Quantification of Effect of Solar Storms on TEC over U.S. sector Using Machine Learning

Disha Sardana

Thesis submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Master of Science
in
Electrical Engineering

Gregory D. Earle, Chair
J. Michael Ruohoniemi
Scott M. Bailey

April 23, 2018
Blacksburg, Virginia

Keywords: Space Weather, Solar Storms, Data Analysis, Machine Learning

Quantification of Effect of Solar Storms on TEC over U.S. sector Using Machine Learning

Disha Sardana

ABSTRACT

A study of large solar storms in the equinox periods of solar cycles 23 & 24 is presented to quantify their effects on the total electron content (TEC) in the ionosphere. We study the dependence of TEC over the contiguous US on various storm parameters, including the onset time of the storm, the duration of the storm, its intensity, and the rate of change of the ring current response. These parameters are inferred autonomously and compared to TEC values obtained from the CORS network of GPS stations. To quantify the effects we examine the difference between the storm-time TEC value and an average from 5 quiet days during the same month. These values are studied over a grid with 1 deg x 1 deg spatial resolution in latitude and longitude over the US sector. Correlations between storm parameters and the quantified delta TEC values are studied using machine learning techniques to identify the most important controlling variables. The weights inferred by the algorithm for each input variable show their importance to the resultant TEC change. The results of this work are compared to recent TEC studies to investigate the effects of large storms on the distribution of ionospheric density over large spatial and temporal scales.

Quantification of Effect of Solar Storms on TEC over U.S. sector Using Machine Learning

Disha Sardana

GENERAL AUDIENCE ABSTRACT

This study analyzes the impact of geomagnetic storms on the electrical properties of the upper atmosphere at altitudes where satellites routinely fly. The storms are caused by bursts of charged particles from the sun entering the Earth's atmosphere at high latitudes, leading to phenomena like the aurora. These fluctuations in the atmospheric electrical properties can potentially have serious consequences for the electrical power grid, the communications infrastructure, and various technological systems.

Given the risks solar storms can pose, it is important to predict how strong the impact of a given storm is likely to be. The current study applies machine learning techniques to model one particular parameter that relates to the electrified atmosphere over the contiguous US sector. We quantify the strength of the fluctuations as a function of various storm parameters, including onset time and duration. This enables us to autonomously infer which storm parameters have the most significant influence on the resultant atmospheric changes, and compare our results to other recent studies.

Acknowledgments

I am very grateful to Dr. Gregory Earle for his immense support throughout my master's. Right from the beginning he has been very supportive, always letting me take time with learning things, and with patience stood by me no matter how long it took.

I want to thank Shantanab Debchoudhury and Lee Kordella for being there during the initial phases and encouraging me to work at Space@VT and for being great colleagues.

I want to thank Dr. Scott Bailey and Dr. Michael Ruohoniemi for their insightful feedback on my work, and challenging me to see things from a different perspective.

I want to thank Mankaran Singh Chhatwal for helping me out with parallel computing, and for the brain storming sessions on machine learning.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The ionosphere is defined as the layer of the Earth's atmosphere that is ionized by solar and cosmic radiation. It lies 75-1000 km (46-621 miles) above the Earth. It is composed of three main parts: the D, E, and F regions. Figure 1.1 shows the normal mid-latitude electron density of the ionosphere as a function of altitude for daytime and nighttime conditions. The electron density is highest in the upper, or F region, which exists during both daytime and nighttime. During the day it is ionized by solar radiation, during the night it decays slowly, and continues to be weakly ionized by cosmic rays. The D region disappears during the night compared to the daytime, and the E region becomes weakened.

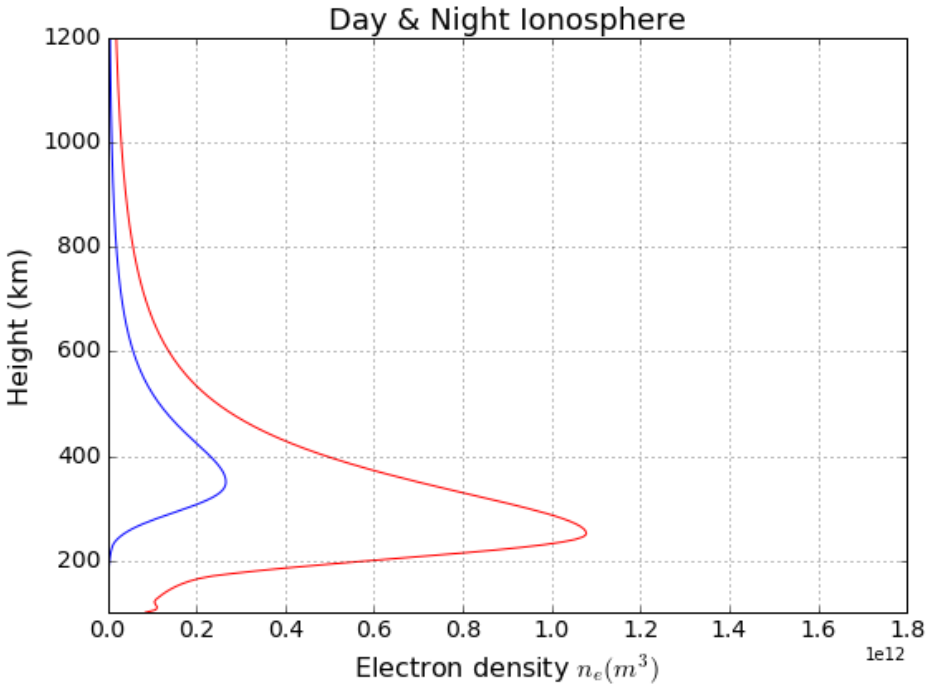


Figure 1.1: The electron density is plotted at location (40°N, 70°W) as a function of height using the International Reference Ionosphere - IRI-2016 model. Red color shows daytime ionosphere at 14 LT and blue color shows nighttime ionosphere at 2 LT on October 28, 2011.

Total Electron Content

Total electron content (or TEC) is an important descriptive quantity for the ionosphere of the Earth. TEC is the total number of electrons integrated between two points (i.e. between the receiver and satellite), along a vertical column of one meter squared cross section. TEC is measured in electrons per square meter. By convention, 1 TEC unit (TECU) = 10^{16} electrons/m^2. By graphically depicting the variations in TEC across broad geographic areas it is possible to identify large-scale ionospheric responses to geophysical events. [1]

Formulation

TEC is path-dependent. By definition, it can be calculated by integrating along the path ds through the ionosphere with the location-dependent electron density $n_e(s)$:

$$\text{TEC} = \int n_e(s)\, ds \tag{1.1}$$

Vertical TEC (VTEC) is obtained by integration of the electron density on a path perpendicular to the ground, while slant TEC (STEC) is determined by integrating over any other straight path. Vertical TEC values in Earth's ionosphere can range from a few to several hundred TECU.

For applications like ground-to-satellite communication and satellite navigation, TEC is a convenient parameter to monitor for possible space weather impacts. [2] The TEC in the ionosphere is known to be modified by changing solar extreme ultra-violet (EUV) radiation, geomagnetic storms, the solar wind and waves that propagate up from the lower atmosphere. The TEC will therefore depend on local time, latitude, longitude, season, geomagnetic conditions, solar cycle and activity, and tropospheric condition.

The propagation of radio waves is affected by the ionosphere. The velocity of radio waves changes when the signal passes through the electrons in the ionosphere. [3] The total delay experienced by a radio wave propagating through the ionosphere is dependent both on the TEC and the frequency of the radio wave between the transmitter and the receiver. [4] At some frequencies the radio waves pass through the ionosphere. At other frequencies, the waves are strongly refracted by the ionosphere. [5] Ham radio and over-the-horizon radar use these refractive effects to communicate and to probe beyond the horizon. The change in the path and velocity of radio waves in the ionosphere has a big impact on the accuracy of satellite navigation systems such as GPS/GNSS. [6] Neglecting changes in the ionosphere TEC can introduce tens of meters of error in the position calculations. [7]

The high density of GPS receivers in first-world countries has created a wealth of TEC data that can help identify the ionospheric phenomenon that affect TEC. [8] In addition, TEC fluctuations can be correlated with other geophysical parameters to gain insight

into possible cause-and-effect scenarios. The study presented here is a first step toward understanding how a particular source of ionospheric perturbations (geomagnetic storms) influences TEC in the US sector. Future extensions of the work presented here may eventually lead to better understanding of the ionosphere and the ways in which it affects GPS and other technological systems.

Geomagnetic Storms

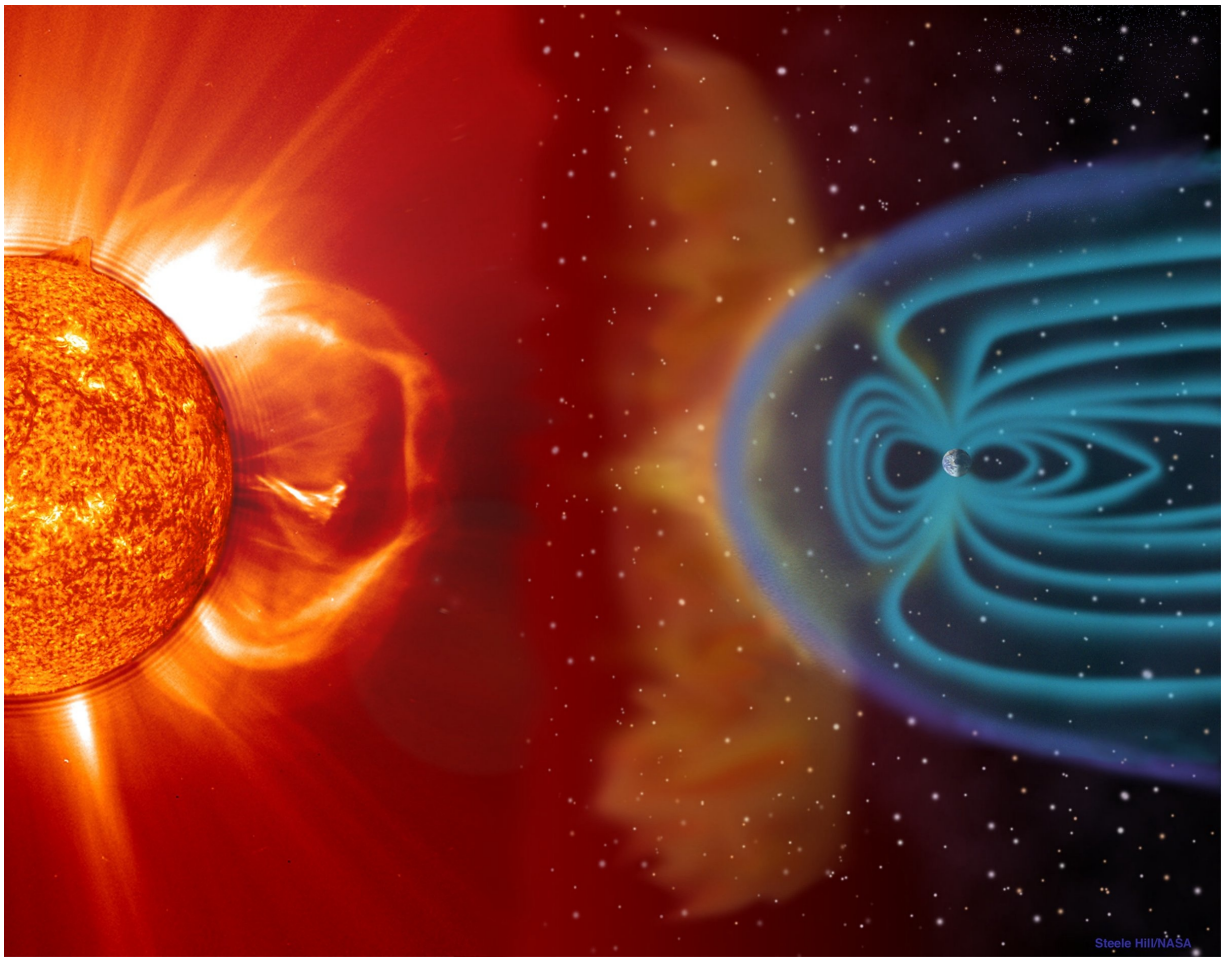


Figure 1.2: Solar storm hitting Earth's magnetosphere. The orange color represents plasma flowing outward from the sun, and the blue region shows the bow shock and magnetic field geometry under normal conditions. **Image Credit: NASA**

"The radiation belts are regions of near-Earth space where charged particles become trapped on geomagnetic field lines." The trapped particles drift around the Earth - positively charged particles travel westward and negatively charged particles drift eastward - and create the ring current which is a permanent feature of the magnetosphere. Figure 1.2 shows a solar image during an active period, and Earth's large scale magnetic field geometry. Due to certain conditions in the interplanetary environment, the ring current occasionally becomes strongly enhanced. The magnetic field created by the ring current is measurable on the Earth's surface, and so this enhancement can be detected on the Earth's surface as a depression of the geomagnetic field at low- and mid-latitudes. A

global system of magnetometers continuously monitors geomagnetic field changes. If the depression of the field is sufficiently strong we refer to the event as a geomagnetic storm. [9]

A geomagnetic storm (GMS) is one of the space weather phenomena that impact TEC greatly. [10] It is defined as the time interval when an intense and long lasting interplanetary convective electric field leads to substantial energization in the magnetosphere-ionosphere system. The intensification of the ring current is measured and used as a detector and quantifier of storm intensity. The ring current measurement is characterized by a parameter called the disturbance storm index, or Dst. [11] Dst variations are measured in magnetic field units, typically nano-tesla (nT).

In terms of time sequence, a GMS can be described in three phases: the initial, main and recovery phases. The initial phase of a storm is often characterized by a short-duration intensification of Dst, followed by a much longer decrease in Dst to large negative values. Abrupt changes in Dst are called sudden commencements. The main phase of a storm is said to begin when the Dst index assumes negative value. This phase ends when Dst reaches its minimum value. The recovery phase, usually the longest one, is characterized by the returning of Dst to its pre-sudden commencement values. During a GMS, the solar wind plasma may penetrate more easily into the magnetosphere, giving rise to changes in both interplanetary space and the magnetospheric and ionospheric plasma. All the perturbations during GMSs involve energy transfer from the solar wind into the magnetosphere-ionosphere system. Efficiency of the energy transfer process seems to depend on the southward component of the magnetic field and the solar wind speed [12]

The study of geomagnetic storms is one of the main ingredients of space weather. These storms can damage the power distribution network, radio communications and spacecraft. Dst, Kp, ap and AE indices are the four most commonly used geomagnetic indices (GI) to study geomagnetic storms. GMSs are usually classified by the Dst indices as intense storms, moderate storms and weak storms as shown in Table 1.1 . [12]

Storm Classification	**Dst Index**
Intense	≤ -100 nT
Moderate	-100 nT $<$ Dst Index ≤ -50 nT
Weak	> -50 nT

Table 1.1: Classification of geomagnetic storms

In our study, results are compiled and studied using various parameters to characterize the storms such as the duration of the storm, its intensity, the rate of change of the ring current response, and the quantified delta TEC value. Our study focuses on major geomagnetic storms during equinox in solar cycles 23 and 24. In Chapter 2, data sources, selection criteria, and classification schemes for GMSs data are presented. In later sections we describe model building algorithms, error plots, results of correlation studies, and analyses of storms over a 16-year period (2000-2015). Discussion of results and final conclusions are presented in the final chapter.

Chapter 2

Data Description

The effects of magnetic storms on the Earth's ionosphere have been a subject of long and intensive investigation. In this chapter we discuss how these storms affect the TEC in the ionosphere. We quantify this effect by studying TEC maps over the continental U.S. We describe how we obtained the data, how we chose our storms based on geomagnetic indices like Dst and SYM-H, and how we characterized the storms based on their temporal occurrences. Once we have characterized a storm, we explore the TEC data and quantify the effect of the storm on TEC.

The majority of the data that are used for this study is sourced from two databases, the Madrigal database, and the Kyoto database, which comes from Data Analysis Center for Geomagnetism and Space Magnetism at Graduate School of Science, Kyoto University. Both of these datasets are regularly used by the international scientific community.

2.1 Dst Index

"By definition, the Dst index is the longitudinally averaged field depression at low latitudes. It provides a simple measure of the strength of the ring current." [9] It is expressed in nanoteslas (nT) and is based on the average value of the horizontal component of the Earth's magnetic field measured hourly at four near-equatorial geomagnetic observatories. Use of the Dst as an index of storm strength is possible because the strength of the northward magnetic field at ground level at low latitudes decreases in proportion to the energy content of the ring current, which increases during geomagnetic storms. As the ring current intensifies its $\overrightarrow{B}$ field subtracts from the normal geomagnetic field at the surface. Dst measures that perturbation, so it is effectively a measure of ring current magnitude. In the case of a classic magnetic storm, the Dst shows a sudden rise, corresponding to the storm sudden commencement, and then decreases sharply as the ring current intensifies. Once the interplanetary magnetic field turns northward again and the ring current begins to recover, the Dst begins a slow rise back to its quiet time level. [13]

In this study we restrict our attention to large storms, where the Dst index is less than -100 nT. The reasoning behind this decision is that signatures of larger storms should be

easier to identify. We also restrict our study to equinox periods, to minimize biases that could be introduced by seasonal variations, trans-equatorial winds, and other effects of a tilted geographic spin-axis.

