DATA601-20A-20X

GNS4

Ionospheric Total Free-Electron Content

—Earthquake Precursor

Project Report

Jiangwei Wang (19364744)
February 05, 2021

Content

1 INTRODUCTION

1.1 Organisation

GNS Science is a New Zealand research organisation, operates as a limited liability company and is owned by the New Zealand government, with an independent Board of Directors. Founded in 1865 as the New Zealand Geological Survey, its research and consultancy area includes geological resources, environment isotopes, industrial ion-beam technologies, and natural hazards, which covers earthquakes, tsunamis, volcanic eruptions, landslides, and all kinds of natural disasters. They aim to understand natural Earth system processes and resources, and to translate these into economic, environmental, and social benefits ("About Us").

1.2 Background

1.2.1 Ionosphere

"Ionospheric is the ionized part of Earth's upper atmosphere, from approximately 48 kilometres to 965 kilometres altitude, a region that includes the thermosphere and parts of the mesosphere and exosphere. The ionosphere is ionized by solar radiation. It plays an important role in atmospheric electricity and forms the inner edge of the magnetosphere. It has practical importance because, among other functions, it influences radio propagation to distant places on earth. The ionosphere is a shell of electrons and electrically charged atoms and molecules that surround the Earth. It exists primarily due to ultraviolet radiation from the Sun ("Ionosphere")."

1.2.2 Ionospheric Total Free-Electron Content

The total free-electron content (TEC) or the ionization of the ionosphere primarily depends on the lunar activity. The amount of TEC varies with the amount of radiation received from the Sun. "Thus there is a diurnal (time of day) effect and a seasonal affect. The local winter hemisphere is tipped away from the Sun, thus there is less received solar radiation. The activity of the Sun modulates following the solar cycle, with more radiation occurring with more sunspots, with a periodicity of around 11 years. Radiation received also varies with geographical location (polar, auroral zones, mid-latitudes, and equatorial regions). Some mechanisms disturb the ionosphere and decrease the ionization. There are disturbances such as solar flares and the associated release of charged particles into the solar wind which reaches the Earth and interacts with its geomagnetic field ("Ionosphere")." In terms of this, TEC has strong daily fluctuations, with high values during the daytime while the Solar activity is strong, and near-zero values at night-time (Rhoades et al. iv).

1.2.3 Relationship between earthquake and TEC

Furthermore, during an earthquake preparation, the earth's crust will be stretched enough and release free-electrons into the atmosphere, this may change the TEC of the local ionosphere and the amount of TEC may break its usual variation routine. This ionospheric

disturbance sometimes occurs as short-term precursors to earthquakes with a precursor time of several days (Rhoades et al. iii).

1.2.4 TEC character

The ionospheric TEC will affect the GPS satellite ranging signals broadcasting at a certain level, depending on the amount of TEC units (TECU). When the signals pass through the ionosphere, the ionospheric TEC imposes a dispersive delay on the signals. Thus, this special character from ionospheric TEC can be used for TECU computation. Ideally, once we measure the routine of TEC variations in spatial and time frames, any TEC anomaly can be detected to identify if an earthquake is approaching in the regions.

1.2.5 Research on the relationship

Research studies on this relationship between ionospheric TEC and earthquakes arising in the last decade (Lim and Leong, 2428). “Retrospective analyses have identified ionospheric anomalies preceding some well-known earthquakes, including the 1999 M7.7 Chi-Chi and other earthquakes in Taiwan, the 2003 M6.6 San Simeon earthquake in California, the 1995 M7.2 Kobe earthquake in Japan, the 2006 Kithara earthquake in Greece, the 2004 M>9 Sumatra and other earthquakes in Indonesia, and the 2008 M7.9 Wenchuan earthquake in China. There is no previous study of ionospheric precursors in relation to New Zealand earthquakes before 2015 (Rhoades et al. 1).”

1.2.6 GPS Toolkit

GPS Toolkit (GPSTk), is an open-source computing suite to the satellite navigation community produced by Space and Geophysics Laboratory within the Applied Research Laboratories of the University of Texas at Austin, to process GPS associated satellite navigation system data. Its existence will enable us to compute TEC value in the ionosphere, combine with the existing substantial network of continuous GPS stations in New Zealand, which creates the opportunity to detect the TEC with spatial and time, to systematically research the relationship between earthquake and TEC, with a reasonable spatial resolution in New Zealand (Rhoades et al. 1).

1.2.7 GPSTk applied in New Zealand

GNS Science initiated an examination of possible ionospheric earthquake precursors in New Zealand, to systematically test their value for short-term earthquake forecasting since 2015. However, their research study towards this area is limited to visualise from daily images only. They employed GPSTk to build a TEC Map throughout the New Zealand region, with different TEC values and corresponding latitudes and longitudes. After this, a Fortran module is applied to calculate the correlation to be a TEC anomaly by identifying if the TEC values deviate from neighbour grids in a small region or grid of the TEC map, or if the values deviate in a large region that most likely caused by solar activities. The normal correlation value is 1, if the values deviate more in a smaller region, the correlation value will be smaller than 1. Finally, an image is created based on the gridded correlation value throughout the New Zealand

region. Figure 1 demonstrates the TEC anomaly presented by the image 3 days before the Christchurch earthquake in 2011. The less the correlation value is, the lighter the colour on the image will be. This image shows an anomaly that covers Christchurch region according to the New Zealand map, and it is within a 3-days range, that mentioned in the relative literature.

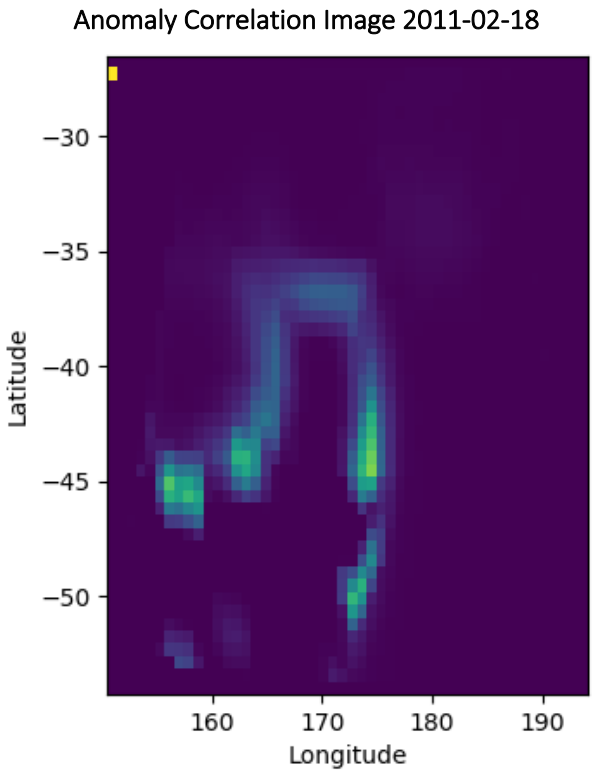


Figure 1 TEC anomaly correlation image 3 days before the Christchurch earthquake on 2011-02-21, magnitude 6.2.

1.3 Goals

1. Build a new program to resume GNS Science pre-earthquake TEC anomaly (PETA) analysis, by using GPSTk to remote sensing the ionosphere TEC, after a recent GPSTk software update caused this process to fail.

2. Implement machine learning or deep learning methods, such as neural network algorithms, to recognize the underlying relationships between ionospheric TEC anomaly and earthquakes in a supervised learning program, by utilizing the power of artificial intelligence.

