongly autocorrelated in space and time, so a random split would distribute neighbouring 15-minute samples across partitions and permit effective lookup of the targets. The test period additionally contains the strongest geomagnetic storm of the interval, making the reported figures conservative.

5.5 Software

The pipeline comprises twelve scripts covering download, parsing, feature construction, training, evaluation and figure generation, plus a local web application for interactive forecasts. Processed data was verified against freshly downloaded source files rather than assumed correct; four defects were found and fixed, described in Section 6.5.

6 Results

6.1 Model comparison

Model	1 h RMSE	2 h RMSE	6 h RMSE	6 h R^2	6 h skill
Persistence (baseline)	4.875	5.565	9.813	−2.105	—
Ridge regression	4.046	4.016	4.856	0.240	+0.505
Neural network (MLP)	4.614	4.534	3.486	0.608	+0.645
Random Forest	4.099	4.009	3.605	0.581	+0.633
XGBoost	4.061	4.131	3.415	0.624	+0.652
HistGradientBoosting	**3.970**	**3.915**	**3.153**	**0.679**	**+0.679**

Table 2: Test-period performance. RMSE in TECU; skill $= 1 - \mathrm{RMSE}_{\mathrm{model}}/\mathrm{RMSE}_{\mathrm{persistence}}$.

All five models beat persistence at every lead time. Skill increases with lead time, from +0.19 at one hour to +0.68 at six hours. This is a physical rather than statistical result: six hours spans a quarter of the diurnal cycle, so the persistence assumption degrades sharply, reaching $R^2 = -2.105$, worse than predicting the sample mean, whereas a model with access to solar geometry does not. An R^2 of 0.68 is a credible value; the circular formulation of Section 5.1 would have produced ≈ 1.0.

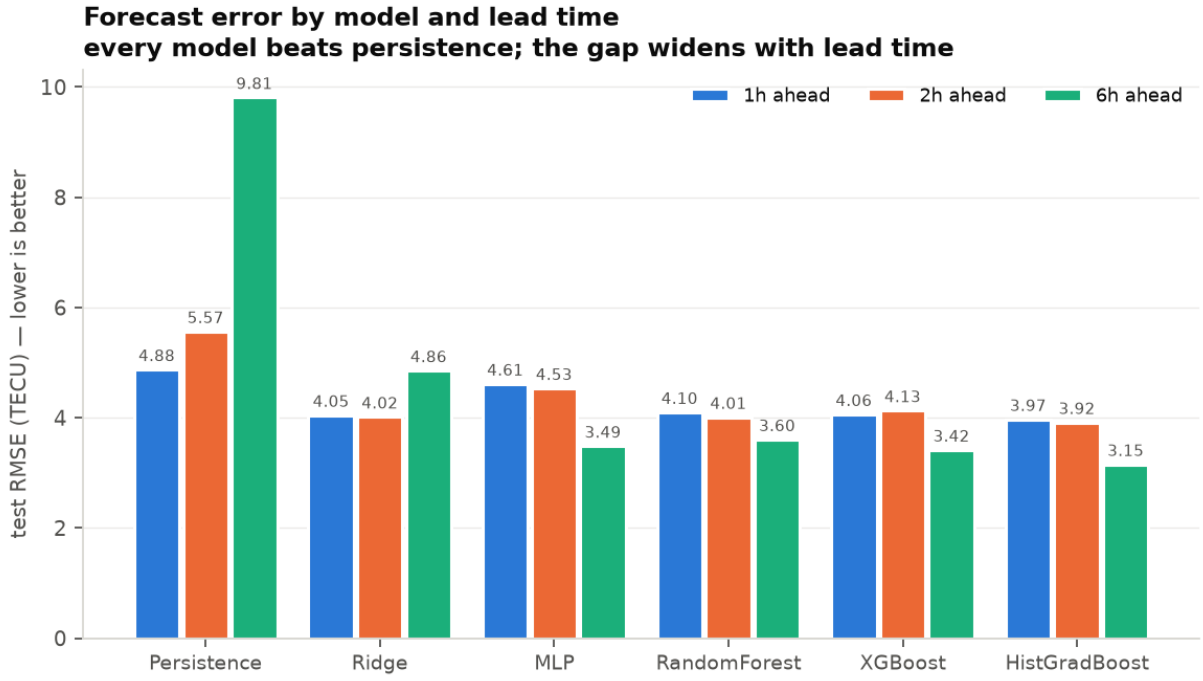


Figure 1: Test RMSE by model and lead time. Persistence degrades sharply with lead time while the learned models do not.