2.1.1 Data Exploration and Statistics

Storms are identified on the basis of hourly Dst index values using the Kyoto database. [14] Those having Dst indices less than -100 nT are identified as large storms and are classified on the basis of the year, month and onset times. In total 91 large storms are identified, and the 11-year solar cycle is covered. Figure 2.1 below represents the hourly Dst index in the years 2000-2015. The statistics of this data are shown in Table 2.1 indicating minimum, maximum, mean, median, mode, and standard deviation. The minimum Dst index value is -422 nT indicating a super storm event. The Dst index falls below -100 nT only rarely, implying that intense storms do not occur often. The vast majority of days are what we classify as 'quiet' periods. Figures 2.2, 2.3, 2.4 are bar graphs showing the storm counts classified on the basis of year, month and onset time respectively.

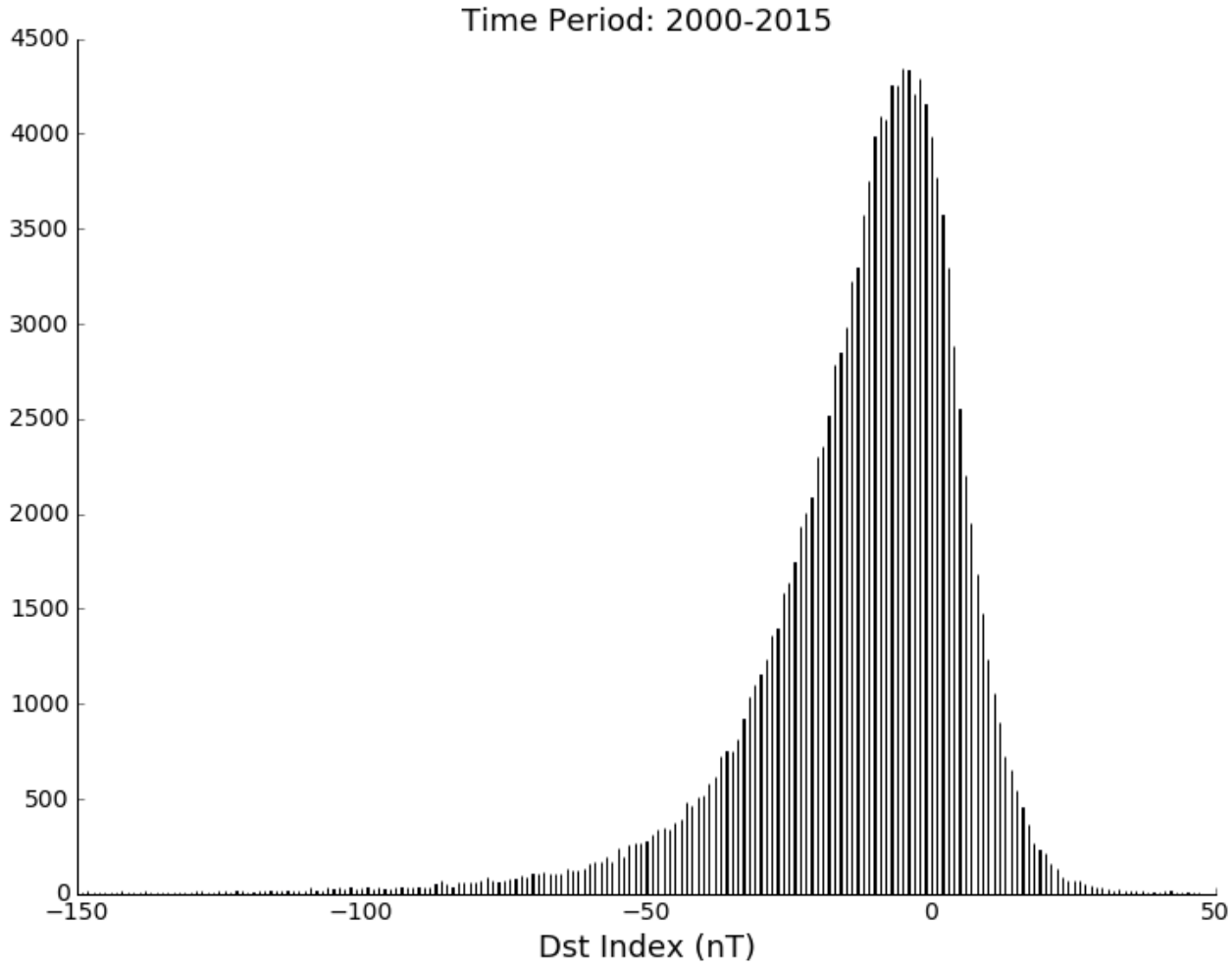


Figure 2.1: The hourly Dst index values during the the time period 2000-2015 are plotted. The Dst index values below -150 nT and above 50 nT are not shown in the plot.

Statistics	**Dst index value (nT)**
Minimum	-422.0
Maximum	77.0
Mean	-12.629
Median	-9.0
Mode	-5.0
Standard Deviation	21.055

Table 2.1: Dst index statistics for the time period : 2000-2015

The main idea is to identify a pattern on the basis of month, year and onset time. In Figure 2.2, it is observed that there are no large storms in years 2007-2010, which is expected since that is a solar minimum phase.

Also, it is observed in Figure 2.3 that the maximum number of large storms occur in October for the years 2000-2015.

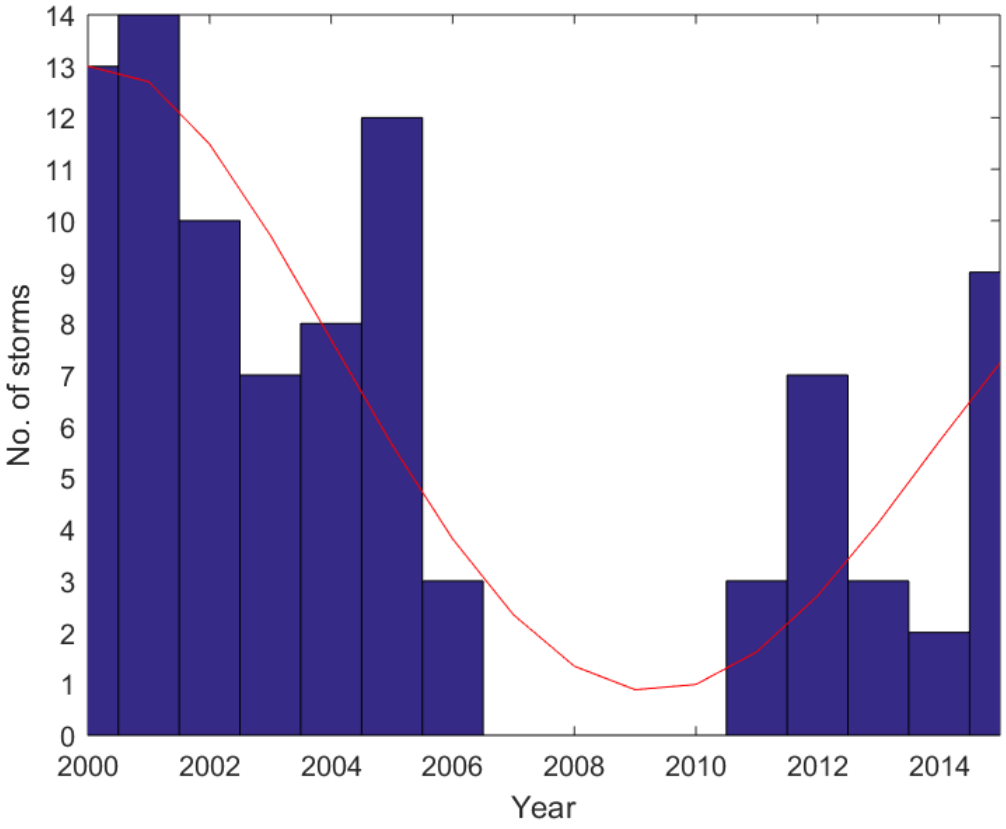


Figure 2.2: Classification of large storms on the basis of year

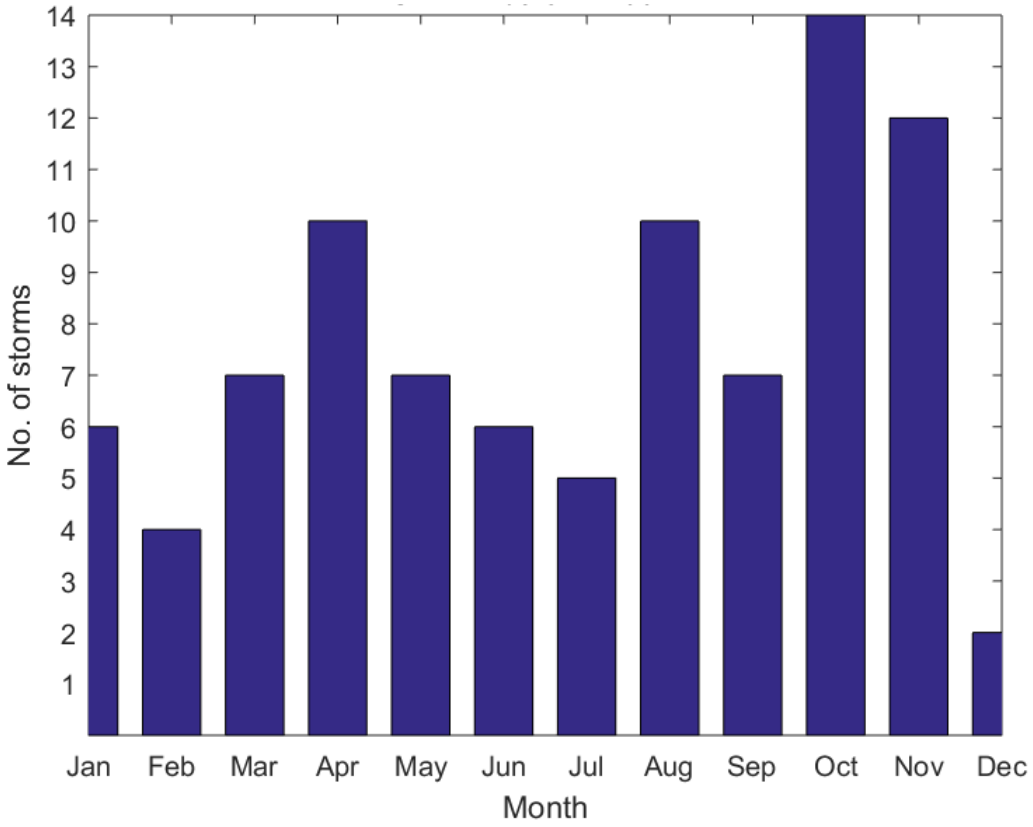


Figure 2.3: Classification of large storms on the basis of month

Figure 2.4 shows the storm count for each hour on the basis of storm onset time (UT). Storms with Dst index $\leq$ -100 but $\geq$ -250 nT, occurring in the equinox periods during the years 2000-2015 are plotted in this Figure. Specifically, equinox interval is referred as the months of February, March, April, August, September, and October. Super storms with Dst index reaching below -250 nT are left out intentionally to avoid any biases and are treated as outliers. A total of 36 storms lie in the specified interval and are shown in Figure 2.4. To summarize, these conditions are explicitly mentioned in Table 2.2

Criteria	**Interval of interest**
Years	2000-2015
Months	Feb, Mar, Apr, Aug, Sep, Oct
Min. Dst index value	$\leq$ -100 nT & $\geq$ -250 nT
Total count	*36*

Table 2.2: Criteria to identify storms of interest

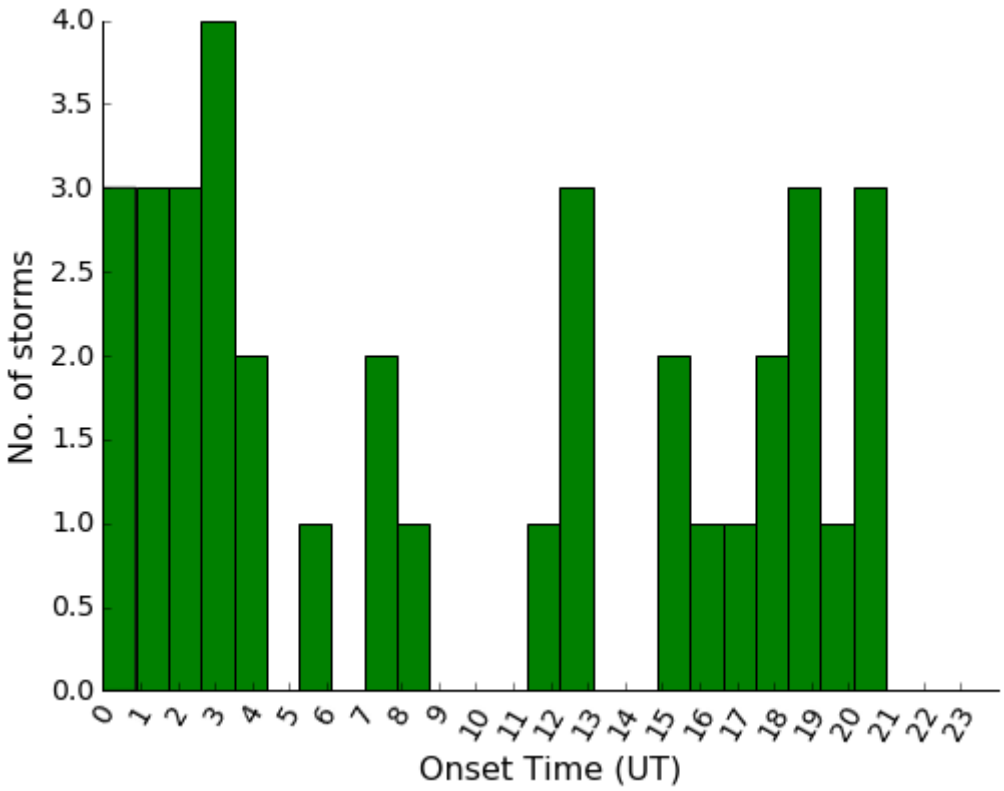


Figure 2.4: Classification of large storms on the basis of onset time in years 2000-2015 in the equinox interval

2.2 SYM-H Index

To describe the geomagnetic disturbance fields in mid-latitudes with high-time (i.e. 1 minute) resolution, a longitudinally asymmetric (ASY) and a symmetric (SYM) disturbance index are introduced and derived for both H and D components. These are defined as the disturbance in the horizontal (dipole pole) direction H (SYM-H, ASY-H) and in the orthogonal (East-West) direction D (SYM-D, ASY-D). [15] The symmetric disturbance field in H, SYM-H, is essentially the same as the hourly Dst index, although 1 minute values from different sets of stations and a slightly different coordinate system are used. Recently many other scientists have begun to use the SYM-H geomagnetic index as a replacement for the classic storm index (Dst) since SYM-H has the distinct advantage of having 1-min time resolution compared to the 1- hour time resolution of Dst. [16]. We use SYM-H to identify and characterize storms for this study, for the same reason. Figure 2.5 shows the SYM-H signature for the major storms in the interval of interest.