1.4 Data

1.4.1 Raw GPS data

Raw GPS data is the primary resource for this whole ionospheric earthquake precursor process. The data are gathered from two channels, one is locally from GeoNet, which is powered by GNS Science and Earthquake Commission (EQC); the other one is from Scripps

Orbit and Permanent Array Center (SOPAC), housed in the University of California, San Diego ("RINEX Data Archive"). The data is in Receiver Independent Exchange Format (RINEX), an interchange format for raw satellite navigation system data, standard satellite navigation data format in the Global Navigation Satellite System (GNSS) community, that enable users to post-process the GPS stations received data and generating more accurate results. It usually contains the satellite's position, speed, or other related physical quantities, such as better models of the atmospheric conditions at the time of measurement ("RINEX").

1.4.2 RINEX data

The RINEX files we gathered from both channels are in three types. One is in "filename.yyo" (yy is the 2 digit year) formated file names, observation data, contains all the related physical quantity observations. These observations are usually recorded in a 30 seconds interval and throughout the 24 hours of the day. One is in "filename.yyd" formated file names, Hatanaka compressed files. There is a RINEX compression software written by Yuki Hatanaka, to deal with RINEX data files compression specialised, and can reach a 35% file size compression capability. The last type is in "filename.yyn" formated file names, navigation files, that accompany the observation RINEX files are broadcast ephemeris data for the GPS constellation ("RINEX Data Archive"). The filename constantly consists of 4 digits GPS receiver station name, three digits of day-of-year, and 1 digit hours range, 0 for 24 hours range, 'a' to 'x' for 1 to 23 hours range, e.g. 'ab012430.19o', station ID is 'ab01', it is on the 243rd day of 2019, and it is a 24 hours range observation file.

1.4.3 Data from GeoNet

Land Information New Zealand (LINZ) is the institution that is getting all the observation RINEX data throughout New Zealand. "The LINZ PositioNZ network consists of 39 continuously operating reference stations (CORS) located throughout New Zealand (including the Chatham Islands) and Antarctica. The network collects Global Navigation Satellite System (GNSS) data to monitor the New Zealand Geodetic Datum 2000 (NZGD2000), updating the national deformation model, and maintaining New Zealand's relationship to the International Terrestrial Reference Frame. LINZ also provides data to the International GNSS Service (IGS) from six of New Zealand stations which contribute to more accurate estimates of GNSS satellite orbits and positions. The PositioNZ Network is run in partnership with GNS Science through the GeoNet Project. The stations are also part of the GeoNet network used to monitor natural hazards in New Zealand ("PositionNZ")." Thus, the observation Rinex data of New Zealand for this project is available through GeoNet.

The files from GeoNet are consists of both observation format and compressed format, there are 370 files for each day with 185 unique station IDs, each station has both observation file and compressed file, which means half of the files are duplicated as a compressed observation file format ("PositionNZ"). From 2019, there are only observation files available from Geonet, no duplicated compressed files. All of the files are either 'gz' or 'Z' zipped files. Figure 2 is a snippet of the unzipped file content of 'tgri2980.18o', its heading contains information about the file type, date, position, and station ID, etc. information until line 32.

The rest is the observation values of each epoch. This is a 24-hour observation file, with 30-second intervals, each time interval is one epoch, there are 24 times 60 times 2 equals 2880 epochs data in each file. Each Hatanaka compressed file size is about 800 KB, each observation file size is approximately 1.7 MB and each unzipped and uncompressed observation file size is approximately 6 MB.

```
     2.11           OBSERVATION DATA    M (MIXED)           RINEX VERSION / TYPE
teqc  2016Nov7                           20181027 03:12:57UTCPGM / RUN BY / DATE
Linux 2.4.21-27.ELsmp|Opteron|gcc -static|Linux x86_64|=+   COMMENT
teqc  2016Nov7      GPS Data Reception  20181027 03:12:56UTCCOMMENT
  0.0550      (antenna height)                              COMMENT
 -38.97712911 (latitude)                                    COMMENT
+175.85849302 (longitude)                                   COMMENT
0520.659      (elevation)                                   COMMENT
BIT 2 OF LLI FLAGS DATA COLLECTED UNDER A/S CONDITION       COMMENT
TGRI (COGO code)                                            COMMENT
TGRI                                                        MARKER NAME
TGRI                                                        MARKER NUMBER
GeoNet              GNS                                     OBSERVER / AGENCY
5033K69573          TRIMBLE NETR9       5.22                REC # / TYPE / VERS
1441031289          TRM57971.00     NONE                    ANT # / TYPE
 -4952363.7236   358596.1257 -3990671.0162                  APPROX POSITION XYZ
        0.0550        0.0000        0.0000                  ANTENNA: DELTA H/E/N
     1     1                                                WAVELENGTH FACT L1/2
     8    C1    L1    S1    P1    C2    L2    S2    P2     # / TYPES OF OBSERV
    30.0000                                                 INTERVAL
    18                                                      LEAP SECONDS
Forced Modulo Decimation to 30 seconds                      COMMENT
Data supplied by the GeoNet project.  GeoNet is core        COMMENT
funded by EQC and is operated by GNS on behalf of           COMMENT
EQC and all New Zealanders.                                 COMMENT
Contact: www.geonet.org.nz  email: info@geonet.org.nz       COMMENT
 SNR is mapped to RINEX snr flag value [0-9]                COMMENT
  L1 & L2: min(max(int(snr_dBHz/6), 0), 9)                  COMMENT
teqc windowed: start @ 2018 Oct 26 00:00:00.000             COMMENT
teqc windowed: delta = 86394.000 sec                        COMMENT
```

Figure 2 Snippet of observation RINEX file of 'tgri2980.18o'

1.4.4 Scripps Orbit and Permanent Array Center data

"LINZ does not collect RINEX navigation data from the PositioNZ network ("PositionNZ") ." There is one navigation RINEX file we need for each day to accompany the observation RINEX files, which broadcasts ephemeris data for the GPS constellation, with 'auto' for its first four digits of the filenames, can be gathered from SOPAC ("PositionNZ"). Figure 3 is a snippet of navigation RINEX file 'auto2990.18n'. The header is much less than observation files, from line 14 is the body of the file. It contains ephemeris observations in each 30-seconds epoch, and there are 2880 epochs in this 24 hours range file. The size of the zipped navigation RINEX file is about 105 KB and the size of the unzipped navigation RINEX file is approximately 380 KB. There is no compressed navigation RINEX data we will handle in this project. The total size of daily raw RINEX zipped data is approximately 2 GB.

```
     2              NAVIGATION DATA                         RINEX VERSION / TYPE
comb_nav.pl         SOPAC      Sat Oct 27 07:40:02 UTC 2018 PGM / RUN BY / DATE
                 BROADCAST EPHEMERIS FILE                   COMMENT
      SOPAC - Scripps Orbit and Permanent Array Center      COMMENT
    IGPP - Institute of Geophysics and Planetary Physics    COMMENT
         SIO - Scripps Institution of Oceanography          COMMENT
        UCSD - University of California, San Diego          COMMENT
                                                            COMMENT
                email: devel@gpsmail.ucsd.edu               COMMENT
                    ftp://garner.ucsd.edu                   COMMENT
                    http://sopac.ucsd.edu                   COMMENT
                                                            COMMENT
                                                            END OF HEADER
 1 18 10 26  0  0  0.0-1.053316518664E-04-5.115907697473E-12 0.000000000000E+00
    3.100000000000E+01 3.525000000000E+01 4.157673183777E-09 1.908124248851E+00
    1.788139343262E-06 8.148038759828E-03 9.708106517792E-06 5.153664724350E+03
    4.320000000000E+05-1.974403858185E-07 9.021410829858E-01-4.470348358154E-08
    9.729067768096E-01 1.988750000000E+02 6.999088193831E-01-7.876756669960E-09
    5.910960500945E-10 1.000000000000E+00 2.024000000000E+03 0.000000000000E+00
    2.000000000000E+00 0.000000000000E+00 5.587935447693E-09 3.100000000000E+01
    4.248180000000E+05 0.000000000000E+00
 1 18 10 26  1 30  0.0-1.053591840900D-04-5.115907697473D-12 0.000000000000D+00
    1.780000000000D+02 2.616406250000D+01 4.068205171361D-09 2.695525989008D+00
    1.348555088043D-06 8.149437722750D-03 9.647570550442D-06 5.153663918530D+03
    4.374000000000D+05 0.000000000000D+00 9.020988865767D-01-1.220032572746D-07
    9.729108854150D-01 1.999453125000D+02 7.001405163485D-01-7.688120882835D-09
    6.094896734056D-10 1.000000000000D+00 2.024000000000D+03 0.000000000000D+00
    5.120000000000D+02 0.000000000000D+00 5.587935447693D-09 1.780000000000D+02
    4.320120000000D+05 4.000000000000D+00
 1 18 10 26  2  0  0.0-1.053684391081D-04-5.115907697473D-12 0.000000000000D+00
```