6.2 Where the advantage originates

Lead	Model wins	Median $\lvert e\rvert$ model	Median $\lvert e\rvert$ pers.	RMSE model	RMSE pers.
1 h	51.9%	1.74	1.70	4.00	4.88
2 h	57.4%	2.05	2.40	3.97	5.57
6 h	**81.1%**	**1.67**	6.10	3.11	9.81

Table 3: Case-by-case comparison. Aggregate RMSE conceals the behaviour at short lead.

At one-hour lead the model is only marginally better than chance on a case-by-case basis, and its *typical* error slightly exceeds that of persistence. Its superior RMSE arises entirely from suppressing large errors; at short lead it is best characterised as a variance-reducer rather than an improved point forecast. At six hours this reverses decisively. Restricted to the storm interval (Kp ≥ 5, $n = 164$ per horizon), errors approximately double and the model loses more often than it wins at short lead (40.9% at 1 h, 46.3% at 2 h, 70.7% at 6 h), while retaining lower RMSE throughout for the same reason.

6.3 Learned feature importance

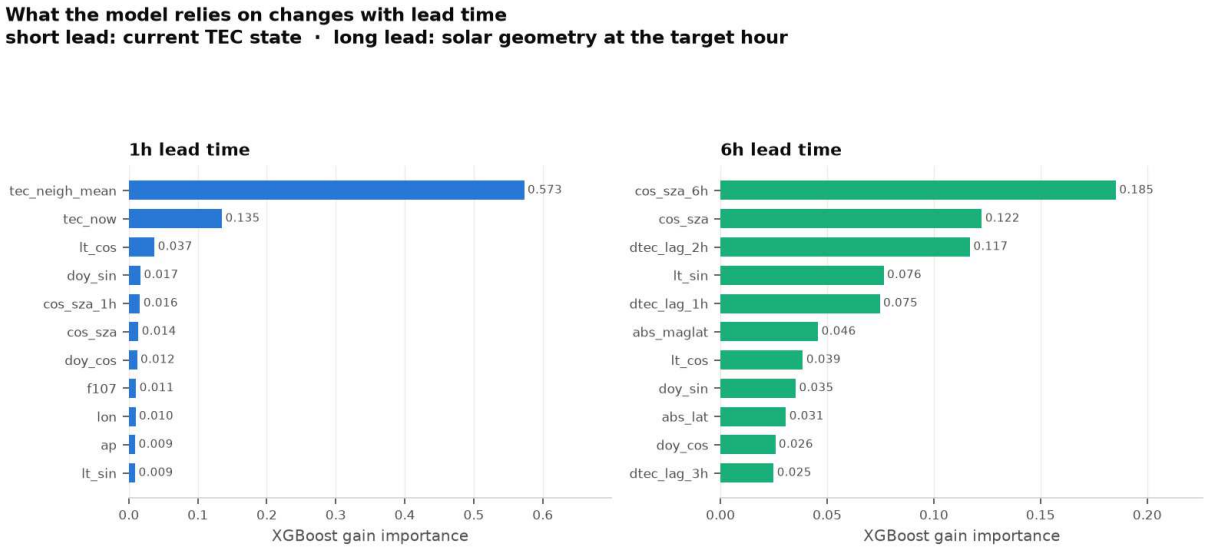


Figure 2: Feature importance at 1 h and 6 h lead. The ranking inverts: current ionospheric state dominates at short lead, solar geometry at the target epoch at long lead.

At one-hour lead the top features are the neighbouring-cell TEC mean (0.573) and local TEC (0.135). At six hours the ranking inverts and solar zenith angle at the target epoch becomes the single most important feature (0.185), followed by the current zenith angle (0.122) and the TEC tendency over two hours (0.117). The model recovers solar illumination as the dominant physical driver without being supplied that relationship.

6.4 A negative result

No lightning feature appears among the important predictors at any lead time. Lightning activity does not measurably influence TEC at these scales. The defensible claim is therefore that this work forecasts the propagation environment **over lightning-active regions**, where the lightning observations define where and when the forecast is issued, rather than forecasting that environment **from** lightning.

6.5 Verification and data-quality findings

Processed data was validated against freshly downloaded source files. Four defects were identified and corrected: (i) the IONEX convention of writing end-of-day midnight as hour 24, which standard date parsers reject; (ii) a silent factor-of-1000 error in the time axis, arising because `pandas` 3.x `DatetimeIndex` may carry microsecond resolution, which placed the entire series in January 1970 without raising any exception; (iii) double application of the netCDF4 `scale_factor`, which is applied on write as well as on read; and (iv) day-boundary duplicate epochs differing by up to 22.2 TECU, because each daily file's 24:00 map is extrapolated beyond its solution window. After correction the stored cube reproduces freshly downloaded source files exactly (0.0000 TECU) at every interior epoch, and the series is gapless: 17,665 epochs, all 17,664 intervals exactly 15 minutes.