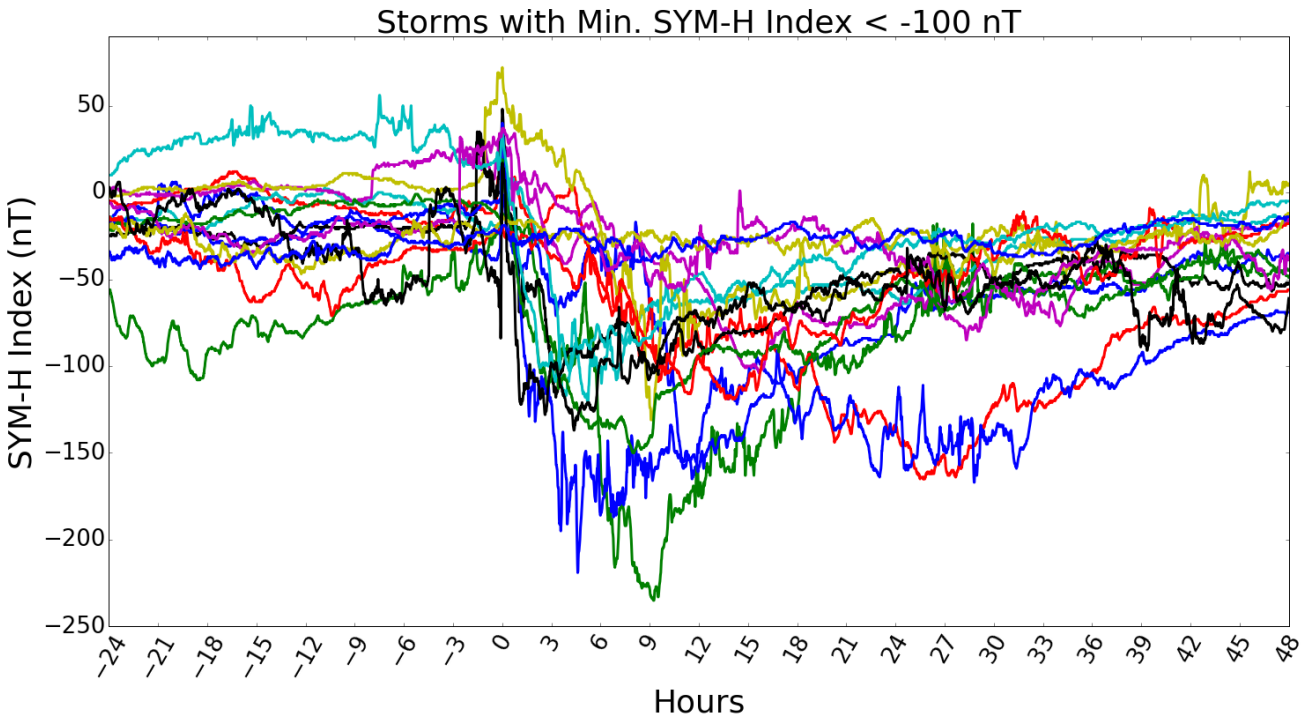


Figure 2.5: Major storms in years 2000-2015 in the equinox interval. Intense storms with minimum SYM-H value below -250 nT are not shown.

2.2.1 Comparison between Dst Index and SYM-H Data

The SYM-H index for years 2000-2015 was procured from the World Data Center for Geomagnetism, Kyoto. Figure 2.6 shows an example of SYM-H data. The parameters of interest for our screening process are date, time and SYM-H (nT).

Format	IAGA-2002
Source of Data	WDC for Geomagnetism, Kyoto
Station Name	ASY/SYM Indices
IAGA CODE	ASY/SYM
Geodetic Latitude	
Geodetic Longitude	
Elevation	
Reported	ASY-D,ASY-H,SYM-D,SYM-H
Sensor Orientation	
Digital Sampling	
Data Interval Type	1-minute
Data Type	Provisional

```
 # Converted to IAGA2002 format                                                  |
 #   by WDC for Geomagnetism, Kyoto. 2017-09-20                                  |
DATE       TIME         DOY     ASY-D     ASY-H     SYM-D     SYM-H  |
2011-10-24 00:00:00.000 297       27.00     11.00     -2.00      4.00
2011-10-24 00:01:00.000 297       27.00     10.00     -2.00      5.00
2011-10-24 00:02:00.000 297       26.00     10.00     -2.00      5.00
2011-10-24 00:03:00.000 297       26.00     10.00     -2.00      5.00
2011-10-24 00:04:00.000 297       25.00     10.00     -2.00      5.00
2011-10-24 00:05:00.000 297       25.00      9.00     -2.00      4.00
2011-10-24 00:06:00.000 297       25.00     11.00     -2.00      4.00
2011-10-24 00:07:00.000 297       26.00     12.00     -3.00      3.00
2011-10-24 00:08:00.000 297       27.00     13.00     -3.00      2.00
2011-10-24 00:09:00.000 297       28.00     13.00     -3.00      1.00
2011-10-24 00:10:00.000 297       29.00     14.00     -4.00      1.00
2011-10-24 00:11:00.000 297       29.00     15.00     -4.00      1.00
2011-10-24 00:12:00.000 297       31.00     16.00     -4.00      0.00
2011-10-24 00:13:00.000 297       33.00     17.00     -4.00     -1.00
2011-10-24 00:14:00.000 297       32.00     16.00     -4.00     -1.00
2011-10-24 00:15:00.000 297       32.00     16.00     -4.00     -1.00
2011-10-24 00:16:00.000 297       29.00     15.00     -3.00     -2.00
2011-10-24 00:17:00.000 297       26.00     14.00     -3.00     -2.00
2011-10-24 00:18:00.000 297       27.00     15.00     -3.00     -3.00
2011-10-24 00:19:00.000 297       30.00     15.00     -4.00     -4.00
2011-10-24 00:20:00.000 297       30.00     16.00     -4.00     -4.00
2011-10-24 00:21:00.000 297       32.00     17.00     -4.00     -6.00
```

Figure 2.6: SYM-H data with 1 minute resolution

WDC for Geomagnetism, Kyoto
Hourly Equatorial Dst Values (FINAL)
OCTOBER 2011

unit=nT

DAY	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	UT 24
1	-13	-18	-13	-19	-29	-41	-36	-35	-39	-37	-33	-28	-24	-26	-27	-27	-24	-20	-18	-13	-24	-30	-32	-25
2	-20	-21	-20	-19	-17	-20	-24	-30	-33	-27	-26	-28	-32	-31	-26	-24	-23	-22	-20	-20	-25	-27	-24	-20
3	-17	-18	-18	-14	-11	-12	-13	-15	-18	-23	-24	-23	-23	-20	-17	-21	-20	-20	-18	-19	-21	-20	-18	-21
4	-23	-21	-19	-18	-17	-15	-14	-18	-22	-22	-20	-16	-14	-12	-18	-15	-17	-17	-20	-21	-21	-21	-16	-13
5	-12	-12	-12	-14	-15	-12	-12	-6	0	-10	-18	-14	-17	-25	-43	-40	-26	-20	-20	-31	-20	-21	-25	-19
6	-14	-11	-9	-14	0	-3	-7	-3	-2	-4	-3	-7	-9	-8	-11	-15	-19	-20	-21	-21	-22	-20	-20	-18
7	-14	-11	-10	-3	-10	-16	-25	-33	-33	-29	-25	-25	-25	-26	-28	-28	-27	-27	-24	-21	-18	-17	-19	-17
8	-15	-11	-10	-11	-12	-13	-10	-6	-3	-1	-1	2	5	8	10	8	4	5	3	0	-10	-6	-1	-10
9	-17	-25	-24	-24	-25	-36	-42	-42	-41	-40	-42	-44	-39	-31	-30	-29	-26	-20	-16	-18	-17	-15	-20	-19
10	-16	-16	-13	-13	-12	-8	-7	-7	-5	-3	-6	-8	-5	-5	-7	-11	-13	-13	-12	-13	-16	-13	-10	-7
11	-6	-10	-10	-11	-8	-7	-4	-2	3	3	1	0	-2	0	3	-2	-5	-3	-5	-4	-4	-7	-5	-5
12	-1	0	1	-2	-5	-5	-4	-8	-7	-3	-4	-4	-3	-3	-3	-4	-5	-4	-6	-9	-11	-8	-6	-5
13	-3	-3	-7	-7	-12	-15	-17	-17	-19	-20	-19	-13	-12	-10	-9	-7	-6	-5	-4	-5	-7	-9	-10	-7
14	-5	-3	-1	0	-1	-3	-4	-2	-1	-1	-2	-3	-4	-3	-1	-2	-2	-4	-6	-3	-1	-3	-3	-2
15	1	-4	-15	-14	-10	-8	-8	-8	-7	-5	-3	-3	-4	-3	-7	-10	-7	-2	-4	-5	-5	-5	-4	-5
16	-8	-8	-11	-19	-17	-11	-9	-10	-12	-11	-7	-6	-7	-7	-11	-11	-9	-7	-6	-4	-6	-8	-14	-18
17	-22	-18	-13	-13	-11	-11	-10	-10	-9	-8	-8	-8	-9	-9	-8	-9	-9	-7	-6	-5	-5	-7	-8	-14
18	-15	-12	-9	-6	-6	-9	-9	-6	-2	-2	-5	-4	-4	-5	-4	-5	-4	-4	-3	-6	-5	-4	-4	-2
19	0	0	-4	-7	-11	-9	-8	-8	-9	-8	-8	-11	-11	-9	-9	-8	-7	-4	-9	-12	-10	-9	-9	-12
20	-12	-15	-12	-9	-9	-12	-14	-13	-12	-12	-11	-11	-11	-9	-8	-9	-11	-12	-16	-19	-15	-12	-12	-12
21	-16	-19	-17	-16	-17	-23	-24	-22	-21	-17	-15	-15	-14	-13	-13	-13	-13	-14	-13	-14	-15	-15	-16	-19
22	-16	-14	-12	-11	-11	-12	-12	-9	-9	-8	-7	-6	-7	-6	-7	-6	-7	-9	-8	-7	-5	-4	-3	0
23	3	5	5	3	3	2	3	2	1	-1	-4	-5	-7	-6	-6	-4	-1	4	8	7	6	6	4	-2
24	-9	-8	-7	-8	-7	-4	-6	-2	-3	-3	1	4	5	4	1	-1	-1	1	20	18	7	7	-22	-79
25	-121	-147	-126	-127	-131	-134	-131	-117	-101	-86	-79	-68	-62	-61	-59	-57	-59	-57	-60	-57	-59	-60	-58	-55

Figure 2.7: Dst index data with 1 hour resolution

Figure 2.7 shows an example of Dst data. The data are available on an hourly basis.

The numeric data, as shown, are plotted for the month of October 2011. Figures 2.8 and 2.9 show a comparison between SYM-H index and Dst index data for the same range of dates. The SYM-H data obviously have more granularity because of the higher sample rate, but all major features are resolved consistently between the two data sets.

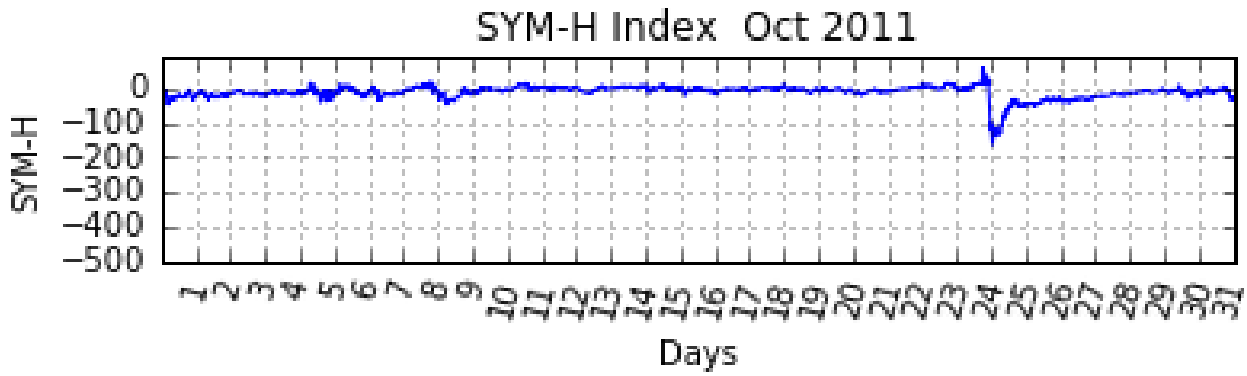


Figure 2.8: SYM-H index plot

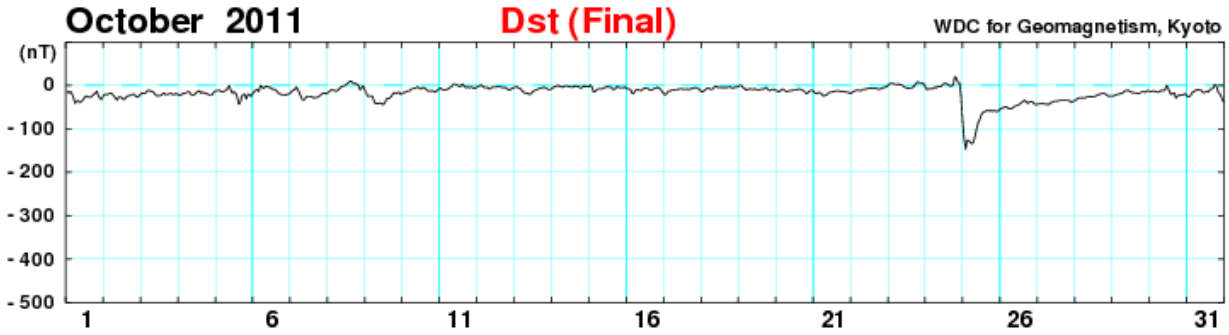


Figure 2.9: Dst index plot

2.2.2 Onset Time

A common set of criteria are used to catalog the onset time of all the storms. First, the data are filtered to identify major storms, and the corresponding dates are stored in a file. SYM-H data are plotted for each storm to provide a visual representation of storm occurrences. Using this database, the exact hour of onset time (in UT) is determined.

As an example, Figures 2.10-2.11 show a GMS occurring on October 24, 2011. In Figure 2.10, there is a sudden spike in the SYM-H index up to 50 nT, and then in the next few hours the index falls to nearly -150 nT, indicating a magnetic storm event. This point of sudden commencement where the SYM-H parameter reaches its maximum positive value

is referred to as the onset time of the storm.

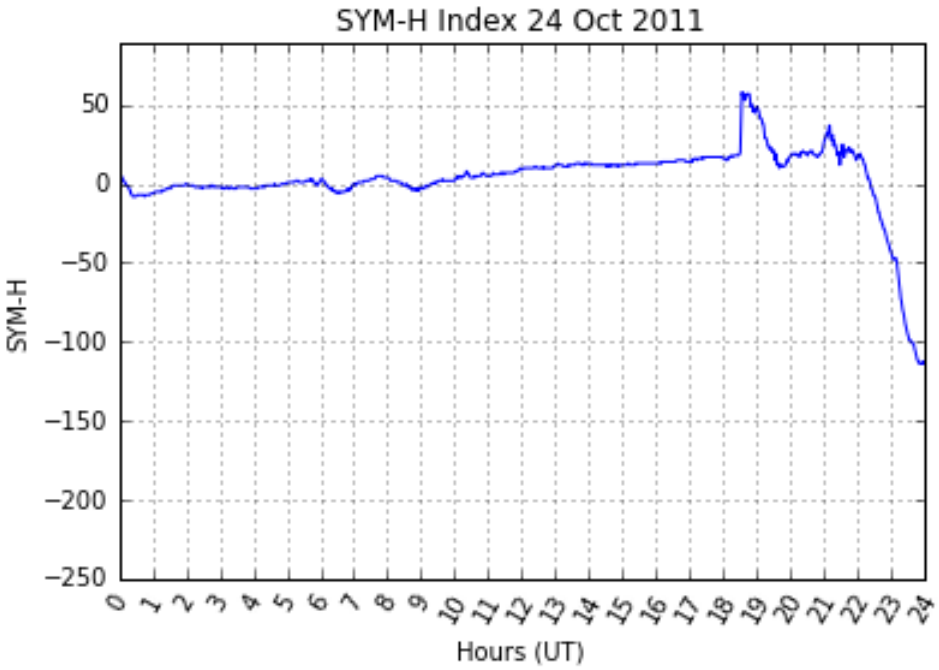


Figure 2.10: SYM-H index on October 24, 2011

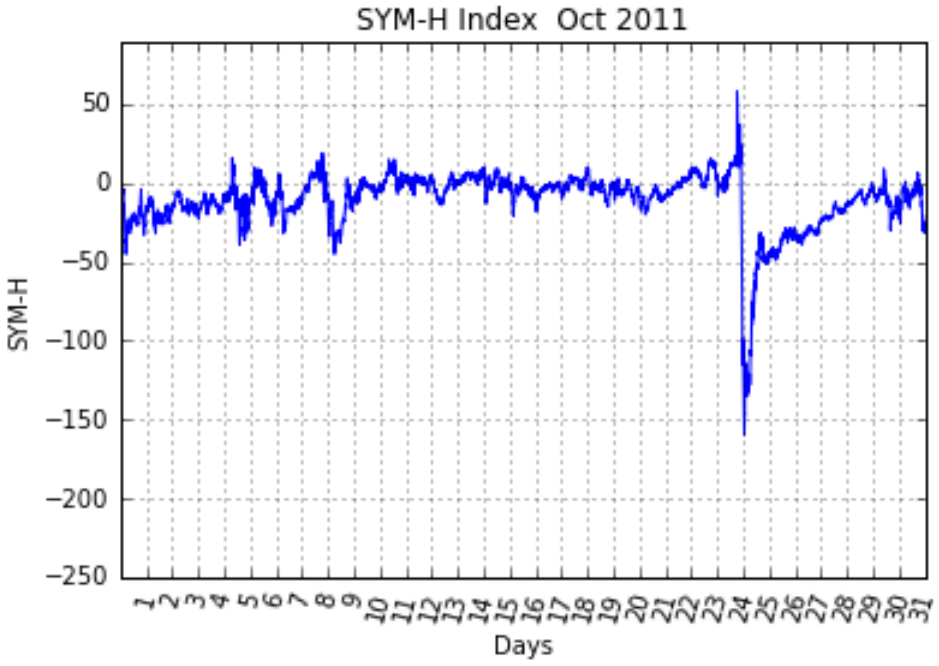


Figure 2.11: SYM-H index in the month of October, 2011

2.2.3 Disturbed Days and Quiet Days

After screening the major storms, we choose quiet days from the same month for comparison. Geomagnetic storms are not the only reason for changes in the TEC level; it can also be affected by seasonal variations, magnetic declination, solar activity, neutral winds, and other geophysical factors. To isolate the impact on TEC that can be attributed to a storm alone, we calculate the difference between TEC levels on a storm day, and the five 'quietest' days in the same month. This ensures we do not confuse effects on TEC from the storm with effects due to unrelated causes.