Figure 3 Snippet of navigation RINEX file 'auto2990.18n'.

1.4.5 Missing data

The output correlation daily CSV files and images from GNS Science's research on PETA are from 2013-01-10 to 2019-09-02, and there is no output file for 2009. There is no raw RINEX observation data available between 2019-12-31 and 2020-04-09 on Geonet. Some of the stations' RINEX observation data is missing for certain days, e.g. yaldhurst GPS receiver station in Christchurch is missing observation RINEX data yald1870.09o on 2009-07-06. Besides, Geonet doesn't have any RINEX observation data available for 2021 by this date 2021-02-02.

1.4.6 Data ethics

All data from both channels are open resources, they are open to the public and free of charge for research use in the GNSS community. There is no personal data or sensitive information involved. GeoNet has a data policy stated clearly as follows:

1. GeoNet project sponsors request to be acknowledged for all users that are using the data.
2. GeoNet project sponsors are not liable for any loss and damage resulting from the use of the data directly or indirectly.
3. GeoNet project sponsors can't guarantee data accuracy, completeness, and fitness for all-purpose use ("Data Policy").

"Please acknowledge usage of SOPAC data and/or data products in your publications, applications and products. Access to the SOPAC archive is intended for non-commercial users but there are no user fees and there is significant usage by the private sector ("Description")." This is also stated clearly on the SOPAC data description web page for their data use policy.

2 METHODOLOGY

2.1 Data retrieval and cleaning

At the beginning of this process, the raw data is getting from both GeoNet and SOPAC. There are two types of data we need, one is observation RINEX data from GeoNet, and the other one is navigation RINEX data from SOPAC. There is a python module built, called 'TECfetchData', one day at a time, to download all the observation data from GeoNet and only one broadcast ephemeris navigation RINEX data file, 'autoddds.yyn', from SOPAC for that day.

After the raw data is collected for one day, first, the 'TECfetchData' module continues its processing to unzip all of the downloaded data, then uncompress any Hatanaka compressed observation RINEX file if there is any. Software 'RNXCMP', written by Yuki Hatanaka should be employed here, to uncompress them. This software also can compress different types of RINEX files and exchange between compressed and uncompressed RINEX files ("Hatanaka Format Information at UNAVCO").

2.2 Using GPSTk to build TEC Maps

2.2.1 Background

Next, 'TECfetchData' module is calling GPSTk to compute the TEC and build TEC maps throughout the New Zealand region. GPSTk software suite is a collection of applications that process GPS associated observations, such as RINEX data, to extract and translate useful information from the observations for further research processing. It consists of a core library, auxiliary libraries, and a set of applications. All of the applications are created by using ISO-standard C++ programming language, to ensure that the code is modular, extensible, and maintainable ("GPSTk Post-Processing").

GPSTk includes a set of applications that can extract information from RINEX data to calculate the estimated TEC of the ionosphere that the RINEX data covered in a spatial and time frame. All of the RINEX data from the geographical region is used to compute a contour map of the ionospheric vertical TEC (Gaussiran et al, 5).

The common model assumption is adopted for TEC Maps analysis, where the ionosphere is a thin shell at a fixed height, usually 350 or 400 kilometres, around Earth. TEC is the two-dimensional density of free electrons in this shell. While GPS satellites broadcast a ranging signal on dual-frequency bands from an altitude of 20,200 kilometres, above the ionosphere, and the signals pass through the ionosphere at an oblique angle in general, the ionospheric TEC imposes a dispersive delay on the signals. The TEC along the line of sight (slant TEC) that produces the delay, can be computed from the ranging signals with the delay together as follows.

$$TEC_{slant} = \frac{(R_1 - R_2)}{\alpha} \bullet TECUperM$$

$$\alpha = \frac{f_1}{f_2} - 1$$

$$TECUperM = \frac{f_1^2}{20.48} \bullet 1.0e-16$$

where pseudorange observations are denoted with R, and TECUperM is a constant that converts meters to TECU, also $\boldsymbol{\alpha}$ is a constant dependent only on the two GPS broadcast frequencies f1 and f2. The TEC signal is measured in TECU where one TECU = 10^{16} electron/m^2 (Gaussiran et al, 5).

Pseudorange is the estimated distance between the GPS satellite and receivers on the ground with receiver clock errors unaccounted since the true range is impossible to be measured, caused by the difference between the transmit time and the receiver time (Rhoades et al. 3).

The slant TEC is related to the vertical TEC by an obliquity factor; from simple geometric considerations, it can be computed as follows.

$$VTEC = TEC_{slant} \bullet obliquity$$

$$obliquity = \sqrt{1 - a^2}$$

$$a = \frac{R\cos(e)}{R + h},$$

where e is the elevation of the satellite measured by the ground receiver, R is Earth's radius and h is ionosphere height (Gaussiran et al, 6).

The point at the GPS satellite signal that crosses the model ionospheric shell is called ionosphere pierce point (IPP). From the ground receiver position and the observed elevation and azimuth angles of the satellite measured by the receiver, plus the height of the ionosphere, corresponding IPP can be computed as follows.

$$p = \frac{\pi}{2} - e - \sin^{-1}\left(\frac{R\cos(e)}{R + h}\right),$$

The latitude Ø and longitude $\boldsymbol{\gamma}$ of IPP also can be computed as follows.

$$\phi_{IPP} = \sin^{-1}\left(\sin(\phi)\cos(p) + \cos(\phi)\sin(p)\cos(a)\right)$$

$$\lambda_{IPP} = \lambda + \sin^{-1}\left(\sin(p)\sin(a)/\cos(\varphi)\right),$$

where receiver position of Ø and $\boldsymbol{\gamma}$ are given, and (e, a) are the measured elevation and azimuth of the satellite. The radius of IPP equals to R + h (Gaussiran et al, 7).

Since both the GPS satellite and receiver have hardware delays, and the delays arise from differing path length for the dual GPS band signal frequencies, the measurement of

ionospheric TEC is complicated. They are essentially constant in short periods within days but vary in a longer period. The sum of the biases from the satellite and receiver that are estimated from the RINEX data, give us the information to build the TEC Maps (Gaussiran et al, 7).

2.2.2 Procedure

Nighttime satellite and receiver biases need to be estimated while the ionospheric TEC is relatively small or near zero in a given day first. Then these estimated biases will be applied to the slant TEC measurements, which computed from vertical TEC. Then fit the result and the biases to a simple model ionosphere at the point of a 2-d grid at the ionospheric height using a weighted least squares estimator, to build a TEC Map, in a fixed height (Gaussiran et al, 7).