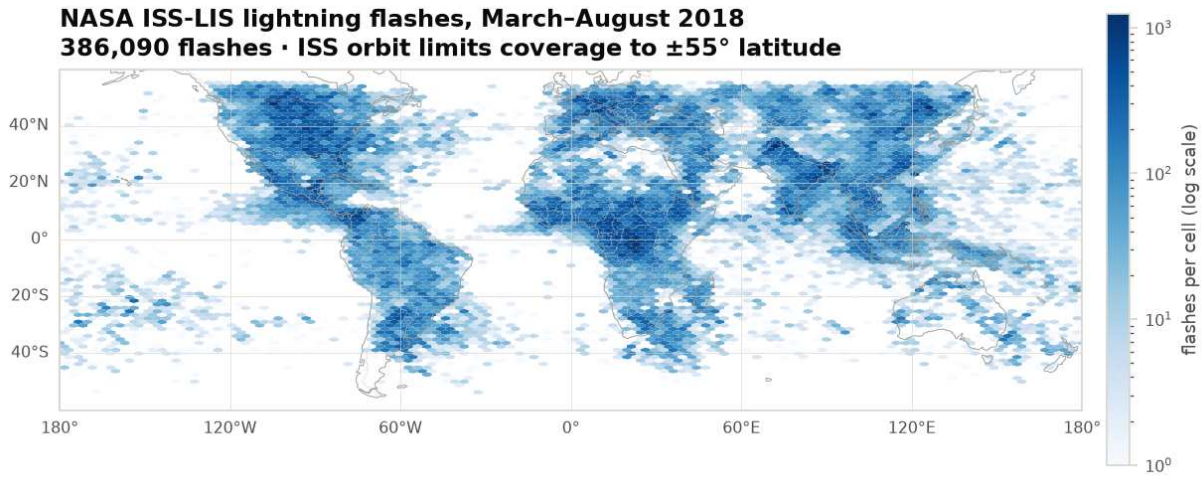


Figure 3: ISS-LIS flash distribution, March–August 2018. The Congo basin, Amazon, Maritime Continent and US Great Plains appear as hotspots and land dominates ocean, reproducing the known global climatology.

6.6 Storm case study

During the 26 August 2018 storm, mean equatorial TEC ($|\text{lat}| \leq 30°$) rises sharply before falling below its pre-storm level. This positive-then-negative phase sequence is the classical storm-time ionospheric response, driven first by storm-time electric fields lifting plasma and subsequently by changes in neutral composition. Its reproduction directly from raw files confirms that the processing chain preserves genuine geophysical signal.

7 Conclusion

A non-circular, physically meaningful machine-learning formulation of the problem was constructed entirely from public NASA data. All five models beat the persistence benchmark; histogram gradient boosting achieved RMSE 3.15 TECU, R^2 0.68 and skill +0.68 at six-hour lead. Forecast skill increases with lead time, as persistence decays across the diurnal cycle while solar geometry does not. The model independently recovered solar illumination as the dominant physical driver. Forecast TEC was converted into f_oF_2, MUF and one-hop ground range. Lightning activity was found not to predict TEC, reported as a genuine negative result.

Is the problem solved? Substantially, with two qualifications stated openly. The title refers to propagation *paths*; what is delivered is the propagation *environment* and the path quantities derived from it, not ray-traced trajectories, which would require a D-region electron-density profile and a ray-tracing model. And because lightning does not predict TEC, the link between storm and forecast is geographic rather than causal.

8 Limitations and future work

1. **TEC is an F2-region quantity while VLF waveguide propagation is governed by the D and E regions.** TEC is a proxy for the propagation environment, not a direct measure of VLF reflection height. This is the most significant caveat of the study. Addressing it would require a lower-ionosphere model.

2. Slab thickness is assumed constant at 300 km; in reality it varies with location, season and solar activity, so f_oF_2 and MUF inherit that systematic uncertainty.

3. Global ionosphere maps are themselves model products interpolated from a finite GNSS network. The processed cube's minimum of −3.6 TECU is unphysical and reflects inversion artefacts.

