



Deposited via The University of York.

White Rose Research Online URL for this paper:

<https://eprints.whiterose.ac.uk/id/eprint/239407/>

Preprint:

Wolniewicz, Linnea M., Kelebek, Halil S., Mestici, Simone et al. (2025) Connecting the Dots: A Machine Learning Ready Dataset for Ionospheric Forecasting Models. [Preprint]

<https://doi.org/10.48550/arXiv.2511.15743>

Reuse

Items deposited in White Rose Research Online are protected by copyright, with all rights reserved unless indicated otherwise. They may be downloaded and/or printed for private study, or other acts as permitted by national copyright laws. The publisher or other rights holders may allow further reproduction and re-use of the full text version. This is indicated by the licence information on the White Rose Research Online record for the item.

Takedown

If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

Connecting the Dots: A Machine Learning Ready Dataset for Ionospheric Forecasting Models

Linnea M. Wolniewicz

Department of Information and Computer Science
University of Hawai'i at Mānoa, USA
linneamw@hawaii.edu

Halil S. Kelebek

Department of Engineering Science
University of Oxford, UK
halil@robots.ox.ac.uk

Simone Mestici

Università degli Studi di Roma Sapienza
Rome, Italy
simone.mestici@uniroma1.it

Michael D. Vergalla

Free Flight Research Lab
Sunnyvale, USA
mike@freeflightlab.org

Giacomo Acciarini

European Space Agency (ESA)
giacomo.acciarini@esa.int

Bala Poduval

University of New Hampshire

Olga Verkhoglyadova

NASA Jet Propulsion Laboratory

Madhulika Guhathakurta

NASA Headquarters
madhulika.guhathakurta@nasa.gov

Thomas E. Berger

Space Weather Technology, Research, and Education Center
University of Colorado Boulder
Thomas.Berger@colorado.edu

Atılım Güneş Baydin

Department of Computer Science
University of Oxford, UK
gunes@robots.ox.ac.uk

Frank Soboczinski

Department of Computer Science
University of York & King's College London, UK
frank.soboczinski@york.ac.uk

Abstract

Operational forecasting of the ionosphere remains a critical space weather challenge due to sparse observations, complex coupling across geospatial layers, and a growing need for timely, accurate predictions that support Global Navigation Satellite System (GNSS), communications, aviation safety, as well as satellite operations. As part of the 2025 NASA Heliolab, we present a curated, open-access dataset that integrates diverse ionospheric and heliospheric measurements into a coherent, machine learning-ready structure, designed specifically to support next-generation forecasting models and address gaps in current operational frameworks. Our workflow integrates a large selection of data sources comprising Solar Dynamic Observatory data, solar irradiance indices (F10.7), solar wind parameters (velocity and interplanetary magnetic field), geomagnetic activity indices (Kp, AE, SYM-H), and NASA JPL's Global Ionospheric Maps of Total Electron Content (GIM-TEC). We also implement geospatially sparse data such as the TEC derived from the

World-Wide GNSS Receiver Network and crowdsourced Android smartphone measurements. This novel heterogeneous dataset is temporally and spatially aligned into a single, modular data structure that supports both physical and data-driven modeling. Leveraging this dataset, we train and benchmark several spatiotemporal machine learning architectures for forecasting vertical TEC under both quiet and geomagnetically active conditions. This work presents an extensive dataset and modeling pipeline that enables exploration of not only ionospheric dynamics but also broader Sun-Earth interactions, supporting both scientific inquiry and operational forecasting efforts.

1 Introduction

Modern society is reliant on complex technological infrastructures, such as space-based navigation and communications systems, Low Earth Orbit (LEO) satellite constellations, aviation networks, and power grids, all of which are highly susceptible to disruptions caused by solar activity. Solar flares, coronal mass ejections, and energetic particles not only represent relevant risks to space operations but can also trigger geo-effective disturbances that directly impact life on Earth [1]. The complex coupling between solar activity, the Earth’s magnetosphere and the ionosphere-thermosphere systems drives geomagnetic storm events, capable of disrupting satellite operations, degrading Global Navigation Satellite Systems (GNSS) accuracy, compromising radio communications, and even precipitating power grid blackouts [2, 3, 4].