2.2.2.1 'ResCor'

First of all, 'ResCor' application from GPSTk will take the dual-frequency range data from the input RINEX data to compute the slant ionospheric delay, ØIPP and ***γ***IPP, and the elevation and azimuth of the satellite. The output intermediate file will include the receiver's ID and its position in the header, in an 'extended' RINEX format file, as shown in figure 4 (Gaussiran et al, 7).

```
     2.10           Observation         M (Mixed)           RINEX VERSION / TYPE
ResCor v.4.1        ARL:UT/SGL/GPSTK    02/01/2021 12:23:15 PGM / RUN BY / DATE
ahti                                                        MARKER NAME
GeoNet              GNS                                     OBSERVER / AGENCY
5522R50046          TRIMBLE NETR9       5.22                REC # / TYPE / VERS
30895627            TRM57971.00     NONE                    ANT # / TYPE
 -5001713.8320   170643.0243 -3941681.7387                  APPROX POSITION XYZ
        0.0020         0.0000         0.0000                ANTENNA: DELTA H/E/N
     1     1                                                WAVELENGTH FACT L1/2
     9    C1    P1    C2    P2    EL    AZ    LA    LO    SR# / TYPES OF OBSERV
    30.000                                                  INTERVAL
  2018    10    26     0     0    0.0000000     GPS         TIME OF FIRST OBS
  2018    10    26    23    59   30.0000000                 TIME OF LAST OBS
AHTI                                                        MARKER NUMBER
    18                                                      LEAP SECONDS
Linux 2.4.21-27.ELsmp|Opteron|gcc -static|Linux x86_64|=+   COMMENT
teqc  2016Nov7      GPS Data Reception  20181027 03:09:03UTCCOMMENT
0.0020      (antenna height)                                COMMENT
-38.41144755 (latitude)                                     COMMENT
+178.04600290 (longitude)                                   COMMENT
0563.221      (elevation)                                   COMMENT
BIT 2 OF LLI FLAGS DATA COLLECTED UNDER A/S CONDITION       COMMENT
AHTI (COGO code)                                            COMMENT
Forced Modulo Decimation to 30 seconds                      COMMENT
Data supplied by the GeoNet project.  GeoNet is core        COMMENT
funded by EQC and is operated by GNS on behalf of           COMMENT
EQC and all New Zealanders.                                 COMMENT
Contact: www.geonet.org.nz  email: info@geonet.org.nz       COMMENT
SNR is mapped to RINEX snr flag value [0-9]                 COMMENT
L1 & L2: min(max(int(snr_dBHz/6), 0), 9)                    COMMENT
teqc windowed: start @ 2018 Oct 26 00:00:00.000             COMMENT
teqc windowed: delta = 86394.000 sec                        COMMENT
Edited by GPSTK Rinex Editor ver 3.5 6/21/2007 on 2021/02/01COMMENT
    55                                                      # OF SATELLITES
   G01  1141     0     0  1130  1141  1141  1141  1141  1130PRN / # OF OBS
   G02   913     0     0   906   913   913   913   913   906PRN / # OF OBS
   G03  1117     0     0  1091  1117  1117  1117  1117  1091PRN / # OF OBS
   G05   868     0     0   841   868   868   868   868   841PRN / # OF OBS
   G06   862     0     0   825   862   862   862   862   825PRN / # OF OBS
   G07   873     0     0   860   873   873   873   873   860PRN / # OF OBS
   G08  1052     0     0   937  1052  1052  1052  1052   937PRN / # OF OBS
   G09  1025     0     0   978  1025  1025  1025  1025   978PRN / # OF OBS
   G10   914     0     0   908   914   914   914   914   908PRN / # OF OBS
   G11  1131     0     0   996  1131  1131  1131  1131   996PRN / # OF OBS
   G12  1091     0     0  1051  1091  1091  1091  1091  1051PRN / # OF OBS
   G13  1127     0     0  1071  1127  1127  1127  1127  1071PRN / # OF OBS
   G14   934     0     0   907   934   934   934   934   907PRN / # OF OBS
   G15   913     0     0   848   913   913   913   913   848PRN / # OF OBS
   G16   909     0     0   800   909   909   909   909   800PRN / # OF OBS
```

Figure 4 Snippet of 'ResCor' RINEX file 'ahti2990.18o.RC'

2.2.2.2 'IonoBias'

Program 'IonoBias' from GPSTk will carry over and read in all the processed files to collect the data during local nighttime by using solar ephemeris from the broadcast ephemeris navigation RINEX file. Both simple models of the nighttime ionospheric TEC and all the biases will be fit to a least squares model.

$$\text{VTEC} = \text{bias} + \text{obliquity} * F(\emptyset, \boldsymbol{\gamma}),$$

where the function F represents the model of the ionosphere, and bias term is the satellite plus receiver biases. GPSTk users can select either a linear, quadratic or cubic model for the fit (Gaussiran et al, 7).

The least square algorithm is an unweighted sequential estimator as follows.

$$\vec{d} = P \bullet \vec{X},$$

where d is the VTEC data, X is the solution vector consisting of the unknowns, and P is the partial matrix relating the two. The partial matrix is computed from the preprocessed RINEX data, at each data point. The following are updated.

$$I\vec{d} = \sum_{data} d \bullet \vec{P}$$

$$IC = \sum_{data} P^T \bullet P$$

After all the VTEC data is computed, the matrix IC is inverted to give the covariance of the solution X, as follows.

$$Cov = inverse(IC)$$

$$\vec{X} = Cov * I\vec{d}.$$

All the estimated biases, with an uncertainty equal to the square root of the corresponding diagonal element of the covariance matrix, are in the output file (Gaussiran et al, 8). Figure 5 is the estimated TEC bias output snippet. Figure 6 is the snippet of satellite plus receiver bias output.

```
01100100001111110010110011010101l
01100100001011110010110011010101l
01100100101111110010110011010101l
01100100101111110010110011010101O
01100100101110110010110001110101l
01100100001111110010110011010101l
01100100101111110010110011010101l
01100100001111110010110011010101l
01100100101110110010110001110101l
01100100101111110010110001110101l
01100100101111110010110011010101l
01100100001111110010110011010101l
00100100101111110010110011010101l
00100100001011110010110011010101l
Npt  5243 Sta 2406 LLH   -38.5042   175.9948   374.6267
2024 462450.0 -39.96400  -51.60528 0.88    2.191      1  2   1
2024 462480.0 -39.96500  -51.46028 0.88    3.867      1  2   1
2024 462510.0 -39.96600  -51.31528 0.88   -6.181      1  2   1
2024 462540.0 -39.96700  -51.17128 0.88    1.638      1  2   1
2024 462570.0 -39.96700  -51.02628 0.88   -0.029      1  2   1
2024 462600.0 -39.96800  -50.88128 0.87    2.753      1  2   1
2024 462600.0 -37.75200  -52.35128 0.96   29.649      1  5   1
2024 462630.0 -39.96800  -50.73628 0.87    2.191      1  2   1
2024 462630.0 -37.76600  -52.22228 0.96   38.021      1  5   1
2024 462660.0 -39.96900  -50.59128 0.87    2.753      1  2   1
2024 462660.0 -37.78000  -52.09328 0.96   33.002      1  5   1
2024 462690.0 -39.96900  -50.44628 0.87   -2.829      1  2   1
2024 462690.0 -37.79400  -51.96428 0.96   35.793      1  5   1
2024 462720.0 -39.97000  -50.30128 0.87   -2.829      1  2   1
2024 462720.0 -37.80700  -51.83528 0.96   26.859      1  5   1
```