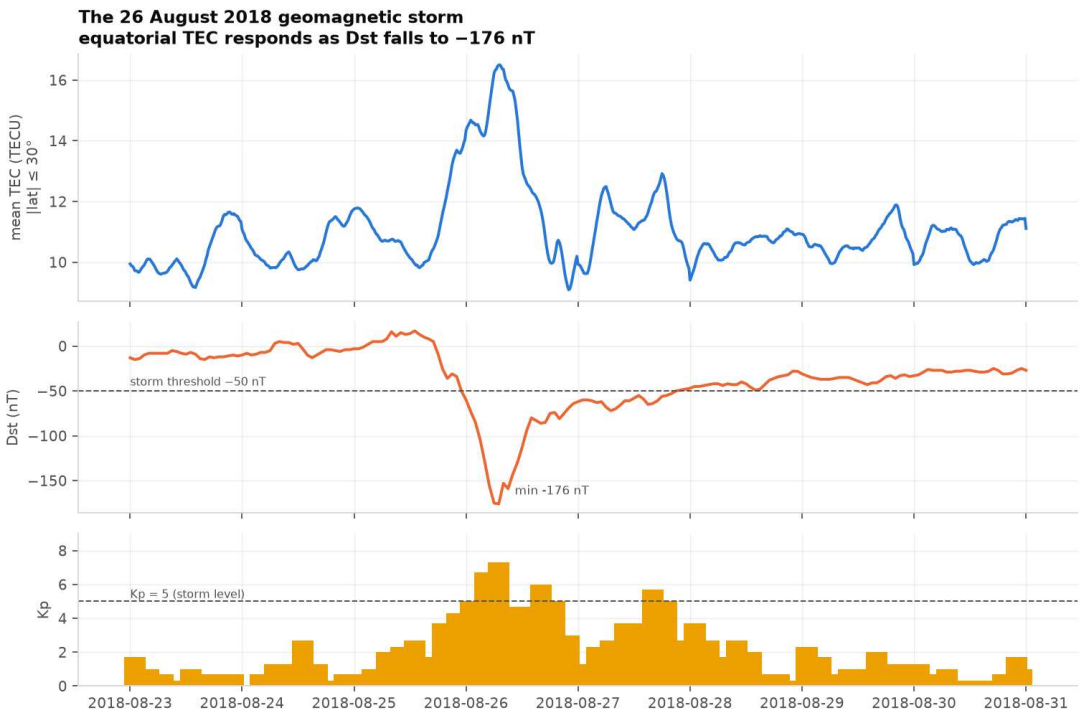


Figure 4: The 26 August 2018 geomagnetic storm. Equatorial TEC rises to ≈ 16.5 TECU (positive phase) then falls below its pre-storm level (negative phase) as Dst reaches −176 nT and Kp 7.3.

4. Coverage is limited to one season of one year. 2018 lay in deep solar minimum (mean F10.7 ≈ 70); solar-maximum conditions, with several times higher TEC, are untested and would likely require retraining.

5. ISS-LIS sampling is intermittent and restricted to ±55° latitude by the ISS orbit, so spatial coverage is non-uniform.

6. At one-hour lead the model functions as a variance-reducer rather than a superior point forecast.

Natural extensions are a live service driven by current NASA feeds rather than archived data, coverage maps produced by running the model across a grid, and validation of the forecast against measured GNSS positioning error.

9 Data and code availability

All data are public. ISS-LIS V3 is distributed by NASA GHRC DAAC; UQRG ionosphere maps by NASA CDDIS; OMNI2 by NASA SPDF. Access to the first two requires a free NASA Earthdata Login. The complete pipeline, the executed notebook, the trained models and the local web application are released with this report.

10 Acknowledgements

I thank Prof. Mayank Goswami for the course and for the project framing. This work uses data from NASA's Global Hydrometeorology Resource Center, the Crustal Dynamics Data Information System and the Space Physics Data Facility, and global ionosphere maps produced by UPC-IonSAT.

11 Video

`https://youtu.be/Zs0UPAre07k`

References

1. NASA GHRC DAAC. *Quality Controlled Lightning Imaging Sensor (LIS) on International Space Station (ISS) Science Data V3* (`isslis_v3_fin`).

2. NASA CDDIS. *UPC-IonSAT UQRG rapid global ionosphere maps*, IONEX format.

3. NASA SPDF. *OMNI2 hourly solar wind and geomagnetic index dataset.*

4. *Lightning Mapping: Techniques, Challenges, and Opportunities.* arXiv:2006.00563.

5. Konan, O.J.E.Y., Mishra, A.K., Lotz, S. (2020). *Machine Learning Techniques to Detect and Characterise Whistler Radio Waves.* arXiv:2002.01244.

6. Cesaroni, C., Spogli, L., Aragón-Àngel, A., Fiocca, M., Dear, V., De Franceschi, G., Romano, V. (2020). *Neural network based model for global Total Electron Content forecasting.* J. Space Weather Space Clim., 10, 11.

7. Liu, L. et al. (2020). *Forecasting Global Ionospheric TEC Using Deep Learning Approach.* Space Weather, 18. doi:10.1029/2020SW002501.

8. Schaer, S., Gurtner, W., Feltens, J. (1998). *IONEX: The IONosphere Map EXchange Format Version 1.* IGS Analysis Center Workshop.

9. Davies, K. (1990). *Ionospheric Radio.* Peter Peregrinus, London.

10. Rakov, V.A., Uman, M.A. (2003). *Lightning: Physics and Effects.* Cambridge University Press.