For these reasons, the past decades have seen a marked increase in missions monitoring near-Earth space. For instance, NASA’s Tandem Reconnection and Cusp Electrodynamics Reconnaissance Satellites (TRACERS) mission [5], launched in 2025, deploys two satellites to study solar wind interactions in the polar cusp, aiming to enhance forecasting of geomagnetic storms. ESA’s forthcoming Vigil mission [6] will be Europe’s first operational space weather satellite, positioned at the Sun–Earth L5 Lagrange point and offering an unprecedented side-view of the Sun enabling early detection of solar events, improve forecast lead times by up to four or five days, and support protection of critical infrastructure.

Other long-term missions such as the Advanced Composition Explorer (ACE), Geotail, IMP and Wind [7, 8, 9] provide continuous measurements of solar wind and interplanetary magnetic fields near the L1 point. Networks of ground-based GNSS stations and radars offer extensive Total Electron Content (TEC) and plasma measurements [10]. This diverse body of data spanning multiple platforms, temporal cadences, and modalities constitutes a heterogeneous observational corpus that is indispensable for both scientific discovery and operational applications.

With the growing availability of ionospheric observations, the field is becoming increasingly well-suited for machine learning (ML) applications. Yet a major bottleneck is the limited availability of machine learning ready datasets: existing products were not designed with ML workflows in mind. Data sources are heterogeneous in resolution and format, often sparse, and require extensive preprocessing before they can be used effectively for training and evaluation. This lack of standardized, ML-ready datasets slows progress and prevents systematic comparison of models.

To address this gap, we focus on building an ML-ready ionospheric dataset that integrates heterogeneous, multi-source observations, harmonizes temporal and spatial scales, and is tailored to the needs of data-driven modeling. This dataset provides the foundation upon which advanced ML architectures can be developed, tested, and benchmarked for global-scale ionospheric nowcasting and forecasting. Ultimately, creating robust ML-ready datasets is a necessary step toward building digital twins of the ionosphere and unlocking the full potential of data-driven space weather research.

2 Dataset

Our dataset aligns heterogeneous data from a diverse set of sources, encompassing multiple modalities, cadences, and start/end dates. Data sources include global ionospheric maps (GIMs) [20], both with sparse and dense measurements, Solar Dynamics Observatory (SDO) extreme ultraviolet flux embeddings, solar driver data, and geomagnetic driver data. Our data product aligns these data

Source	Features	Cadence	Date Range	Description
NASA/GSFC OMNI2 via OMNI- Web [7, 9, 8]	AU, AL, AE (nT) SYM-D, SYM-H, ASY-D (nT) $B_x GSE, B_y GSE, B_z GSE$ (nT) Solar wind speed, v_x, v_y, v_z (km/s)	1 min	2010-05-13 - 2024-08-01	Geomagnetic indices related to magnetospheric/ionospheric currents and solar wind/IMF measurements.
NOAA SWPC/GFZ via Celestrak [11]	Ap (nT) Kp (no unit)	3 hours	1997-01-01 - 2025-10-12	Geomagnetic indices describing the state of magnetospheric and ionospheric currents.
JPLD [12, 13, 14]	Dense TEC GIM ($1^\circ \times 1^\circ$ grid TECU)	15 min	2010-05-13 - 2024-07-31	TEC maps generated by JPL based on daily operational GIM processing for the Deep Space Network (DSN).
Madrigal [15]	Sparse TEC GIM ($1^\circ \times 1^\circ$ grid TECU)	5 min	2010-01-01 - 2024-08-01	TEC maps derived from GNSS receivers, hosted by MIT Haystack Observatory.
SDO-FM [16]	EUV irradiance embeddings (unitless)	15 s	2010-05-13 - 2024-08-01	Full solar disk irradiance observations, reduced with NVAE embeddings for ML applications.
Space Environment Technologies [17]	F10.7, S10.7, M10.7, Y10.7 (sfu) JB08 dSt/dt (K)	Daily	1997-01-01 - 2025-10-12	Solar flux indices across multiple wavelengths; includes JB08 thermospheric heating rate.
Orbital Mechanics [18]	Solar and lunar zenith angle ($^\circ$) Subsolar and sublunar point (lat, lon) Solar and lunar antipode point (lat, lon) Earth–Sun (AU) and Earth–Moon (LD) dist.	Variable	Variable	Geometrical features derived from solar and lunar positions relative to Earth. Computed with the same temporal resolution as input features.
Quasi Dipole [19]	Latitude and longitude in quasidipole frame	Yearly	2010-01-01 - 2024-12-31	Yearly quasidipole maps: projection of Earth’s magnetic field lines into a geographic reference frame.