Figure 5 Snippet of 'IonoBias' TEC bias output 'aTestFile_2018-10-26'

```
IonoBias, built on the GPSTK ToolKit, Ver 1.0 6/25/04, Run 2021/01/31 23:52:30
3249  Number of SPR biases
BIAS    1  2406 G02 5243     7.928916  9.450e-02 ../2018-10-26/alldata/24062990.18o.RC
BIAS    2  2406 G05 5243    33.161939  5.894e-02 ../2018-10-26/alldata/24062990.18o.RC
BIAS    3  2406 G12 5243    37.394359  9.990e-02 ../2018-10-26/alldata/24062990.18o.RC
BIAS    4  2406 G25 5243    62.881943  7.376e-02 ../2018-10-26/alldata/24062990.18o.RC
BIAS    5  2406 G29 5243    36.446889  4.763e-02 ../2018-10-26/alldata/24062990.18o.RC
BIAS    6  2406 G15 5243    33.341572  4.635e-02 ../2018-10-26/alldata/24062990.18o.RC
BIAS    7  2406 G13 5243    30.325440  6.056e-02 ../2018-10-26/alldata/24062990.18o.RC
BIAS    8  2406 G21 5243    26.335029  4.068e-02 ../2018-10-26/alldata/24062990.18o.RC
BIAS    9  2406 G20 5243    28.795591  3.987e-02 ../2018-10-26/alldata/24062990.18o.RC
BIAS   10  2406 G10 5243    56.807910  4.538e-02 ../2018-10-26/alldata/24062990.18o.RC
BIAS   11  2406 G27 5243    54.633195  5.543e-02 ../2018-10-26/alldata/24062990.18o.RC
BIAS   12  2406 G32 5243    55.748647  5.378e-02 ../2018-10-26/alldata/24062990.18o.RC
BIAS   13  2406 G08 5243    56.714625  8.643e-02 ../2018-10-26/alldata/24062990.18o.RC
BIAS   14  2406 G14 5243    30.575819  6.165e-02 ../2018-10-26/alldata/24062990.18o.RC
BIAS   15  2406 G18 5243    35.886646  7.381e-02 ../2018-10-26/alldata/24062990.18o.RC
BIAS   16  2406 G11 5243    23.034660  8.090e-02 ../2018-10-26/alldata/24062990.18o.RC
BIAS   17  2406 G31 5243    27.616944  1.100e-01 ../2018-10-26/alldata/24062990.18o.RC
BIAS   18  2406 G01 5243    61.925465  1.349e-01 ../2018-10-26/alldata/24062990.18o.RC
BIAS   19  ahti G02 4715     8.440992  1.478e-01 ../2018-10-26/alldata/ahti2990.18o.RC
BIAS   20  ahti G05 4715    32.865471  6.644e-02 ../2018-10-26/alldata/ahti2990.18o.RC
BIAS   21  ahti G12 4715    36.619140  9.535e-02 ../2018-10-26/alldata/ahti2990.18o.RC
BIAS   22  ahti G25 4715    63.736183  7.224e-02 ../2018-10-26/alldata/ahti2990.18o.RC
BIAS   23  ahti G29 4715    37.327538  4.722e-02 ../2018-10-26/alldata/ahti2990.18o.RC
BIAS   24  ahti G15 4715    34.724981  5.192e-02 ../2018-10-26/alldata/ahti2990.18o.RC
BIAS   25  ahti G21 4715    27.516123  4.039e-02 ../2018-10-26/alldata/ahti2990.18o.RC
BIAS   26  ahti G26 4715    40.197843  1.085e-01 ../2018-10-26/alldata/ahti2990.18o.RC
BIAS   27  ahti G20 4715    31.254896  4.233e-02 ../2018-10-26/alldata/ahti2990.18o.RC
BIAS   28  ahti G10 4715    59.554928  4.532e-02 ../2018-10-26/alldata/ahti2990.18o.RC
```

Figure 6 Snippet of 'IonoBias' satellite plus receiver bias output 'satBias_2018-10-26'

2.2.2.3 'TECMaps' program

The output files from 'IonoBias', which include the satellite plus receiver biases, will be feed into program 'TECMaps' from GPSTk, as well as the output intermediate files from 'ResCor', to estimate the ionospheric TEC on each point of a horizon grid at a fixed height of 350 kilometres. Boundaries of a region in latitude and longitude over New Zealand Map and grid resolution type is also fed into the program, to compute all the TECU values at each time point and grid (Gaussiran et al, 8).

It computes the average VTEC value at each time point first, then executes a double loop, ove rall gridded points and all data values, to compute each grid value. The IPP is already computed, along with the obliquity and the VTEC, the uncertainty in the VTEC can be computed as follows.

$$VTECerr = \sqrt{d^2 + r^2},$$

where d is the uncertainty in the VTEC measurements, and r is the decorrelation error that equal to the range times a decorrelation rate input (Gaussiran et al, 8).

Between the current grid point and the current IPP, a bearing and range of the separation are computed. Then the separation vector is broken into is 2-d X and Y and weighted by the input decorrelation rate.

Once the computation loop passes through all the data, at the current grid, the final value is computed via a least squares fit of the data to a model (f) of the surface, weighted by the uncertainty VTECerr. By minimizing the function, we get follows.

$$\sum_{i=data} \left((d_i - f_i) / \sigma_i \right)^2 .$$

(Gaussiran et al, 9).

The output from 'TECMaps' is 2881 files, 2880 of them are the estimated TEC values of each grid with corresponding latitudes and longitudes. Each of the 2880 files is one of the 30 seconds time intervals TEC values throughout the 24 hours range. The snippet of one of these files is shown in figure 7, the three columns represent latitude, longitude and TECU. The 2881st file is an output grid file with a ".LL" suffix.

```
-52.122 150.562 1.948
-50.361 151.292 2.176
-48.574 151.978 2.373
-46.764 152.624 2.528
-44.934 153.233 2.649
-43.089 153.808 2.746
-41.234 154.352 2.829
-39.372 154.866 2.903
-37.507 155.354 2.969
-35.645 155.817 3.029
-33.790 156.257 3.081
-31.945 156.676 3.125
-30.115 157.075 3.162
-28.303 157.455 3.190
-26.513 157.818 3.211

-52.385 151.923 1.863
-50.612 152.617 2.129
-48.812 153.268 2.350
-46.989 153.881 2.517
-45.147 154.458 2.644
-43.290 155.002 2.745
-41.422 155.517 2.832
-39.548 156.003 2.908
-37.672 156.464 2.978
-35.799 156.901 3.040
-33.932 157.317 3.095
-32.077 157.712 3.141
-30.236 158.088 3.179
-28.414 158.446 3.207
```

Figure 7 Snippet of 'TECMaps' output file 'igs.0296'

2.3 TEC anomaly correlation

'TECfetchData' is called by 'dataProcessing' python module to download daily data and pip it through GPSTk. Following that, 'dataProcessing' calls a Fortran module, written by David Rhoades (GNS Science staff), to read all of the 2880 'TECMaps' output files, and determine if each grid's TECU value, likely to be an anomaly. The output content sample is shown in Figure 8. There are 4 columns, which are latitude, longitude, correlation to be an anomaly, and the number of grids is taken into account for such correlation calculation. There is only 1 file for 1 day, it compares each grid's TECU values to neighbour grids'. If one or fewer grids' TECU values are standing out to its or their neighbour grids', there is more likely to be a TEC anomaly caused by local earthquake preparation. If a large number of grids' TECU values are standing out from their neighbour grids, this is more likely caused by solar activities or other geomagnetic effects. There is a TEC anomaly if the value is closer to 0, or 1 on the other side.

lat	lon	corr_k	nd_k
-52.122002	150.561996	NaN	4
-50.361	151.292007	NaN	6
-48.574001	151.977997	NaN	6
-46.764	152.623993	NaN	6
-44.933998	153.233002	NaN	6
-43.089001	153.807999	0.91372406	6
-41.234001	154.352005	0.97785163	6
-39.372002	154.865997	0.98638111	6
-37.507	155.354004	0.98656416	6
-35.645	155.817001	0.97686702	6
-33.790001	156.257004	0.98372561	6
-31.945	156.675995	0.99544036	6
-30.115	157.074997	0.99810314	6
-28.302999	157.455002	0.99915373	6
-26.513	157.817993	0.99862462	4

Figure 8 Snippet of TEC anomaly correlation file '2011-02-18.csv'

2.4 TEC anomaly image

By scanning through the latitudes, longitudes and correlation values, 'dataProcessing' will call 'MakeImage' python module to produce gridded daily images, to visualize the anomaly correlation values in these spatial structured images. As shown in figure 1, the smaller the correlation value is, the lighter colour stood out from the image and is likely to be a TEC anomaly or earthquake forewarning, near the location.