The Geomagnetic Data Service at Kyoto has a database for "The international 5 and 10 quietest and 5 most disturbed days [1932 -]". [17] The 5 quietest days selected for comparison with storm days are chosen using this database. Figure 2.12 shows an example from the database for the year 2011. The 5 quietest days in the month of October are :- q_1 : 10-28-2011, q_2 : 10-29-2011, q_3 : 10-22-2011, q_4 : 10-14-2011 & q_5 : 10-23-2011.

```
YYYY MM q1q2q3q4q5 q6q7q8q9q0 d1d2d3d4d5
2011 01 3023 52721 2226 1 229  7141319 8
2011 02  3 9271328 2423251722  418 5 614
2011 03 1526162718 2831142925 11 110 2 3
2011 04 2726102816 1723211514 30 612 220
2011 05 20 81225 9 13 6191423 2829 2 131
2011 06 29 3281927 183016 625  5232224 8
2011 07 2728162417 2915 22618 203019 111
2011 08 311819 321  230131211  6 5151423
2011 09 2319 116 8  224221425 1027261712
2011 10 2829221423 1013261118 25 124 5 2
2011 11 1914 92013 1810 5 316  12930 224
2011 12 27162617 6 15 7232418  329101130
```

Figure 2.12: Excerpt of the Kyoto database for the quietest and most disturbed days in each month of 2011

Figures 2.13 and 2.14 show the SYM-H data for a storm day and a quiet day, respectively. As the Figures show, the SYM-H index remains fairly constant on quiet days, while it can significantly deviate from the baseline on storm days.

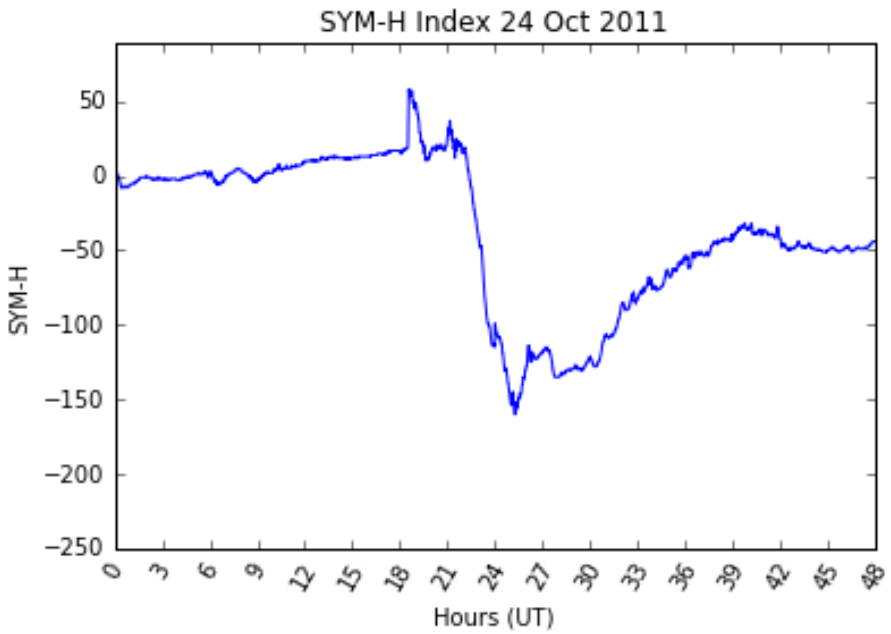


Figure 2.13: SYM-H index on a storm day

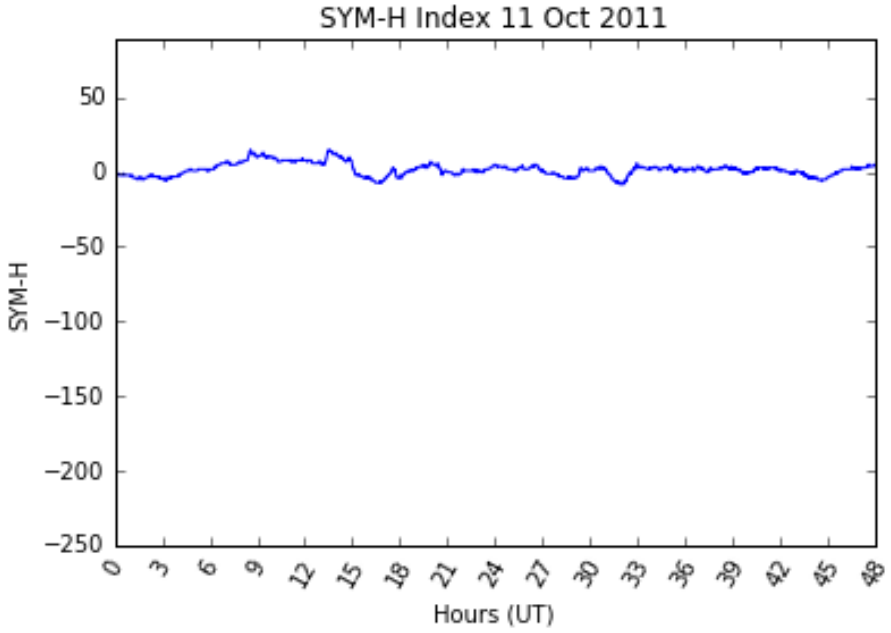


Figure 2.14: SYM-H index on a quiet day

2.3 Total Electron Content (TEC)

Ionospheric scientists use GPS observables to measure properties of the electron density such as the total electron content (TEC). The TEC is a measure of the total number of electrons that would be contained in a cylinder with a 1 m^2 cross-section that extends up vertically above a given point on the earth all the way through the ionosphere. By incorporating data from multiple receivers (greater than 2000) distributed over the globe, scientists are able to generate wide-ranging spatial maps of the TEC. The deployment

of these receivers is rapidly increasing and some areas already have very dense networks (e.g. Japan, North America, and Europe). [18]

For our purposes, we use the data from the Madrigal site at the MIT Haystack observatory. MIT Haystack has automated the process of downloading and processing GPS data to produce globally gridded TEC data. The algorithms used in the MIT automated processing of GPS (MAPGPS) software package have been described by Rideout and Coster [2006]. Processed TEC data are available to the entire scientific community via MIT Haystack's Madrigal database. [19]

2.3.1 Data Exploration

Figure 2.15 shows the format of these data. The parameters of interest are date, time, geodetic latitude (deg), geographic longitude (deg), vertically integrated electron density (TECU) and error in vertically integrated electron density (TECU).

YEAR	MONTH	DAY	HOUR	MIN	SEC	GDLAT	GLON	TEC	DTEC
2017	8	21	0	2	30	-90.00	-102.00	1.50000e+00	1.40000e+00
2017	8	21	0	2	30	-90.00	-101.00	1.50000e+00	1.40000e+00
2017	8	21	0	2	30	-90.00	-99.00	1.40000e+00	1.40000e+00
2017	8	21	0	2	30	-90.00	-98.00	1.30000e+00	1.40000e+00
2017	8	21	0	2	30	-90.00	-97.00	1.30000e+00	1.40000e+00
2017	8	21	0	2	30	-90.00	-95.00	1.30000e+00	1.40000e+00
2017	8	21	0	2	30	-90.00	-94.00	1.30000e+00	1.40000e+00
2017	8	21	0	2	30	-90.00	-93.00	1.30000e+00	1.40000e+00
2017	8	21	0	2	30	-90.00	-92.00	1.30000e+00	1.40000e+00
2017	8	21	0	2	30	-90.00	-90.00	1.30000e+00	1.40000e+00
2017	8	21	0	2	30	-90.00	-89.00	1.30000e+00	1.40000e+00
2017	8	21	0	2	30	-89.00	-126.00	2.50000e+00	1.30000e+00
2017	8	21	0	2	30	-89.00	-125.00	2.70000e+00	1.30000e+00
2017	8	21	0	2	30	-89.00	-123.00	2.70000e+00	1.30000e+00
2017	8	21	0	2	30	-89.00	-122.00	2.70000e+00	1.30000e+00
2017	8	21	0	2	30	-89.00	-121.00	2.90000e+00	1.30000e+00
2017	8	21	0	2	30	-89.00	-119.00	3.30000e+00	1.30000e+00
2017	8	21	0	2	30	-89.00	-118.00	3.20000e+00	1.30000e+00
2017	8	21	0	2	30	-89.00	-117.00	3.40000e+00	1.30000e+00
2017	8	21	0	2	30	-89.00	-103.00	2.50000e+00	1.30000e+00

Figure 2.15: An example of Madrigal data.

After selecting storm days and the 5 quietest days in the same month, TEC data are downloaded from the Madrigal website for all those days. The resolution of data in terms of latitude and longitude chosen for this study is $1° \times 1°$, and the range of latitude is from 20° N to 50° N. The range of longitude is 60° W to 125° W. The range is chosen such that it covers most of the continental U.S. Figure 2.16 shows a U.S. map and the grid we apply. In the latitude/longitude bins for which data are not available, the value NaN is assumed so that those bins show up as white spaces in the 2D ΔTEC maps.

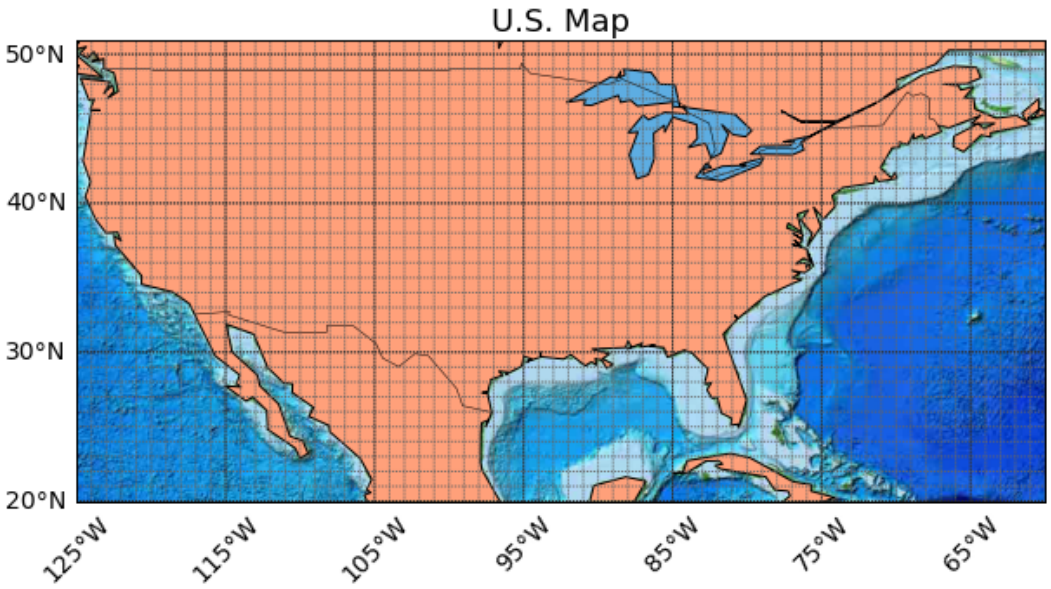
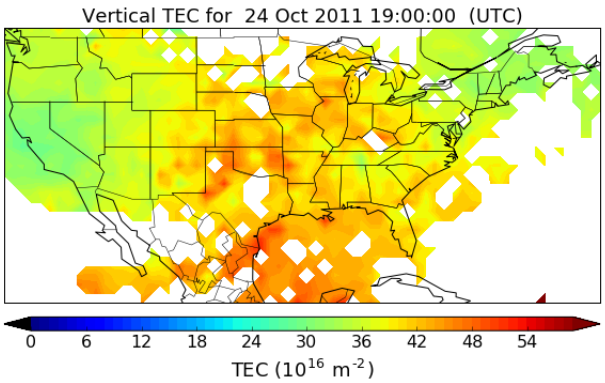
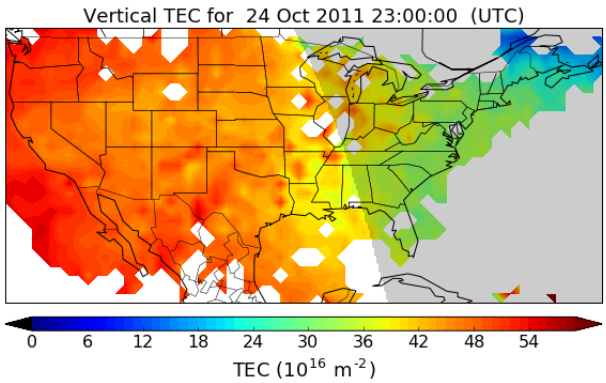


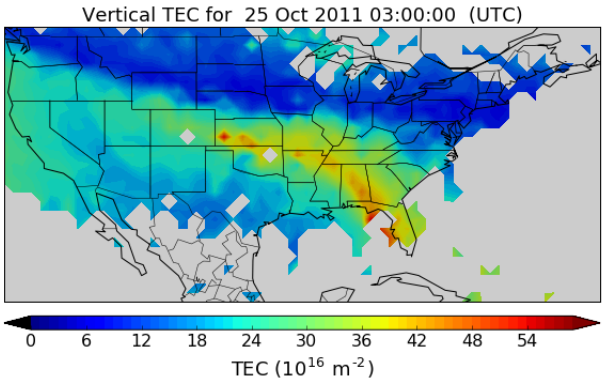
Figure 2.16: $1° \times 1°$ latitude-longitude resolution for our study



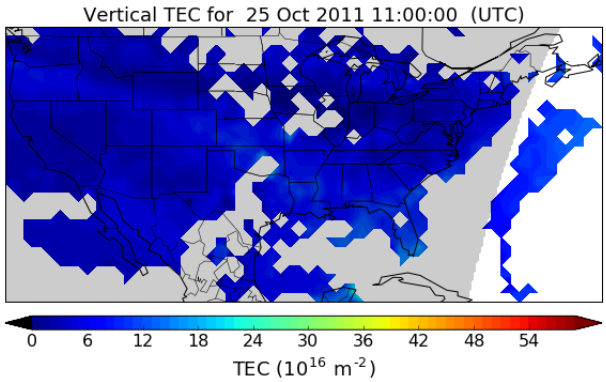
(a) At onset time T_0



(b) At T_0+4 hours



(c) At T_0+8 hours



(d) At T_0+16 hours

Figure 2.17: An example of TEC for a storm day in October 2011

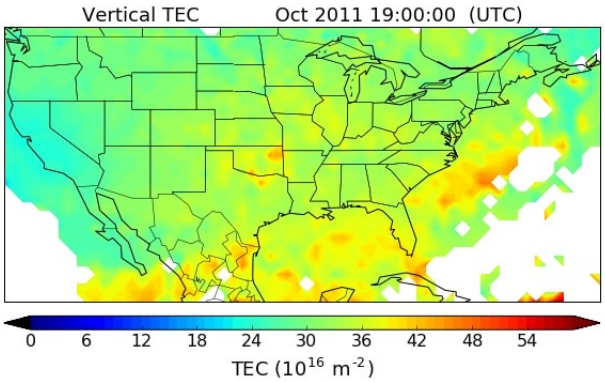
Once the data are downloaded for the 5 quiet days, the average of the TEC data every 5 minutes from the onset time T_o to T_o + 18 hours is taken, per grid cell.

$$\mathrm{TEC}_q = \frac{\mathrm{TEC}_{q1} + \mathrm{TEC}_{q2} + \mathrm{TEC}_{q3} + \mathrm{TEC}_{q4} + \mathrm{TEC}_{q5}}{5} \tag{2.1}$$

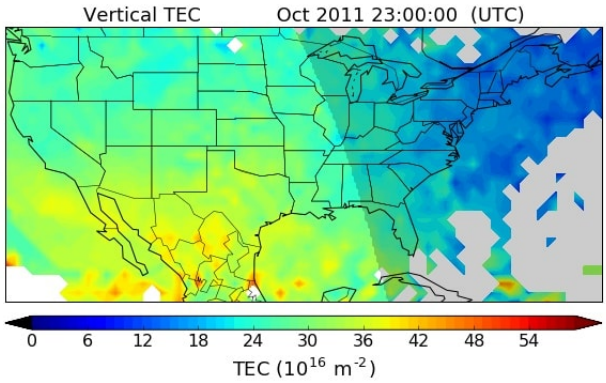
This value is then subtracted from the corresponding TEC value for the storm day to yield a ΔTEC for the storm.

$$\Delta\mathrm{TEC} = \mathrm{TEC}_{storm\ day} - \mathrm{TEC}_q \tag{2.2}$$

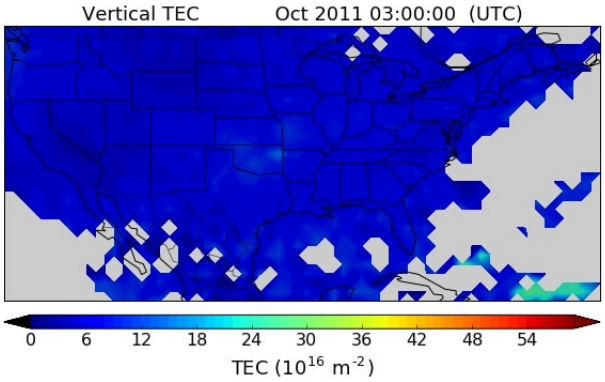
This technique helps ensure that the effect of the storm on TEC for each grid point is quantified.