Table 1: Summary of data sources, their channels, cadence, date ranges available in the data product and descriptions. The date ranges for certain datasets were selected to match the date range of SDO [16].

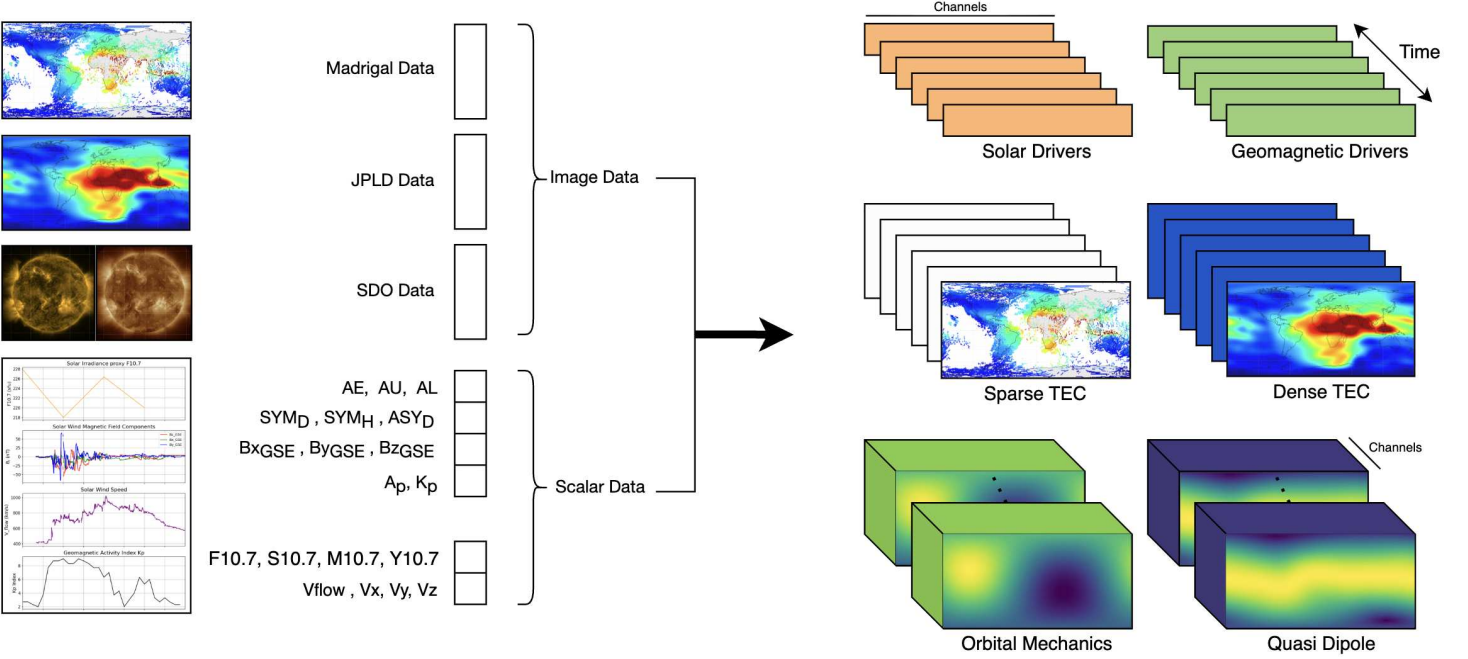


Figure 1: Visualization of dataset inputs and alignment in time and dimension. Output dataset incorporates solar and geomagnetic driver data, sparse and dense TEC maps, and orbital mechanics and quasi-dipole data calculated over a latitude-longitude grid.

sources in time and incorporates relevant orbital mechanics and quasi-dipole features. The aligned data product is visualized in Figure 1, and is publicly available as a Google Cloud bucket ¹.