2.5 Program result validation

After the PETA analysis by using GPSTk software is resumed, validation is necessary to define if the first goal is reached. In figure 9, the comparison between the two images is identical, the research towards pre-earthquake TEC anomaly is resumed, it can be carried on or extended. All of the missing data will be fulfilled if there is raw RINEX data available.

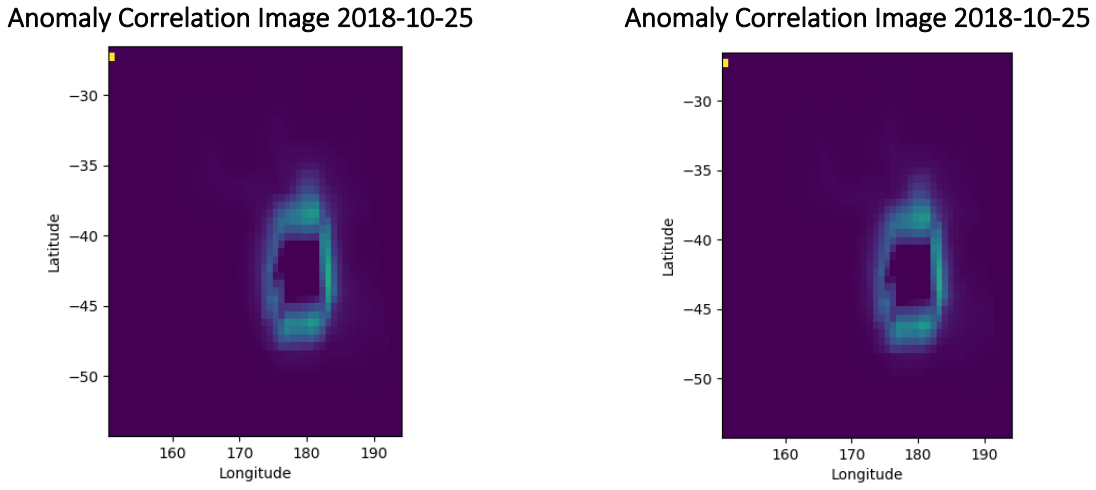


Figure 9 TEC anomaly correlation image(5 days before Taranaki earthquake on 201-10-30, magnitude 6.16). Left: new image created after GPSTk resumed; Right: original image created by GNS Science.

2.6 Temporal visualization analysis

Appendix 2 listed the patterns shown on the images in temporal only, based on an image set of 21 days before and 2 days after the 18 major earthquake events. It showed there were a lot of noises in the dataset, no regular pattern that can be temporally captured by visualization. The TEC anomaly showed on the day of the earthquake event in half of the image set, some events had TEC anomalies up to 5 days before the event, some only had the anomaly on the 6th day before, some of them had no anomaly, some events had anomaly 21 days before. The anomaly strengths were varied too. Figure 10 showed there were anomalies 20 days and 17 days before the Invercargill earthquake.

From the literature, the PETA is up to two weeks before the event (Lim and Leong, 2436). If PETA only occurs within two weeks and only within two weeks before the events, most of the 18 events have outliers. This visualization analysis was only based on a temporal factor, if more rigorously add spatial factor into this analysis, it would be even challenger. Such as GNS Science reported TEC anomaly analysis on 4 major earthquakes with both spatial and temporal factors (Rhoades et al. 23), only the Christchurch earthquake occurred in 2011 met both criteria that mentioned in the literature. May supervised deep learning methods would give us some clues on the underlying relationships.

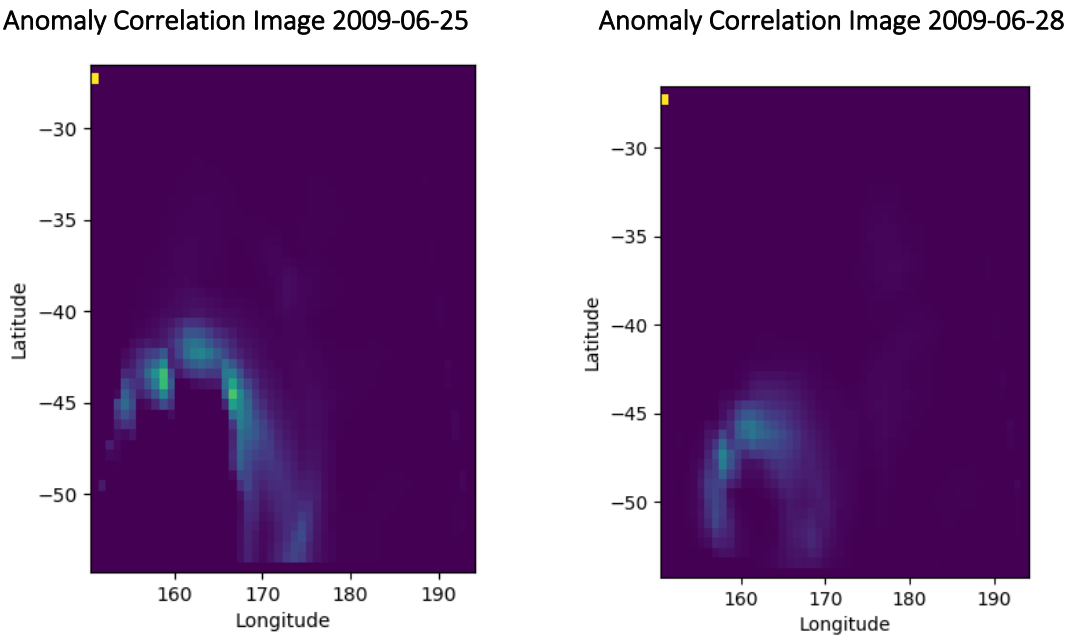


Figure 10 TEC anomaly correlation images. Left: 20 days before the Invercargill earthquake on 15-07-2009, magnitude 5.63; Right: 17 days before the Invercargill earthquake.

Figure 11 shows a weak anomaly signal that could easily be missed out by visualization. Even weaker anomalies would be impossible to detect only by visualization.

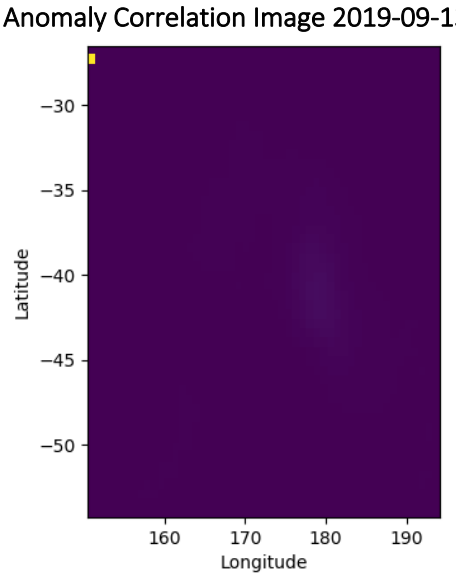


Figure 11 TEC anomaly correlation image. 18 days before 8.2 km southwest of Nuhaka earthquake on 01-10-2019, magnitude 5.65. Weak TEC anomaly signal around latitude -40, longitude 180.