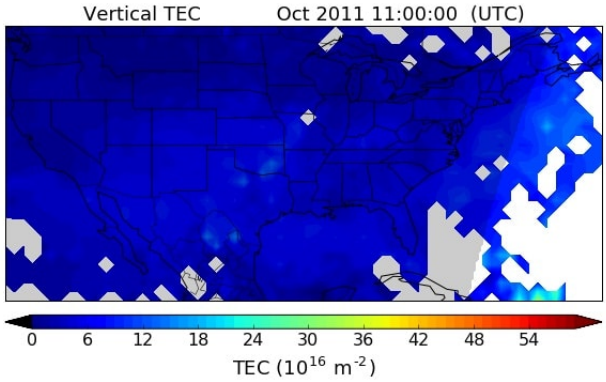
(a) At onset time T_0



(b) At T_0+4 hours



(c) At T_0+8 hours

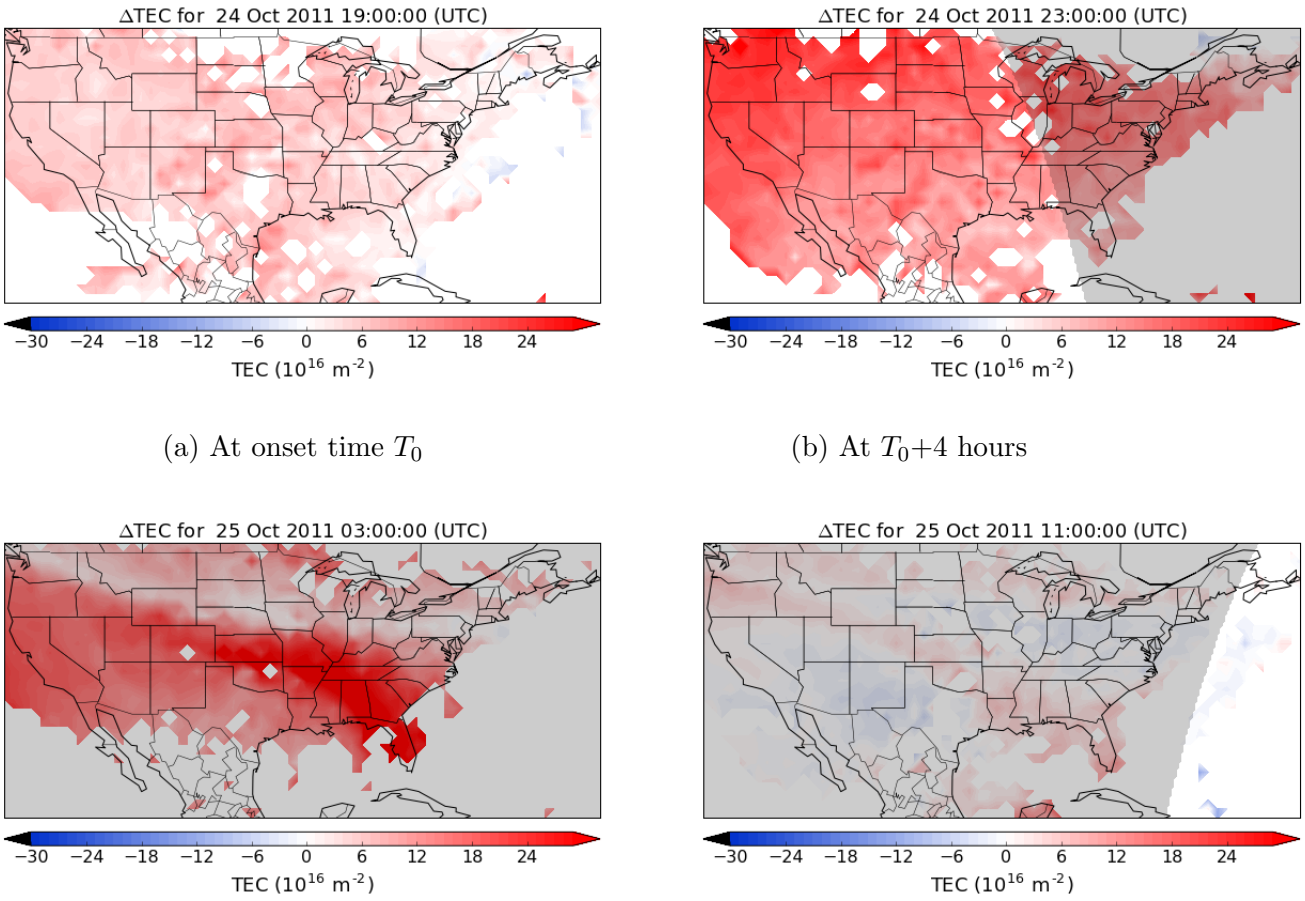


(d) At T_0+16 hours

Figure 2.18: An example of TEC_q for 5 quiet days in October 2011

As an example of the process, two dimensional vertical TEC maps for October 24, 2011 are given in Figure 2.17. The 4 panels show data plotted at different time intevals during the storm. The quietest days in the month of October in 2011 are October 28 (q_1), October 29 (q_2), October 22 (q_3) , October 14(q_4) , and October 23(q_5). In Figure 2.18, the average vertical TEC for the 5 quiet days (TEC_q) is shown for the same times as the storm day.

Figure 2.19 shows the ΔTEC for the same time intervals corresponding to Figures 2.17 and 2.18.



(a) At onset time T_0

(b) At T_0+4 hours

(c) At T_0+8 hours

(d) At T_0+16 hours

Figure 2.19: An example of ΔTEC for a storm day in October 2011

2.3.2 Quantifying the Effects of Solar Storms on TEC

Once we have calculated the ΔTEC for each five minute time interval, we average the ΔTEC for the bins in which it is available. Let us call this quantity ΔTEC_{avg}.

We sum the ΔTEC_{avg} over the 18 hours following storm onset to obtain a single number that quantifies the effect of the storm on TEC in the atmosphere over the geographic region of interest. This number is referred to as quantified delta TEC throughout the study. Note that ΔTEC_{avg} is defined for each grid point, but quantified delta TEC is a single number representing the net change in TEC over the entire grid area.

To allow study of possible magnetic aspect angle effects, we segregate the positive and negative declination ΔTEC grid values in this process. Figure 2.20 shows a magnetic field declination map of North America to illustrate the regional differences in the magnetic

field.

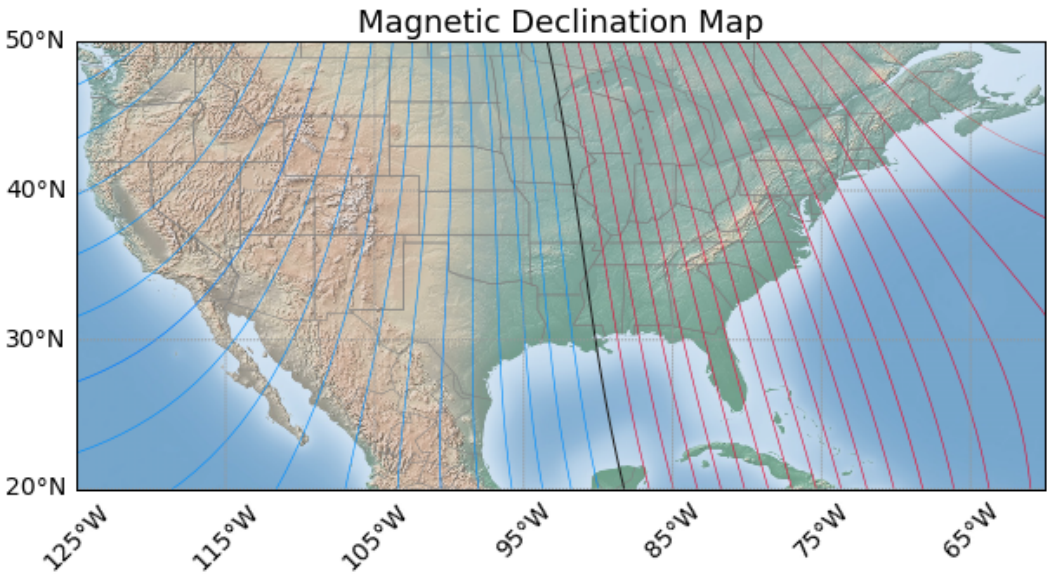


Figure 2.20: Magnetic declination map of North America for the year 2010. Red lines show regions with negative declination, and blue lines show regions with positive declination.

This gives us quantified delta TEC for both negative and positive magnetic declination regions, thus converting the dataset into 72 data points from 36 storms.

It is important to study magnetic declination regions separately because a significant ionospheric longitudinal variation at mid-latitudes over the continental US has been found recently. A higher west-side electron density in the morning and higher east-side electron density in the evening is observed, with seasonal and solar activity dependencies. This is explained by a combination of geomagnetic declination and changing zonal winds. The study by Zhang, Coster et al. [2012] confirms the declination-zonal wind mechanism and explains the longitudinal variations at mid latitudes for other geographic sectors. [20]

After separating the regions for both positive and negative magnetic declination, the time series plots of ΔTEC_{avg} are shown in Figures 2.21-2.24.

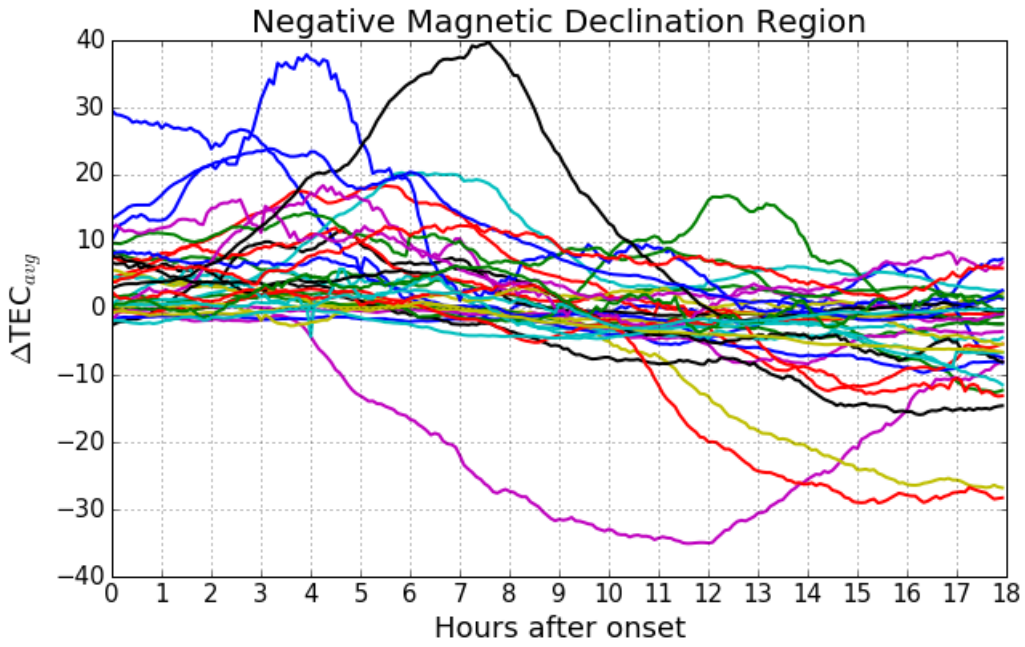


Figure 2.21: Evolution of ΔTEC_{avg} for 18 hours after storm onset taken at five minute intervals for negative magnetic declination region

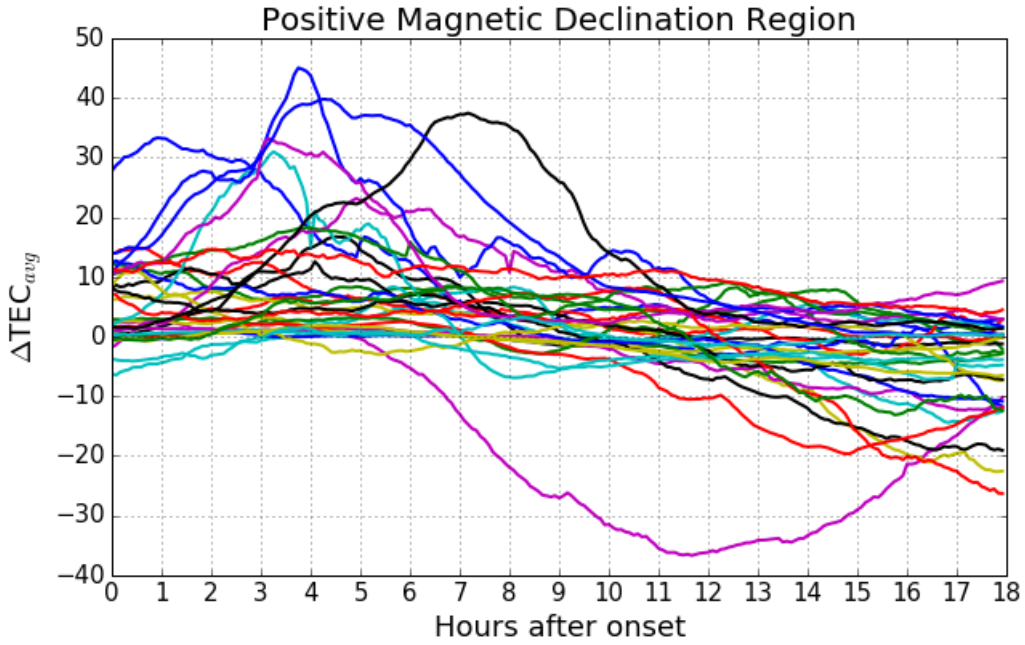


Figure 2.22: Evolution of ΔTEC_{avg} for 18 hours after storm onset taken at five minute intervals for positive magnetic declination region

To illustrate the sensitivity of TEC measurements to large-scale geophysical variations we present Figures 2.23 and 2.24. These figures show some very interesting geophysical traits. In the American sector, storms with onset times prior to 13 UT have a weak positive phase (enhanced TEC) for the first half of the storm period, followed by a stronger negative phase (decreased TEC) that commences about 10 hours after the initial compression phase of the storm (which we use as our reference start time). Storms with onset

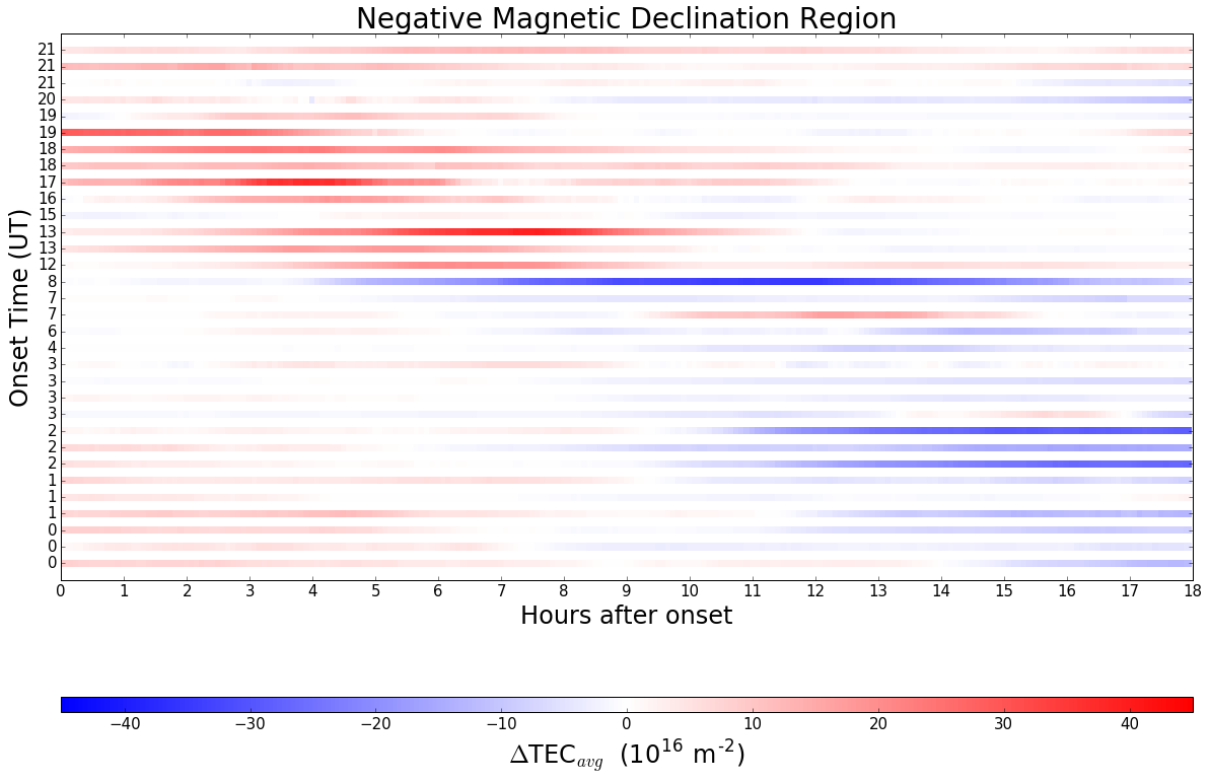


Figure 2.23: Time series plot of ΔTEC_{avg} with respect to the storm onset time for negative magnetic declination region