We align our dataset to the start and end dates of the SDO Foundation Model [16], which are 2010-05-13T00:00:00 to 2024-08-01T00:00:00. Our dataset will be available with multiple cadences according to the dataset sources detailed in Table 1. Multimodal data is processed in our codebase, provided publicly on GitHub², which also handles the alignment in time and processing of data features. The dataset is structured and queried by time.

As an additional product, we provide an event catalog which uses a simple threshold on the Kp time series to identify periods of enhanced geomagnetic activity. In particular, this catalog divides the entire time interval into sub-intervals associated with a specific geomagnetic storm flag using the NOAA G-levels. This classification criteria also take into account the duration of the event period (see Table 2). A schematic view of the event distribution in the considered time interval is shown in Appendix 2. This physics-based classification was prepared to ensure proper model validation and mitigate data leakage (where portions of the same geomagnetic storm event are scattered across training and validation sets) for the IonCast model [21], which aim to accurate forecast TEC global maps for all geomagnetic conditions.

3 Challenges

A key challenge in constructing our dataset is the presence of missing values and inconsistent cadences across the underlying data streams, posing a challenge in temporally aligning the datasets. Different sources also adopt non-standard conventions for encoding missing values. For example, the OMNI dataset [7, 9, 22] marks gaps with sentinel values that differ over channels, making it difficult to detect. To standardize across datasets, we represent all missing values as NaNs. The OMNI dataset also contains certain features with multiple years of data missing. Any columns containing major gaps were removed. To deal with small holes, we use a simple forward-filling approach to fill in

¹https://console.cloud.google.com/storage/browser/ionosphere_data_public

²<https://github.com/FrontierDevelopmentLab/2025-HL-Ionosphere>

Event ID	NOAA G-Level (Kp)	Duration (hours)
G0H ℓ	$K_p < 5$ (Calm)	ℓ
G1H ℓ	$5 \leq K_p < 6$ (Minor)	ℓ
G2H ℓ	$6 \leq K_p < 7$ (Moderate)	ℓ
G3H ℓ	$7 \leq K_p < 8$ (Strong)	ℓ
G4H ℓ	$8 \leq K_p < 9$ (Severe)	ℓ
G5H ℓ	$K_p \geq 9$ (Extreme)	ℓ

Table 2: The geomagnetic storms catalog classification scheme. The event-ID combines NOAA G-levels (defined by Kp) with storm duration ℓ in hours. For example, G2H6 means an event that reached the G2 level lasting at least 6 hours.

short gaps, using the most recent valid sample to fill NaN values. To determine whether to fill or skip a short gap, for each data stream, we define a maximum rewind time. For most data streams, the maximum rewind time is set equal to the native cadence of the dataset, to ensure only minor interruptions are filled. The one exception is OMNI, which has a rewind time of 50 minutes. If a gap exceeds the rewind time, the corresponding timestamps are skipped to avoid propagating stale data. The choice of rewind time is flexible and can be updated by the user if the default values are not suitable for the end user. This same forward-filling logic is also used as a simple interpolation strategy to resample all features to a standard temporal cadence.

4 Baseline Results

Our machine learning ready data product has enabled the training of a suite of global ionospheric forecast models showing promising results on autoregressive forecasts of TEC. These models are trained on a 15-minute cadence aligned data product, with dense TEC maps from JPL serving as prediction targets. Our models, named IonCast [21], outperform baseline persistence TEC forecasts and produce accurate forecasts up to 12-hour lead times. The IonCast models include an LSTM [23] baseline model, a Spherical Neural Operator Model (SFNO) [24], and a GraphCast [25] model inspired by recent advancements in weather modeling.