2.7 Deep Learning

2.7.1 Algorithm

Convolutional Neural Network (CNN) can assign important learnable weights and biases to different aspects of the input images to distinguish them. It requires less pre-processing for the input data, and the convolutional filters can be learnt through the training, to capture the spatial and temporal dependencies of the images more effectively (Saha). In consideration of the data images that are going to be fed into the algorithm in a time sequence, CNN long short term memory algorithm (CNN LSTM) would be suitable here to develop the analysis. The standard LSTM neural network algorithm could not easily handle Input with spatial structure, like images. The CNN LSTM is an LSTM architecture specifically designed for sequence prediction problems with spatial inputs, like images (Brownlee). Tensorflow is better for high-performance models and large datasets than Keras ("Keras vs Tensorflow"), in this case, the CNN LSTM algorithm will be imported from Tensorflow.

2.7.2 Input Data

Statistically, all data is needed to achieve a reasonable result. However, existing image data is limited, downloading and processing the raw data is computationally expensive, and training a CNN LSTM based on a large dataset is also computationally expensive, by considering the limited time for this project and this is the first attempt, this deep learning research project is based on 29 largest earthquake events since 2003 as shown in appendix 1. These 29 large earthquakes are filtered out from all the earthquake events by magnitude equals or greater

than 5.0 and not an aftershock within 3 days. There is no raw data available on Geonet around 2020-02-04, thus, this large event is excluded.

2.7.3 Data splitting

Stratified data splitting strategy is used to split the data as training and test by earthquake events. This is to minimize the data similarity between training and test data, to prevent data leakage. The 29 earthquake events are going to be split as 26 events for training data and 3 events for test data, this is approximately 90% and 10% split. The three events in the test dataset are on the 2013-12-16, 2015-05-04 and 2017-10-22 events. This split by event is random, but the events in the test dataset are well spread out to maximise the difference between hold out samples.

2.7.4 Data Pre-processing

2.7.4.1 NumPy array data

The 'MakeImage' python module is extended to output those 50 by 50 by 1 NumPy array data, which are used to make the correlation images.

2.7.4.2 Sunspot number

Ionospheric TEC variation is highly correlated with solar radiation. To include this affection in the dataset, an extra layer of the NumPy array data is added with a daily sunspot number. The sunspot numbers are downloaded from the Sunspot Index and Long-term Solar Observations (SILSO), an internationally well known Solar Influences Data analysis Center ("Sunspot Number"). The new daily NumPy array data now has a new shape of 50 by 50 by 2, with a constant sunspot number for one day, and it is scaled between 0 and 1. This layer is reflecting on how much effect the ionospheric TEC is received from solar radiation.

2.7.4.3 Day of the year

Day of the year will be added to the third layer of the NumPy array data, it is also a constant number for a day, and it is scaled between 1 and -1. The first day of the year, which is summer, is scaled as 1, and the middle day of the year, which is winter, is scaled as -1. This figure is to reflect how likely the ionosphere is affected by solar radiation, as the local winter hemisphere is tipped away from the Sun, thus there is less received solar radiation. 'MakeImage' will finally output this 50 by 50 by 3 NumPy array data.

2.7.4.4 Temporal sequences

An Ipython notebook 'TECAnomaly' is loading the NumPy array data and carrying on the rest of this project.

To find out the correlation between temporal and spatial with TEC anomaly, temporal sequence structured dataset is essential. Ten days is the time sequence length we are going

to use, for each earthquake event, and 15 sequences can cover 24 days, thus, 21 days before the event, 2 days after the event, and the 22nd day is the event day. The sequences are day 1 to day 10, day 2 to day 11, and so on, until day 15 to day 24, for each earthquake event. This imbalanced choice is because we are more concentrated on the days that are leading the earthquake, which will give us forewarning. The new NumPy array date shape is 10 by 50 by 50 by 1 now.

2.7.4.5 Labeling

This deep learning analysis is predicting a percentage of the occurrence of earthquake events, then, this is a regression problem. Each of the sequences will be assigned with a single percentage value, the list of value [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0.05, 0.25, 0.4, 0.25, 0.05] is going to match with the 15 sequences of each event. It is a right-skewed bell-shaped value distribution when the earthquake event occurs at the end of the sequence, the highest value 0.4 is assigned; if the end of the sequence is one day apart from the event (one day before or after), lower 0.25 is assigned, if 2 days apart from the event at the end of the sequence (2 days before or after), 0.05 is assigned, otherwise 0 is assigned for the rest of the sequences. This will create a 2-dimensional NumPy array dataset to label our training dataset and test dataset.

2.7.4.6 Resampling

The dataset is large 'class' imbalanced, according to figure 12, label 0 is the majority 'class' in the training dataset, label 0.4 is the smallest 'class'. We need our minority 'classes' to be extremely accurate. However, the majority of our training dataset is 'class' 0, our model will be biased to learn 'class' 0 better since it has more data, and will perform poorly on the small 'classes'. Thus, 'class' balance is crucially important to put more weight on the minor 'classes' to have balanced 'classes'.

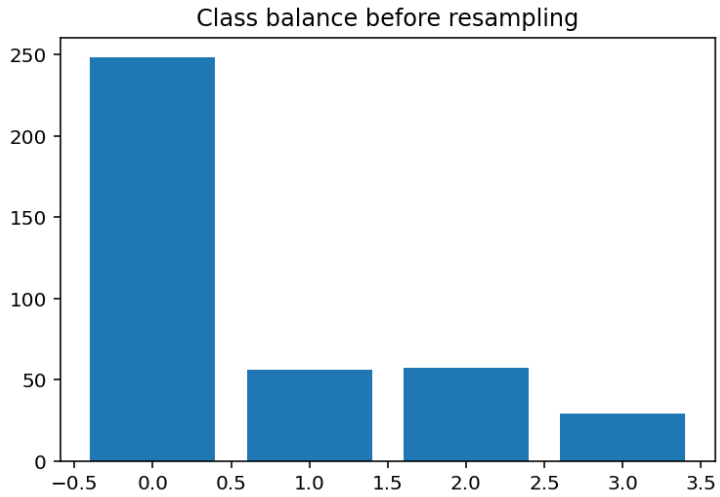


Figure 12 0, Class=0, n=248 (63.590%)
0.05, Class=1, n=56 (14.359%)
0.25, Class=2, n=57 (14.615%)
0.4, Class=3, n=29 (7.436%)

Upsampling strategy is used to duplicate the minority 'classes' randomly to have an equal amount of samples across 'classes'. The resampled result is showing in figure 13.

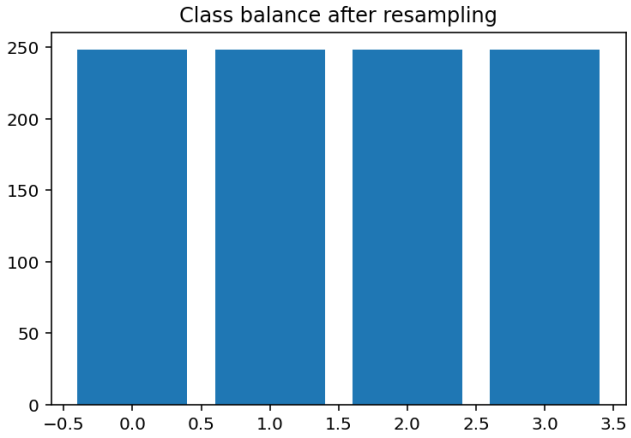


Figure 13 0, Class=0, n=248 (25.000%)
0.05, Class=1, n=248 (25.000%)
0.25, Class=2, n=248 (25.000%)
0.4, Class=3, n=248 (25.000%)

2.7.4.7 Data shuffling

The batch of 992 training dataset is shuffled manually, rather than shuffled during the training. Usually, when the data is shuffled during training, they are only shuffled within the batch size, which is the sample size the training data is sliced and repeatedly iterated over the entire dataset for a given number of epochs. If the batch size with the training is small, e.g. 4, the training data will be shuffled within the slice of 4 sequences each time, however, the neighbour sequences are still close to each other, this manual shuffle will maximise the distance between neighbour sequences, to avoid model will learn the distribution of the sequences. Also, a set seed is implemented before the shuffling.