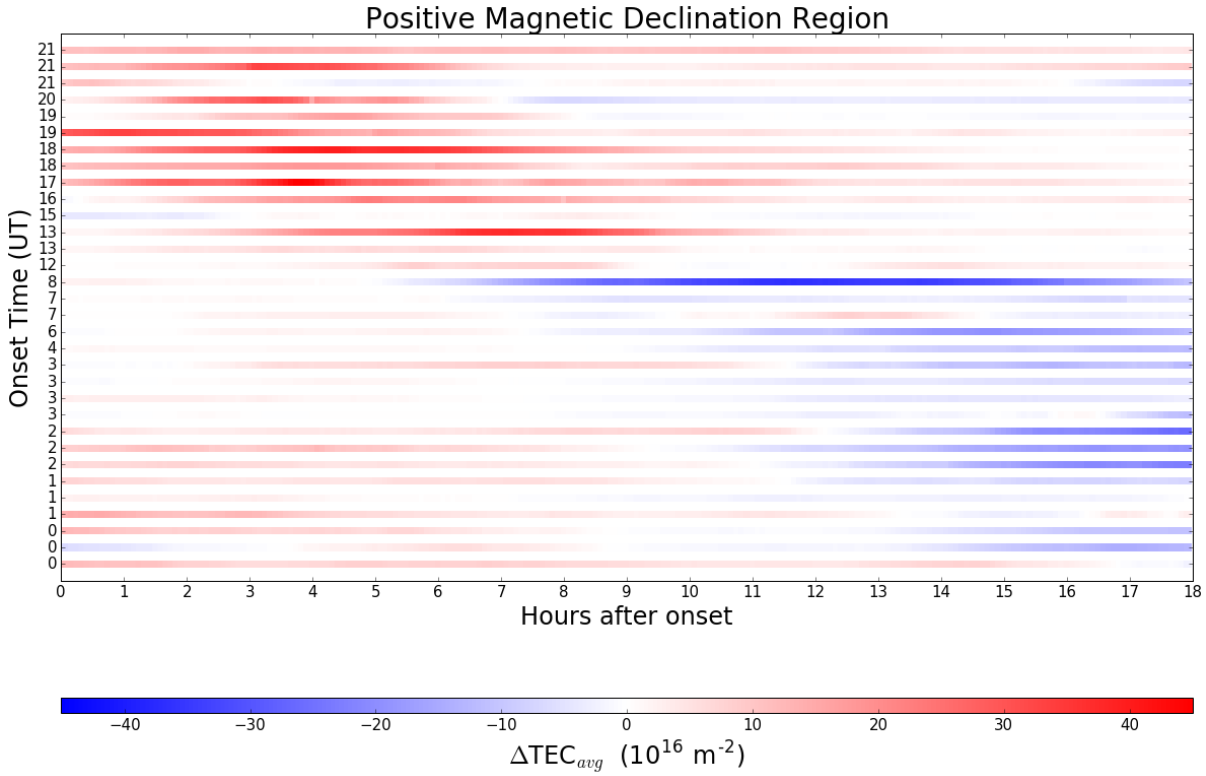


Figure 2.24: Time series plot of ΔTEC_{avg} with respect to the storm onset time for positive magnetic declination region

times after 13 UT have an initial strong positive phase, followed by a weaker phase that is usually still positive. This weaker phase commences 10 hours after our designated start time. These effects appear to be independent of magnetic declination angle, since the two figures are remarkably similar.

2.4 Effects of Other Natural Phenomena on TEC in the Ionosphere

Figure 2.25 shows 4 panels to illustrate normal variations in TEC associated with sunrise (a & b) and sunset (c & d).

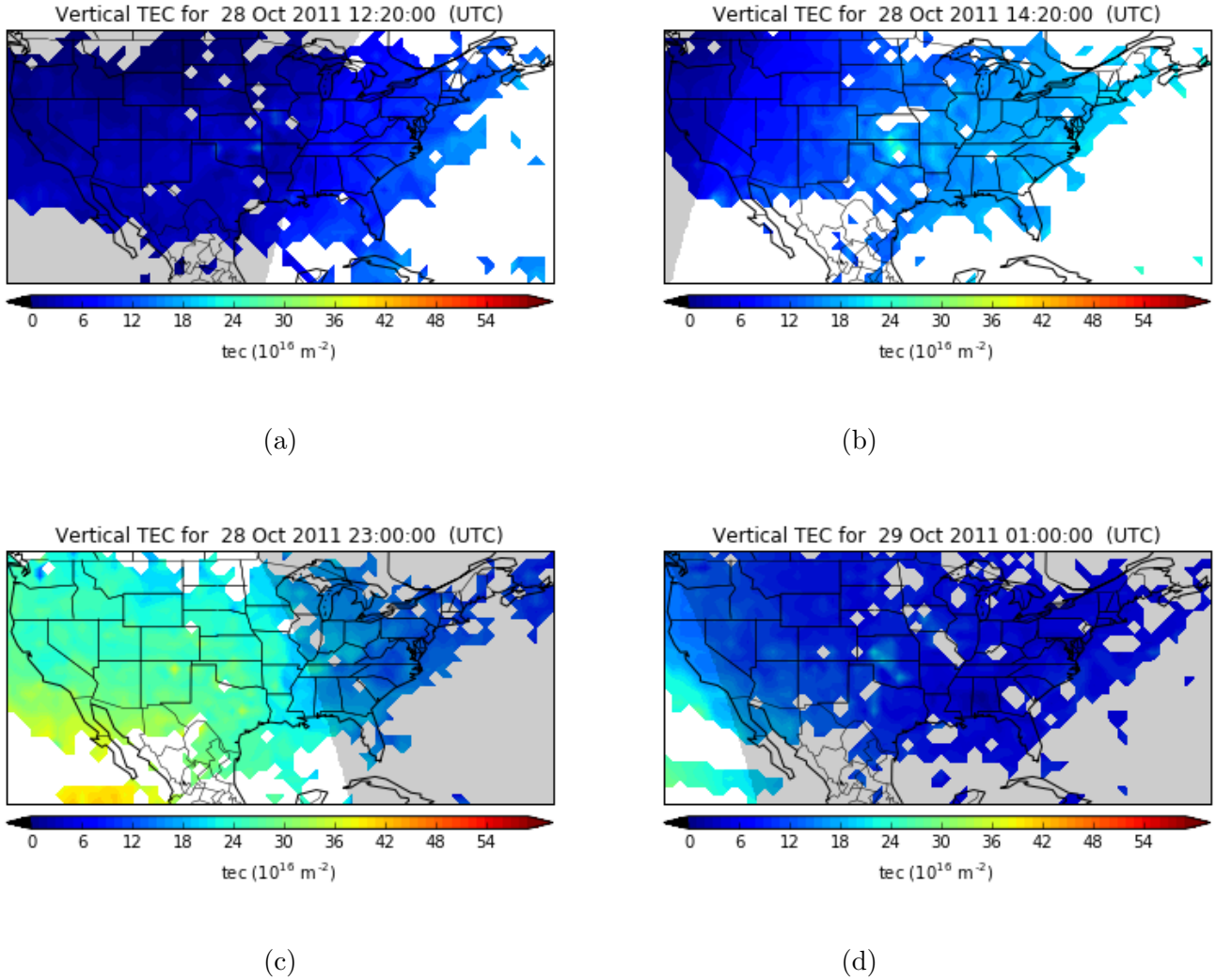


Figure 2.25: Variations in TEC associated with sunrise (a & b) and sunset (c & d) on a quiet day

Figure 2.26 shows a similar set of plots that highlight the west-to-east motion of an elliptical patch of reduced TEC corresponding to the penumbra of the August 2017 solar eclipse. In both cases the gradients in TEC reliably indicate the location of regions where TEC is changing due to changes in direct solar facing. Our approach to studying ΔTEC on large geomagnetic storm days exploits the sensitivity of the TEC technique to study the variations related to this alternative and sporadic type of geophysical forcing.

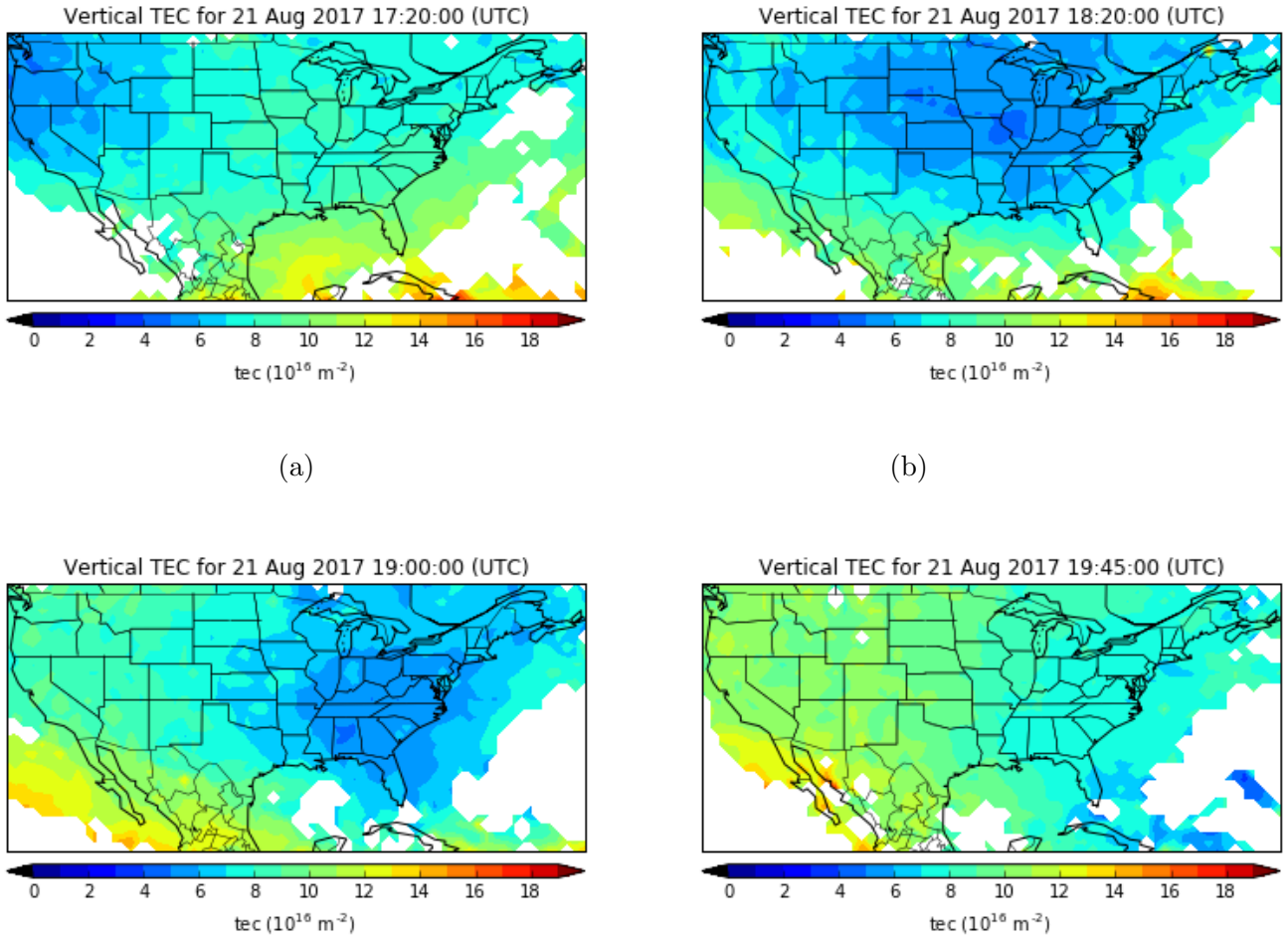


Figure 2.26: Total electron content during the solar eclipse in August of 2017 at four different times.

Chapter 3

Analysis using Machine Learning

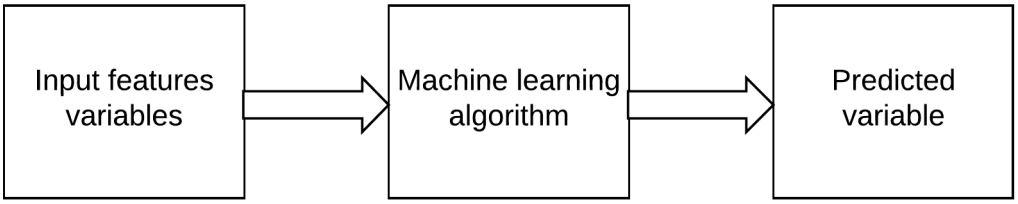


Figure 3.1: Schematic diagram of the process used to study major geomagnetic storms.

To gain further insight into how geomagnetic storms influence the TEC, we wish to quantify the impact of various storm parameters. The ideal scenario would be to find an analytical formula relating storm parameters to TEC. However, since no such clear cut relationship could be discovered, we turn to machine learning techniques.

Machine learning allows us to autonomously infer complex, non-linear relationships between a given set of independent variables and the dependent variable. As an additional advantage, most machine learning techniques can quantify the relative importance of each of the independent variables to the response exhibited by the dependent variable.

In this chapter we describe the application of machine learning to our dataset of 36 storms. We describe how the algorithm for this study was chosen, and explain its theory, application, evaluation and challenges. The underlying idea is that the weights autonomously inferred by the algorithm for each input variable show its importance to the resultant TEC change. Figure 3.1 shows a block diagram to illustrate the process. The machine learning algorithm is designed to generate a mapping between the inputs and the outcomes. The approach is described in more detail in the remainder of this chapter.

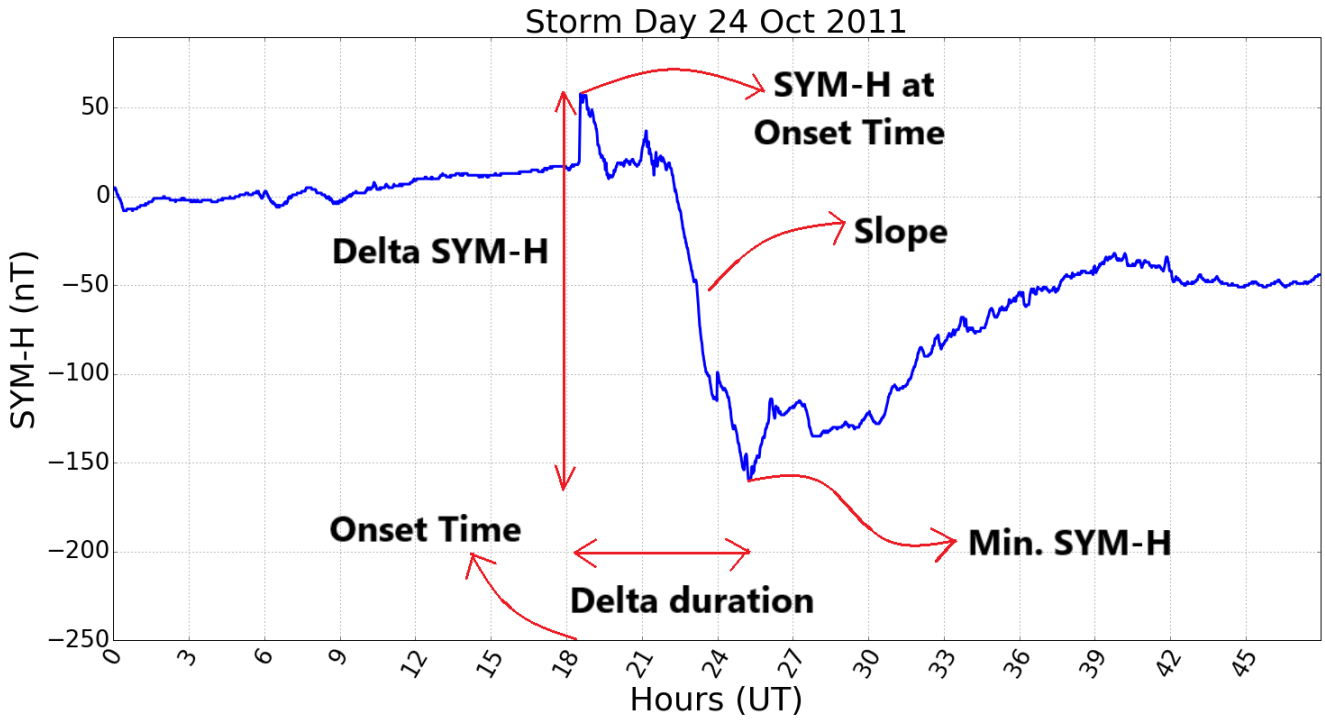


Figure 3.2: Definitions of input feature variables studied

3.1 Input Features

The input features are selected as quantified values related to physical characteristics of the ring current's reaction to the storm, as measured by the SYM-H signatures. The variables are defined below and graphically depicted in Figure 3.2.