Our data product is available publicly as a Google Cloud bucket, along with the codebase used to align, process, and split data based on geomagnetic storm events³. Our codebase includes PyTorch datasets that prepare training data according to user-specified start and end ranges, dataset-specific normalization schemes, and example PyTorch model training code.

The growing availability of data in the fields of space weather and heliophysics has made it possible to curate large datasets of heterogeneous data sources for machine learning model training. Yet, no existing resource aligns ionospheric TEC maps (both sparse and dense) with solar and geomagnetic driver data to enable ionospheric modeling with machine learning. To this end, we present a novel dataset that integrates data from diverse modalities, sources, and cadences into a single, machine-learning-ready product.

Acknowledgments

This research is the result of the Frontier Development Lab, Heliolab a partnership between NASA, Trillium Technologies Inc. (USA), Google Cloud, NVIDIA and Pasteur Labs, Contract No. 80GSFC23CA040. A portion of research was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with NASA. The authors thank Andrew Smith and Umaa D. Rebbapragada for their valuable insights, NASA’s Goddard Space Flight Center, and NASA’s Jet Propulsion Laboratory for their continuing support.

³https://console.cloud.google.com/storage/browser/ionosphere_data_public

References

- [1] Thomas E Berger, MJ Holzinger, EK Sutton, and JP Thayer. Flying through uncertainty. *Space Weather*, 18(1):e2019SW002373, 2020.
- [2] Jr. Kintner, P. M. Observations of velocity shear driven plasma turbulence. *Journal of Geophysical Research*, 81(A28):5114–5122, October 1976.
- [3] Ryuho Kataoka, Daikou Shiota, Hitoshi Fujiwara, Hidekatsu Jin, Chihiro Tao, Hiroyuki Shinagawa, and Yasunobu Miyoshi. Unexpected space weather causing the reentry of 38 starlink satellites in february 2022. *Journal of Space Weather and Space Climate*, 12:41, 2022.
- [4] Antti Pulkkinen, E Bernabeu, A Thomson, A Viljanen, R Pirjola, D Boteler, J Eichner, PJ Cilliers, D Welling, NP Savani, et al. Geomagnetically induced currents: Science, engineering, and applications readiness. *Space weather*, 15(7):828–856, 2017.
- [5] Steven M Petrinec, CA Kletzing, DM Miles, Stephen A Fuselier, IW Christopher, Danielle Crawford, Sanny Omar, Scott R Bounds, John W Bonnell, Jasper S Halekas, et al. The tandem reconnection and cusp electrodynamics reconnaissance satellites (tracers) mission design. *Space science reviews*, 221(5):1–23, 2025.
- [6] JP Eastwood, P Brown, W Magnes, CM Carr, M Agu, R Baughen, G Berghofer, J Hodgkins, I Jernej, C Möstl, et al. The vigil magnetometer for operational space weather services from the sun-earth l5 point. *Space Weather*, 22(6):e2024SW003867, 2024.
- [7] Edward C Stone, AM Frandsen, RA Mewaldt, ER Christian, D Margolies, JF Ormes, and F Snow. The advanced composition explorer. *Space Science Reviews*, 86(1):1–22, 1998.
- [8] Lynn B Wilson III, Alexandra L Brosius, Natchimuthuk Gopalswamy, Teresa Nieves-Chinchilla, Adam Szabo, Kevin Hurley, Tai Phan, Justin C Kasper, Noé Lugaz, Ian G Richardson, et al. A quarter century of wind spacecraft discoveries, 2021.
- [9] A Nishida, K Uesugi, I Nakatani, T Mukai, DH Fairfield, and MH Acuna. Geotail mission to explore earth’s magnetotail. *Eos, Transactions American Geophysical Union*, 73(40):425–429, 1992.
- [10] Norbert Jakowski, C Mayer, MM Hoque, and V Wilken. Total electron content models and their use in ionosphere monitoring. *Radio Science*, 46(06):1–11, 2011.
- [11] J. Matzka, C. Stolle, Y. Yamazaki, O. Bronkalla, and A. Morschhauser. The Geomagnetic Kp Index and Derived Indices of Geomagnetic Activity. *Space Weather*, 19(5):e2020SW002641, May 2021.
- [12] AJ Mannucci, BD Wilson, DN Yuan, CH Ho, UJ Lindqwister, and TF Runge. A global mapping technique for gps-derived ionospheric total electron content measurements. *Radio science*, 33(3):565–582, 1998.
- [13] Léo Martire, Thomas F Runge, Xing Meng, Siddharth Krishnamoorthy, Panagiotis Vergados, Anthony J Mannucci, Olga P Verkhoglyadova, Attila Komjáthy, Angelyn W Moore, Robert F Meyer, et al. The jpl-gim algorithm and products: multi-gnss high-rate global mapping of total electron content. *Journal of Geodesy*, 98(5), 2024.
- [14] Olga Verkhoglyadova and Xing Meng. Global ionospheric maps for research – jpld data product. https://sideshow.jpl.nasa.gov/pub/iono_daily/gim_for_research/jpld/, April 2024. Last updated: 8 Apr 2024. Government sponsorship acknowledged.
- [15] Cariglia K. Rideout W. The open madrigal initiative.
- [16] James Walsh, Daniel G Gass, Raul Ramos Pollan, Paul J Wright, Richard Galvez, Noah Kasmanoff, Jason Naradowsky, Anne Spalding, James Parr, and Atılım Güneş Baydin. A foundation model for the solar dynamics observatory. *arXiv preprint arXiv:2410.02530*, 2024.
- [17] W Kent Tobiska, BR Bowman, and SD Bouwer. Solar and geomagnetic indices for thermospheric density models. *COSPAR International Reference Atmosphere, edited by Rees D. and Tobiska WK*, 2012.