Onset Time: The point in time when there is a sudden increase in the SYM-H index, followed by a sharp decrease.

Min. SYM-H: The minimum value of SYM-H index (nT) that is achieved during an 18-hour interval after the onset time.

Delta duration: The time that it takes a storm to achieve its Min. SYM-H value after the onset time (T_0).

Delta SYM-H: The difference between the SYM-H index value at the onset time (T_0) and the min. SYM-H value achieved during the storm.

Slope: The rate of change of SYM-H as follows:

$$\text{Slope} = \frac{\text{Delta SYM-H}}{\text{Delta duration}}$$

These five quantities represent the intensity, onset time, abruptness, and duration of the initial phase of each storm.

Negative Declination: We define this value to be equal to 1 in regions where the magnetic declination is negative, and 0 where the declination is positive. This additional feature allows us to account for possible magnetic declination effects.

Each of these quantities are calculated for all the 36 storms that satisfy the conditions specified in Table 2.2.

3.2 Exploratory Analysis & Feature Engineering

The value of interest (predicted variable) in the final data set is the quantified ΔTEC, computed separately for the positive and negative declination regions, for all 36 large storms. The chosen set of features is used as an input to a variety of machine learning algorithms including Support Vector Machines [21], Feed-forward Neural Networks [22] and Random Forests [23]. The goal here is two-fold: to see if our input data set contains any 'signal' that can explain the response, and to then choose the algorithm that performs best.

To enable quick experimentation with a large number of algorithms, the GUI driven Weka machine learning software [24] is used for exploratory analysis. We tried a variety of different algorithms, and went with the one that performs the best. The most common metric for measuring the performance of regression models (models that predict a quantity on the real line) is known as R^2. The mathematical equation for calculating R^2 is given as:

$$R^2(y, \hat{y}) = 1 - \frac{\Sigma_{i=0}^{n-1} \ (y - \hat{y})^2}{\Sigma_{i=0}^{n-1} \ (y - \overline{y})^2} \tag{3.1}$$

where $\hat{y}_i$ is the predicted value, y_i is the corresponding true value, and $\overline{y} = \frac{1}{n} \ \Sigma_{i=0}^{n-1} \ y_i$, the mean of the true values.

The R^2 metric is a measure of "goodness of fit", and quantifies how close the model's predictions are to the true values of the dependent variable. Higher values of R^2 are better. The maximum value is 1 (perfect model), and the minimum value can be arbitrarily negative. After trying several different algorithms, Random Forests were found to perform the best on our data.

In an attempt to further improve model performance, a new feature is created from the dataset as follows:

- 'Night Fraction': A new feature is created that signifies what fraction of the storm from the onset time T_o to $T_o + 18$ lies in the nighttime. Local time is used for creating this feature, taking the eastern time zone for the region of negative declination

and the mountain time zone for the positive declination region, and referring to sunset and sunrise times for the day of storm onset in these time-zones. Starting from the onset time of the storm T_o to $T_o + 18$, the fraction of this duration which lies during the nightfall (i.e.after local sunset time and before local sunrise time) is calculated.

After adding this feature, Random Forest still gave the best results as compared to the other algorithms. To improve our understanding of how each feature contributes to the final model, we switched from Weka to 'Scikit Learn' [25]. This is a machine learning library for Python that enabled us to proceed further with our study using the Random Forest algorithm.

3.3 Machine Learning Techniques

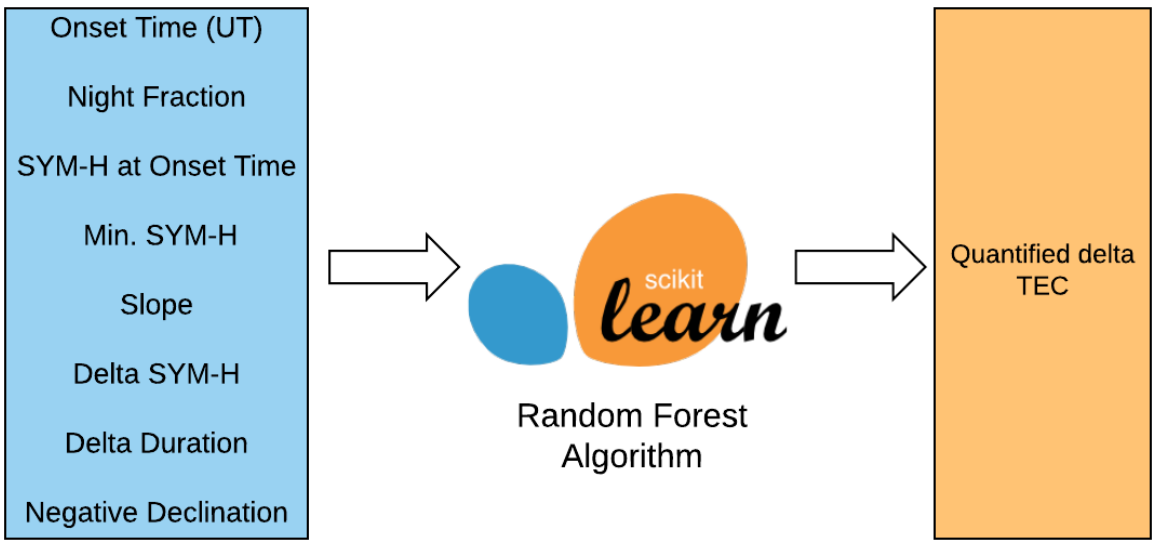


Figure 3.3: Model building using scikit learn in Python

Random Forest is a learning algorithm that uses a series of bifurcating branches to autonomously classify the input parameter space. The user controls both the input parameters (Figure 3.2) and elements of the tree topography. Figure 3.3 shows a block diagram of our approach for the specific input parameters identified as important for our study.

Random Forest does not give the weights of features in the model, but it does provide an accepted way to compute variable importance, which is the goal of the research. Since the Random Forest technique was found to provide the best performance, it should also give the closest approximation to the 'truth' about variable importances.

The fundamental unit of a Random Forest - "Regression Tree" is explained in the following subsection.

3.3.1 How Regression Trees Work

Prediction trees are a kind of non-linear predictive model. They are of two types: classification trees and regression trees. For making quantitative predictions, we are familiar with the idea of simple linear regression in which a dependent variable Y is modeled as linear variation of an independent variable X plus noise.

$$Y = \beta_0 + \beta_1 X + \epsilon \tag{3.2}$$

For multiple linear regression, the dependent variable $\mathbf{Y}$ is modeled as linear combination of various independent input variables X_1, X_2, X_3, . . . $X_p \equiv \mathbf{X}$,

$$\mathbf{Y} = \beta_0 + \beta^T \mathbf{X} + \epsilon \tag{3.3}$$

Sometimes the interaction terms between various X_i are also incorporated, making the model complicated and non-linear.

$$\mathbf{Y} = \beta_0 + \beta^T \mathbf{X} + \gamma \mathbf{X}\mathbf{X}^T + \epsilon \tag{3.4}$$

Linear regression is a global predictive model, where once the model is created, it spans the entire space. With multiple features, making a global model or finding a non-linear relation that is true for the entire domain becomes very difficult. An alternative approach is partitioning the space, so that a simpler model can effectively describe the subspace.

The basic idea is the following: If the space of all possible variable values can be partitioned 'logically', the instances that end up in a certain partition will be very similar. These subdivisions can be partitioned further until we reach a point where the data can be fitted using very simple models; this is called recursive partitioning. [26] This approach has two parts: recursive partitioning, and fitting simple models at the terminal nodes of each partition.

Recursive partitions are typically represented with a prediction tree, where at each of the terminal nodes, i.e. leaves of the tree, prediction is made. Each of the leaves of the tree represent a cell of the partition, and assigned to it is a simple prediction model that applies to that cell only. For partitioning, we start from a root node, and make a split at that node. That split could be binary or could have more than two branches.

Figure 3.4 shows an example set of branch attributes leading to prediction of the marital status of a person. In general, the data could be continuous or discrete, but ordered, or categorical. The root node for this example tree is whether a person owns a car or not. Attributes like car type or gender represent categorical data. Taxable income is a continuous data type and split is made on the condition whether the people are earning $<$ 80K or $\geq$ 80K. At each terminal node, a decision is made on the basis of majority voting and the label (marital status) which wins is attached to that node. In this example, there are 7 people who own a sports car, 5 of them are single and 2 are married; using this training dataset, the label 'single' is assigned to that particular cell.

For regression trees, data are usually continuous, and prediction is made by calculating the average of dependent variable Y values for all the instances that end up at a cell. The

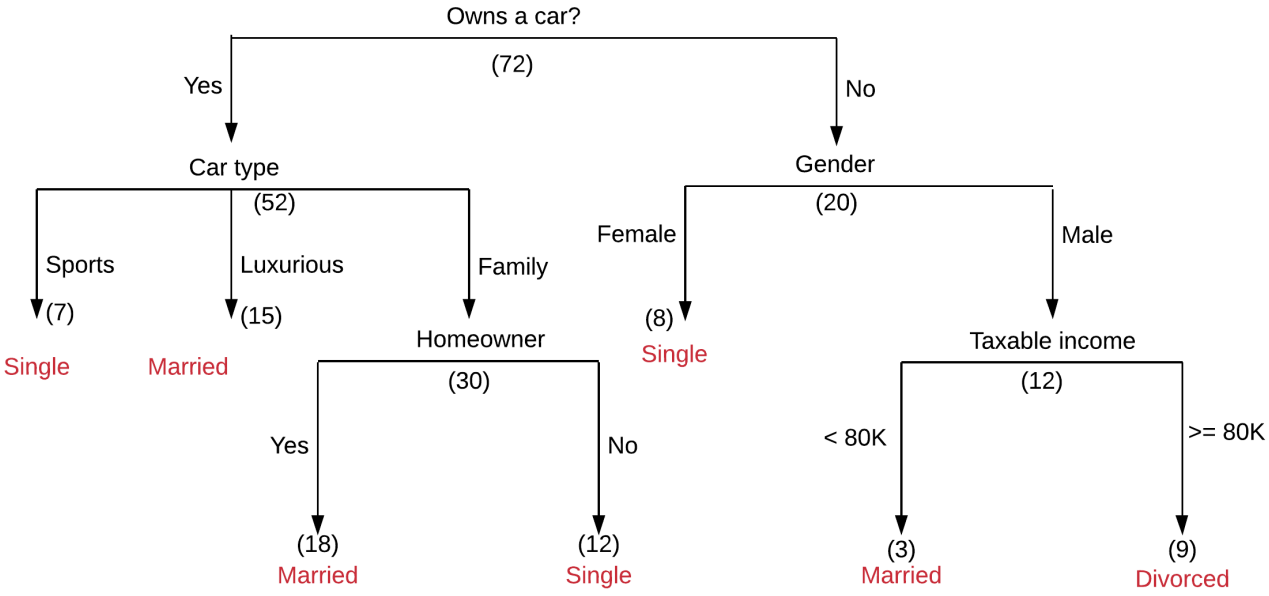


Figure 3.4: A simplified example of a prediction tree for predicting the marital status of a person: single, married or divorced. The numbers in parentheses at the nodes indicate how many data points belong to that node

advantages of using a sample mean model (i.e., calculation of mean for the dependent variable that belongs to a leaf) for making predictions are:

- It's fast, because no complicated model is needed at the end to make predictions.
- It's easy to interpret the model and determine which features are important by looking at the tree.

Once the tree is created, making a model isn't difficult. Apart from the sample mean model, some other simple regression model could also be used at each of the leaves. What matters is how we compute the partitions. What's the best way to split the data into subspaces: binary or multi-way split, and how do we specify the attributes' test conditions? Therefore, the process for making a good prediction tree requires care in finding good partitions.

There are several ways to go about constructing the partitions. At each node, we need to figure out how good the prediction is. For example, there are 8 females in the example who do not own a car. If there are 4 single females, and 4 married females, then attaching 'single' label to this node is actually not accurate. Such nodes are called non-homogeneous nodes and they have high degree of impurity. However, if there are 7 females who are single and 1 who is is married, then 'single' is a good prediction. Such nodes are called homogeneous nodes, and are said to have a low degree of impurity. There are various measures for calculating node impurity. Some of them are gini index, information gain,

gain ratio, and sum of squared errors. [27]

The algorithm picks the split that gives the lowest average impurity at the resulting nodes. For continuous variables, it's important to figure out which condition it should make a split on. Why does the split have the condition $< 80\text{K}$ & $\geq 80\text{K}$, but not another value like $< 55\text{K}$ & $\geq 55\text{K}$ or $< 90\text{K}$ & $\geq 90\text{K}$? This is also determined by taking into account all the possible splits, and the split which gives the least node impurity is chosen.

For regression trees, the typical method to discover the best partitions, especially for continuous variables, is to look at the sum of squared errors. [28] i.e. for a given tree T, we calculate the sum of squared errors S at each cell between y_i (true value at a cell) and m_c (predicted value at a cell),

$$S = \Sigma_{c \in leaves(T)} \; \Sigma_{i \in C} \; (y_i - m_c)^2 \tag{3.5}$$

where $m_c = \dfrac{1}{n_c} \; \Sigma_{i \in C} \; (y_i)$, the mean of y_i values, is the prediction made by the regression tree at leaf c. In terms of variance, this can be written as:

$$S = \Sigma_{c \in leaves(T)} \; n_c \; V_c \tag{3.6}$$

where $V_c = \dfrac{1}{n_c} \; \Sigma_{i \in C} \; (y_i - m_c)^2$ and is referred to as the within-leaf variance. Splits are made to minimize this variance at each step.

Our implementation uses mean squared error as the metric for discovering the best partitions, which is just a scaled version of the sum of squared errors.

There are several possible stopping criteria that can be used to halt the growth of the tree. For example, if S is less than a certain value, there is no need to grow a tree further. Alternatively, it can require:

- Each leaf should have a minimum number of data points
- A node must have a minimum number of data points to further allow a split
- The tree may not be more than a certain number of levels deep

Our implementation uses specific values for all three of these criteria, the specific values being set empirically.

The tree-growing algorithm can be described in three basic steps:

1. Start with all the features at the root node containing all data points. Calculate m_c, the prediction for the leaf and S, the sum of squared errors.
2. Search over all the features for splits which will reduce S as much as possible. If all the points have the same value for all the independent features, or any of the stopping criteria from the above list are satisfied, stop. If none of these conditions are satisfied, make the split, creating new nodes.
3. At each new node, go back to Step 1.

3.3.2 Random Forest: Multiple Regression Trees to Reduce Variance

Regression trees are notorious for overfitting and high variance. In machine learning terms, this is how much the model predictions depend on which subset of all possible training samples it was trained on. Random Forest is a robust machine learning technique that is based on two key ideas that solve the problems with regression trees:

- **Bootstrapping**: Sample the given dataset with replacement to create several versions of the training data set. A data point that is already chosen is put back in the original pool of the dataset so that it is equally likely to be chosen again. Data points which are not included in the bootstrap sample are used as test set. About one third of the data points from the original dataset are left out while creating a bootstrap sample. For example, imagine there are n data points in the original dataset. While sampling, each data point has a probability of being chosen equal to $1/n$. This means that the probability that a data point won't be picked is $(1-1/n)$. Since the bootstrap sample is the same size as the original dataset, the probability of a data point not being in the bootstrap sample is :

  $$\left(1-\frac{1}{n}\right)^n \approx e^{-1} \approx 0.368$$

  Sampling is then repeated and a different tree is built for each version of the bootstrap sample.

- **Attribute selection**: While growing each individual tree, only consider a random subset of features at each step.

The above techniques lead to the formation of a **'forest'** of trees that avoids overfitting [23]. For making final predictions, the individual predictions of each tree are averaged. The Random Forest technique is very effective in practice because it can implicitly model strong non-linearities.

3.3.3 Tuning the Random Forest

A Random Forest model has several hyper-parameters that must be specified, and are not directly 'learned' as part of the model building process. The values of these parameters are chosen empirically by building a model for each possible combination of values of the parameters of interest, and then testing their performance on unseen data. The combination of values that yields the best performance is the one used to build the final model.