- [18] JD Giorgini, DK Yeomans, AB Chamberlin, PW Chodas, RA Jacobson, MS Keesey, JH Lieske, SJ Ostro, EM Standish, and RN Wimberly. Jpl's on-line solar system data service. In *AAS/Division for Planetary Sciences Meeting Abstracts# 28*, volume 28, pages 25–04, 1996.
- [19] Karl Magnus Laundal and Arthur D Richmond. Magnetic coordinate systems. *Space science reviews*, 206(1):27–59, 2017.
- [20] Manuel Hernández-Pajares, JM Juan, J Sanz, R Orus, A Garcia-Rigo, J Feltens, A Komjathy, SC Schaer, and A Krankowski. The igs vtec maps: a reliable source of ionospheric information since 1998. *Journal of Geodesy*, 83(3):263–275, 2009.
- [21] Halil S. Kelebek, Linnea M. Wolniewicz, Michael D. Vergalla, Simone Mestici, Giacomo Acciarini, Bala Poduval, Umaa Rebbapragada, Olga Verkhoglyadova, Madhulika Guhathakurta, Thomas Berger, Frank Soboczinski, and Atılım Güneş Baydin. Ioncast: A deep learning framework for forecasting ionospheric dynamics. In *Proceedings of the Machine Learning for the Physical Sciences Workshop at the 39th Conference on Neural Information Processing Systems (NeurIPS 2025)*, Vancouver, Canada, 2025. Neural Information Processing Systems Foundation. Accepted; to appear.
- [22] JH King and NE Papitashvili. Solar wind spatial scales in and comparisons of hourly wind and ace plasma and magnetic field data. *Journal of Geophysical Research: Space Physics*, 110(A2), 2005.
- [23] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 11 1997.
- [24] Boris Bonev, Thorsten Kurth, Ankur Mahesh, Mauro Bisson, Jean Kossaifi, Karthik Kashinath, Anima Anandkumar, William D. Collins, Michael S. Pritchard, and Alexander Keller. Four-castnet 3: A geometric approach to probabilistic machine-learning weather forecasting at scale, 2025.
- [25] Remi Lam, Alvaro Sanchez-Gonzalez, Matthew Willson, Peter Wirsberger, Meire Fortunato, Ferran Alet, Suman Ravuri, Timo Ewalds, Zach Eaton-Rosen, Weihua Hu, Alexander Merose, Stephan Hoyer, George Holland, Oriol Vinyals, Jacklynn Stott, Alexander Pritzel, Shakir Mohamed, and Peter Battaglia. Learning skillful medium-range global weather forecasting. *Science*, 382(6677):1416–1421, 2023.

Appendix

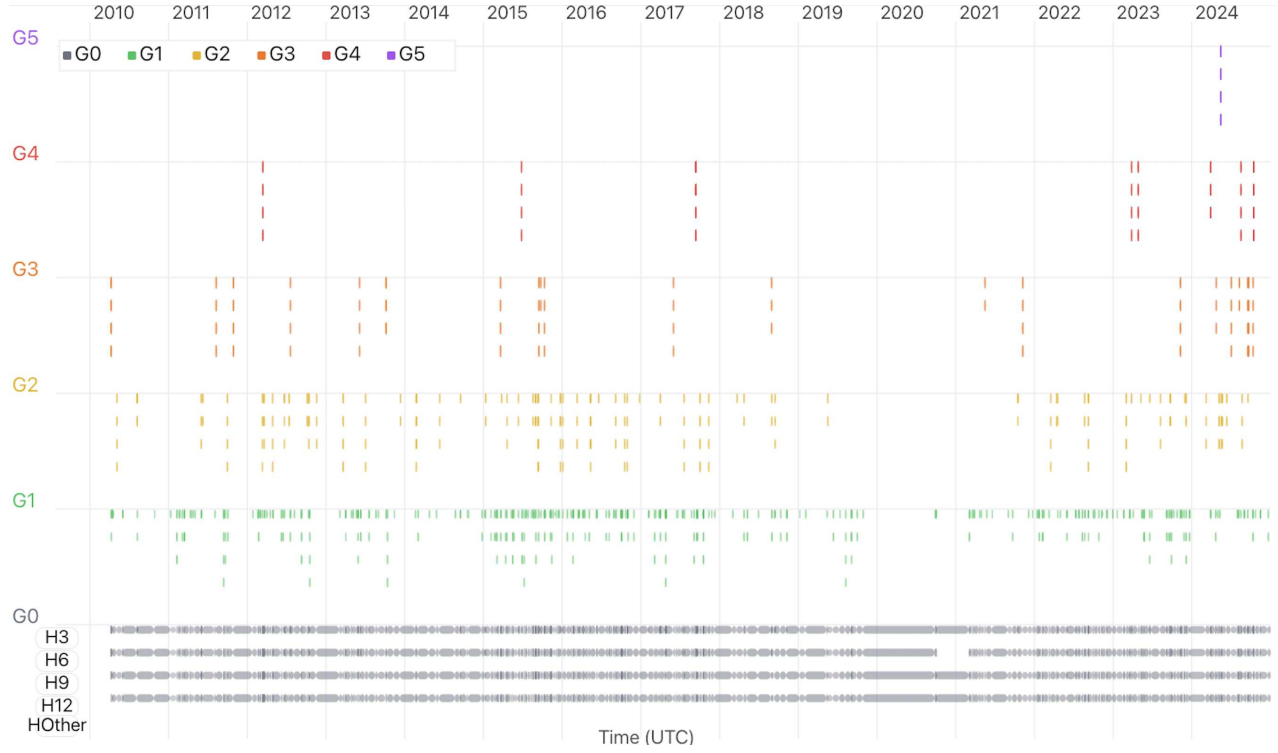


Figure 2: Visualization of the 'Monitoring Event Space-weather TEC Ionospheric Catalog Index' (the MESTICI scale) showing temporal distribution of the Event class for the entire dataset time interval (2010-2024). The x and y axes represent the time (years) and the intensity of the event (G-level), respectively. Each class bin in the y-axis is then divided into four segments, which correspond to the event duration, as shown in the lower part of the plot.