The values for each model parameter to be tried are a matter of judgment, and there is no one process to choose them. Generally, a 'reasonable' range is chosen, with what is reasonable being guided by experience and experimentation. The most important hyperparameters that must be tuned for a Random Forest model are the following: n_estimators, max_depth, min_samples_split, min_samples_leaf and max_features. [29] The definitions, reasonable ranges of these parameters, and step sizes used while iterating through these ranges are given in Table 3.1 below:

Model Parameter	**Definition**	**Range**	**Step Size**
n_estimators	The number of trees in forest	(100,600)	100
max_depth	The maximum allowed depth of each tree	(2,11)	1
min_samples_split	The minimum number of samples required to split an internal node	(2,11)	1
min_samples_leaf	The minimum number of samples required to be at a leaf node	(1,6)	1
max_features	The number of features to consider while finding the best split	(1,8)	1

Table 3.1: Hyperparameter tuning

3.3.4 Performance Evaluation

For Random Forest, the performance metric chosen to compare the performance of each model constructed during hyperparameter tuning is the **Out-Of-Bag score**, or the OOB score. The OOB score for a particular model is estimated during the run internally, and is calculated as follows:

- Each tree in the forest is constructed using a different bootstrap sample from the original data.
- While constructing the kth tree, using a bootstrap sample, about one third of the original data points are left out.

- Each data point p would be included in the training set for building about two-third of the total trees, and in about one-third it will be left out.
- The predicted value for p is taken to be the average of the values predicted by each of the trees that did not use data point p for training.
- Once predictions are computed like this for each point in the original data, the R^2 of these predictions is computed with respect to the actual values. This R^2 value is the OOB score for the forest. [30]

The R^2 metric can be interpreted as the fraction of variance in the dependent variable that is explained by the model. Higher R^2 values indicate a model that is a better approximation of the true relation between the dependent variable and the independent variables. The maximum possible value of R^2 is 1. The value of R^2 could be negative too which is obtained if the model does even worse than a simple mean prediction.

The results obtained using this approach for the TEC prediction problem are given in Table 3.2, and the OOB Score obtained is **0.74789**.

Model Parameter	**Optimum Value**
n_estimators	600
max_depth	10
min_samples_split	2
min_samples_leaf	1
max_features	4

Table 3.2: Optimum parameters which yield the best performance after grid search over hyper-parameter space. OOB score performance metric is used for tuning the hyper-parameter space.

3.3.5 Computing Feature Importances from a Random Forest Model

Storm Parameters	**Weighted Importance**
Onset Time UT	0.4035
Night Fraction	0.1724
Sym-H at Onset Time	0.1132
Min. Sym-H	0.1020
Slope	0.0733
Delta SYM-H	0.0621
Delta Duration	0.0574
Negative declination	0.0161

Table 3.3: Feature importances

Each time a question is asked in a regression tree, the resulting split causes a reduction in the variance of the tree. The key point is that each question is associated with one particular variable. If we average the variance reductions achieved by each variable across all trees, this gives us an estimate of how much, on average, that feature helped with the model building process. Computing this metric for each variable and normalizing to 1 gives a relative weight of the importance of each variable. [23] Table 3.3 shows the results of this computation for our study. In the next chapter we discuss and interpret the physical meaning of these relative weights.

Chapter 4

Results & Future Work

4.1 Discussion

The results shown in Figure 4.1 reveal that the onset time of the storm (UT)(see Figure. 3.2) is the most important parameter controlling the overall change in TEC over the US sector. This finding extends a larger study by Thomas et al. [2016] by quantitatively identifying the relevance of characteristic geophysical storm variables. In future work results such as these may be useful to formulate near-term predictive models.

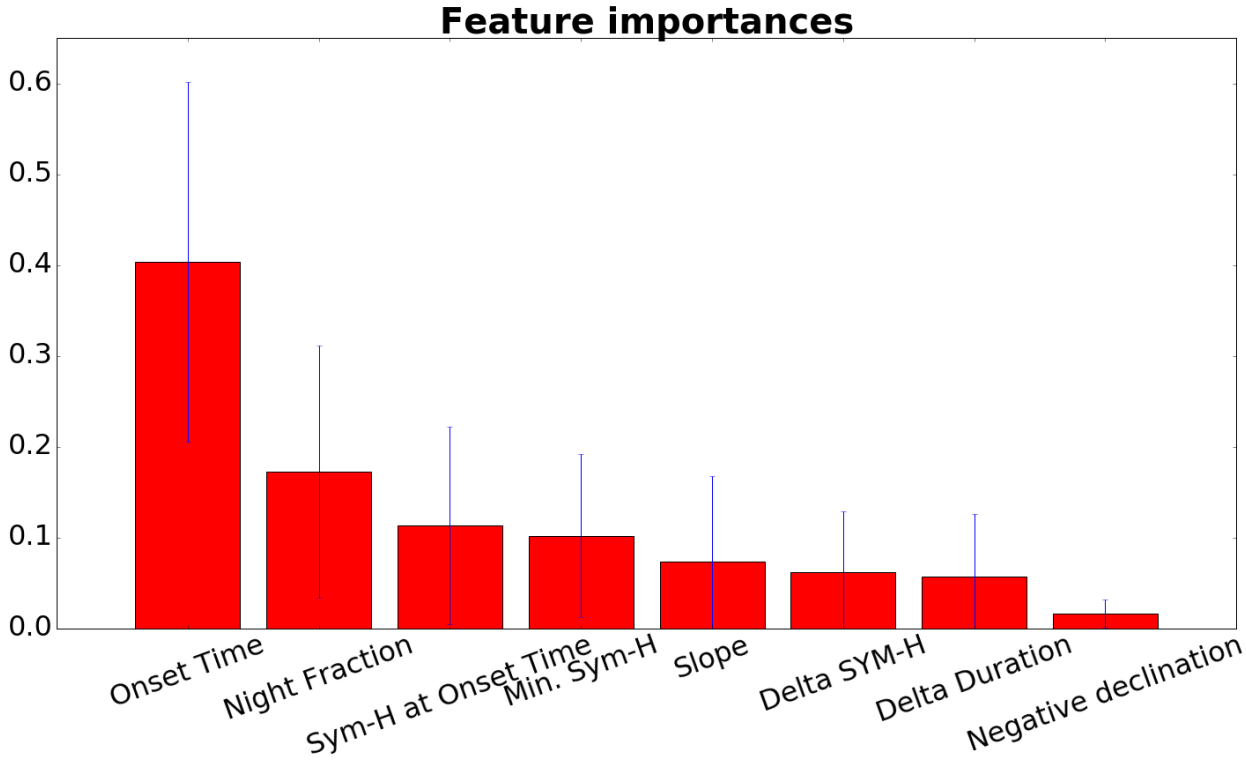


Figure 4.1: Feature importance as a result of Random Forest

OOB Score : 0.7479

The strong dependence of the ionospheric TEC response on the onset time suggests two possibilities:

1. Earth's magnetic field geometry relative to the solar wind during the onset of the storm may exert a large controlling influence on the ability of precipitating particles to produce enhanced ionization over the U.S. sector. For example, the orientation of the cusp region relative to the magnetic field carried by the solar wind varies substantially with UT. It is possible that this geometry exerts a controlling influence on the energy deposition in the high latitude ionosphere during a storm, which in turn affects the TEC enhancement at lower latitudes as the heat moves equatorward.
2. There may be a response time for storm perturbations created by solar wind or magnetotail energy dissipation to affect the longitude and latitude regions covered by this study.

Both of these ideas have been suggested by S. Zhang, A. Coster et al. [2012], and Earle and Kelly [1987] study shows hints at the second possibility. [31]

E. P. Szuszczewicz et al. [1998] investigate the worldwide responses of F region heights h_mF_2 and densities N_mF_2 as a function of universal and local times, latitudinal domains, and storm onset-times. They observed h_mF_2 to respond quickly to the storm onset (pointing to the importance of electric fields) with enhanced values in all latitudinal domains; thus showing the importance of storm onset time on the ionospheric responses. [32]

4.2 Future Work

The model created in this study could benefit from an expanded data set, which could be obtained in several ways:

1. Repeat analysis for non-equinox conditions:

   In this study the storms are chosen only from the equinox interval. Storms can be chosen from the non-equinox interval as well.

2. Repeat analysis for more years:

   Data were obtained for the years 2000-2015. Storms from 2015-2018, as well as before the year 2000 could be chosen as well, thus expanding the scope of the work.

3. Repeat analysis to include weaker storms:

   For this study only storms with minimum SYM-H below -100 nT were chosen. For future studies, to have a larger dataset, weaker storms with a higher min. SYM-H threshold (for instance -50 nT) could be chosen as well.

4. Add another independent variable and repeat the analysis:

   Other input features that could be used as independent variables include the slope of the recovery phase of the storm, or the integral of SYM-H over the duration

of the storm. The latter value is likely indicative of the total power input to the magnetosphere from the solar wind, so it may have a quantifiable effect on the ΔTEC values.

In terms of geography, the current study limited itself to modeling the TEC response over the U.S. sector. Other researchers have previously investigated solar storm effects on TEC over other geographies like Europe [33], Antarctica [34], Africa [35] etc. Studies could be done to determine if onset time is the largest controlling factor of ΔTEC in both hemispheres and at all the longitudes.

The range of controlling factors studies could also be expanded by including other storm parameters like solar wind data. This could potentially yield insight into new dependencies in the TEC response.

Bibliography

[1] E. G. Thomas, J. B. H. Baker, J. M. Ruohoniemi, A. J. Coster, and S. Zhang, "The geomagnetic storm time response of gps total electron content in the north american sector," *Journal of Geophysical Research: Space Physics*, vol. 121, no. 2, pp. 1744–1759.

[2] A. Coster and A. Komjathy, "Space weather and the global positioning system," *Space Weather*, vol. 6, no. 6.

[3] R. D. Hunsucker and J. K. Hargreaves, *The high-latitude ionosphere and its effects on radio propagation*. Cambridge University Press, 2003.

[4] B. Arbesser-Rastburg and N. Jakowski, *Effects on satellite navigation*, pp. 383–402. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007.

[5] E. E. Johnson, *Advanced high-frequency radio communications*. Artech House, 1997.

[6] P. M. Kintner and B. M. Ledvina, "The ionosphere, radio navigation, and global navigation satellite systems," *Advances in Space Research*, vol. 35, no. 5, pp. 788 – 811, 2005. Fundamentals of Space Environment Science.

[7] N. Jakowski, M. M. Hoque, and C. Mayer, "A new global tec model for estimating transionospheric radio wave propagation errors," *Journal of Geodesy*, vol. 85, pp. 965–974, Dec 2011.

[8] A. J. Mannucci, B. D. Wilson, D. N. Yuan, C. H. Ho, U. J. Lindqwister, and T. F. Runge, "A global mapping technique for gps-derived ionospheric total electron content measurements," *Radio Science*, vol. 33, pp. 565–582, May 1998.

[9] "Dst models." `http://www.lund.irf.se/rwc/dst/techniques.html`. Accessed: 2018-01-30.

[10] M. Mendillo, "Storms in the ionosphere: Patterns and processes for total electron content," *Reviews of Geophysics*, vol. 44, no. 4.

[11] N. C. Joshi, N. S. Bankoti, S. Pande, B. Pande, and K. Pandey, "Behavior of plasma and field parameters and their relationship with geomagnetic indices during intense geomagnetic storms of solar cycle 23," *arXiv preprint arXiv:1003.2868*, 2010.

[12] W. D. Gonzalez, J. A. Joselyn, Y. Kamide, H. W. Kroehl, G. Rostoker, B. T. Tsurutani, and V. M. Vasyliunas, "What is a geomagnetic storm?," *Journal of Geophysical Research: Space Physics*, vol. 99, no. A4, pp. 5771–5792.

[13] P. N. Mayaud, *The Dst Index*, ch. 8, pp. 115–129. American Geophysical Union (AGU), 2013.

[14] http://wdc.kugi.kyoto-u.ac.jp/wdc/Sec3.html.

[15] "Sym-h definition." `http://wdc.kugi.kyoto-u.ac.jp/aeasy/asy.pdf`. Accessed: 2018-01-30.

[16] J. A. Wanliss and K. M. Showalter, "High-resolution global storm index: Dst versus sym-h," *Journal of Geophysical Research: Space Physics*, vol. 111, no. A2.

[17] "5 quietest days." `http://wdc.kugi.kyoto-u.ac.jp/qddays/index.html`. Accessed: 2018-01-30.

[18] "Gps." https://www.haystack.mit.edu/atm/arrays/gps/index.html.

[19] W. Rideout and A. Coster, "Automated gps processing for global total electron content data," *GPS Solutions*, vol. 10, pp. 219–228, Jul 2006.

[20] S. Zhang, J. C. Foster, J. M. Holt, P. J. Erickson, and A. J. Coster, "Magnetic declination and zonal wind effects on longitudinal differences of ionospheric electron density at midlatitudes," *Journal of Geophysical Research: Space Physics*, vol. 117, no. A8.

[21] A. Shmilovici, *Support Vector Machines*, pp. 231–247. Boston, MA: Springer US, 2010.

[22] T. M. Mitchell, *Machine Learning*, ch. 4. New York, NY, USA: McGraw-Hill, Inc., 1 ed., 1997.

[23] L. Breiman, "Random forests," *Machine Learning*, vol. 45, pp. 5–32, Oct 2001.

[24] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: An update," *SIGKDD Explor. Newsl.*, vol. 11, pp. 10–18, Nov. 2009.

[25] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[26] W. Loh, "Classification and regression trees," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 1, no. 1, pp. 14–23.

[27] L. Rokach and O. Maimon, *Data Mining With Decision Trees: Theory and Applications*, ch. 5. River Edge, NJ, USA: World Scientific Publishing Co., Inc., 2nd ed., 2014.

[28] http://www.stat.cmu.edu/ cshalizi/350-2006/lecture-10.pdf.

[29] "3.2.4.3.2. sklearn.ensemble.randomforestregressor¶."

[30] L. Breiman, "Out-of-bag estimation."

[31] G. D. Earle and M. C. Kelley, "Spectral studies of the sources of ionospheric electric fields," *Journal of Geophysical Research: Space Physics*, vol. 92, no. A1, pp. 213–224.

[32] E. P. Szuszczewicz, M. Lester, P. Wilkinson, P. Blanchard, M. Abdu, R. Hanbaba, K. Igarashi, S. Pulinets, and B. M. Reddy, "A comparative study of global ionospheric responses to intense magnetic storm conditions," *Journal of Geophysical Research: Space Physics*, vol. 103, no. A6, pp. 11665–11684.

[33] J. Lastovicka, "Monitoring and forecasting of ionospheric space weather—effects of geomagnetic storms," *Journal of Atmospheric and Solar-Terrestrial Physics*, vol. 64, no. 5, pp. 697 – 705, 2002. Space Storms and Space Weather.

[34] Z. A. A. Rashid, M. A. Momani, S. Sulaiman, M. A. M. Ali, B. Yatim, G. Fraser, and N. Sato, "Gps ionospheric tec measurement during the 23rd november 2003 total solar eclipse at scott base antarctica," *Journal of Atmospheric and Solar-Terrestrial Physics*, vol. 68, no. 11, pp. 1219 – 1236, 2006.

[35] M. Mendillo, J. A. Klobuchar, R. B. Fritz, A. V. da Rosa, L. Kersley, K. C. Yeh, B. J. Flaherty, S. Rangaswamy, P. E. Schmid, J. V. Evans, J. P. Schödel, D. A. Matsoukas, J. R. Koster, A. R. Webster, and P. Chin, "Behavior of the ionospheric f region during the great solar flare of august 7, 1972," *Journal of Geophysical Research*, vol. 79, no. 4, pp. 665–672.

Appendix A

Online Databases

A.1 Steps to download SYM-H index data:

- Go to the website : http://wdc.kugi.kyoto-u.ac.jp/index.html
- Go to the option Geomagnetic Data Service
- Click on the option: "Plot and download of ASY/SYM [since 1981] and AE [since 1975] indices"
- Add date and duration for how much data is needed. For this study, I downloaded 16 datasets for each year from 2000 to 2015.
- Select output type as "ASY and SYM output" and format type as "IAGA2002-like format"

A.2 Steps to download data from madrigal website:

To manually download the data, the following steps have to be taken:

- Go to madrigal website: http://madrigal.haystack.mit.edu/madrigal/
- Select simple local data access
- Add your details: Name, email id, Affiliation
- Choose instrument type to be: Distributed Ground Based Satellite Receivers
- Choose instrument to be: World Wide GPS receiver Network
- Select date
- Download data

Madrigal also has its API for Python script. For our case, since number of storms that we are studying is around 40, I used Python script to automatically fetch data for the specified date, time and duration.

The file is downloaded in .txt format. Since there are no missing values or (####/ 99999) values, there is no data pre-processing required.