2.7.5 Model training

A simple CNN LSTM model is tried the first time, with 2 hidden layers, 1 is ConvLSTM2D layer, and the other is a flatten layer, which is a fully-connected layer that allows the model to output one dimensional output. The model is compiled with mean squared error as the loss function, stochastic gradient descent optimizer and accuracy evaluation metric. The result isn't meaningful. During the training, all training accuracy, training loss, validation accuracy and validation loss are constant. Rectified Linear Unit (ReLU) activation function is used to regularize the output with a positive number. Validation split is 10% of the training dataset, and the model performance is evaluated based on the holdout sample, unseen test dataset. This model is trying to predict every test data as value 0.

More complex models with more layers need to have experimented, and hyperparameters need to be tuned. The Bayesian Optimizer (BO) package is helpful with grid search to determine hyperparameters. By setting a range for each hyperparameter, BO will try with different combinations of hyperparameters, to find the best result. The more

hyperparameters are put into the BO grid search, the more computational cost it will be. In this case, the number of kernels, layers, neurons and neuron shrinkage is the hyperparameters we are going to tune. The range of kernel numbers is 3 to 11, the range of layers is 0 to 4, the range of neurons is 1 to 50 and the range of neuron shrinkage is 0.1 to 1. The grid search will always set hyperparameters as continuous numbers, we need to convert them to discrete numbers to use.

Table 1 is the result of BO tuning the above 4 hyperparameters within the range. It grid searched 8 points within each of the ranges with 8 iterations. Due to the previous model only predicting every test data as value 0, the optimizer's target is set as the maximum value of model prediction on the test dataset, to try if the model can predict values greater than 0. A default ConvLSTM2D hidden layer, which will not return sequences to reduce 1 dimension is added before the fully-connected layer, and a dense layer is also added before the output layer.

	kernels	layers	neurons	shrinkage	target
trial 1	2.498	3.803	36.87	0.197	0.054
trial 2	1.624	0.624	3.846	0.732	0.05415
trial 3	3.404	2.832	2.009	0.94	0.05431
trial 4	4.33	0.8494	9.909	0.367	0.5449
trial 5	2.217	2.099	22.17	0.582	0.05437
trial 6	3.447	0.558	15.32	0.733	0.05441
trial 7	2.824	3.141	10.78	0.028	0.05448
trial 8	3.37	0.1858	30.77	0.341	0.05481
trial 9	1.26	3.796	48.32	0.617	0.05492
trial 10	2.218	0.3907	34.53	0.88	0.05482
trial 11	1.488	1.981	2.685	0.819	0.05474
trial 12	2.035	2.65	16.27	0.04	0.05484
trial 13	2.453	0.2091	34.59	0.872	0.05498
trial 14	2.942	0	34.44	0.245	0.05506
trial 15	3.573	0.1856	35.23	0.639	0.05526
trial 16	3.672	0	34.32	1	0.05534
trial 17	4.987	0.0214	34.07	0.956	0.05536
trial 18	4.646	1.431	34.25	0.038	0.05537
trial 19	5	0.9348	32.69	1	0.05561
trial 20	4.943	2.558	32.3	0.955	0.05565

Table 1 Bayesian Optimizer with 12 points grid search on 8 iterations

The results are similar, approximately 0.05, which are better than 0, but close to each other with 3 decimal place differences. The last trial has the highest target value, by converting the hyperparameters into discrete numbers, the best performing model so far should have a kernel size with 9 by 9, 2 layers, 32 neurons in the first hidden layer, and other hidden layers' neurons are should be similar while the shrinkage value is 0.955. Even the result of BO hasn't improved significantly, but it informs us of some useful trends as follows.

1. Kernel size 9 might be the optimal choice out of the range from 3 to 11.
2. A better model might be searching for a more complex model with multilayers.
3. 32 might be the best number for neurons throughout the ConvLSTM2D hidden layers.
4. The model performance is improving, it might still have potentials.

According to above, after quite a few attempts, there are 2 CNN LSTM models developed as shown in appendix 3 and 4.

Appendix 3 is a CNN LSTM model with 2 ConvLSTM2D layers, 1 Conv2D layer, all of them are followed with a batch normalization layer, a max pooling layer and a 10% drop out layer. It also has 1 time distributed flatten layer, 1 fully-connected layer, 2 LSTM hidden layers, and 3 dense layers. All LSTM layers and dense layers are followed with a 10% drop out layer.

Appendix 4 is a CNN LSTM model with 5 ConvLSTM2D layers, each layer followed by a batch normalization layer, a max pooling layer, and a 10% dropout layer. 2 dense layers before the output layer with a fully-connected layer in between.

Deep neural network training is complicated when each layer's inputs change during training when the parameters in the previous layer change. It requires lower learning rates and careful parameter initialization and slows down the training with these nonlinearities. This is referred to as internal covariate shifting, batch normalization allows the model to use much higher learning rates and be less careful about initialization to help with this unstable internal covariate shifting and increase the training speed (Ioffe and Szegedy, 1). Batch normalization can also normalize the training parameters, by re-centring and re-scaling.

Max pooling layer is useful in preventing overfitting by providing abstraction from the representation. It is a feature commonly imbibed into Convolutional Neural Network architectures. It accumulates features from maps generated by convolving a filter over images and progressively reduces the spatial size of the presentations and computation cost ("Max Pooling").

Dropout is another useful layer in preventing overfitting. It simply drops out a certain percentage of nodes randomly by its sparse activation, which means some of the random nodes are ignored during training, and the model is learning on a sparse presentation. This can avoid overfitting especially training small datasets in deep neural networks (Brownlee).

Mean squared error, stochastic gradient descent, and accuracy are still used as loss function, optimizer and evaluation metrics for both models. Different options are tested, especially the optimizer, nadam, adam, rmsprop, adadelta and adagrad are also tested. Batch size is 10, for a training dataset with a total batch size of 992, each epoch takes 90 iterations to go the entire dataset. The epoch number is 50, validation split is still 10% of the training dataset.

CNN LSTM is a method learning correlations between temporal and spatial of the presentations together, CNN2D and LSTM methods were used to find the correlations with the presentations in turn. Neither less than 5 ConvLSTM2D layers models or 2 ConvLSTM2D layers without Conv2D and LSTM models were able to predict values rather than 0.

3 RESULTS

3.1 CNN LSTM model with Conv2D and LSTM layers

Figure 13 showed the loss and accuracy for both training and in sample validation figures during the training with corresponding epochs. It showed all of them are constant, especially both accuracy, fixed at approximately 2.0 and 2.5 at all epochs. Training loss and validation loss had minor fluctuations around 0.16 and went downward slightly. This implied the model struggled to learn from the training presentations, which can only reduce the loss by a percentage in decimal places and cannot improve the accuracy. We can see the validation loss fluctuated around a constant value after about epoch 30 and, indicating 50 epochs training might be enough to reach its optimal performance.

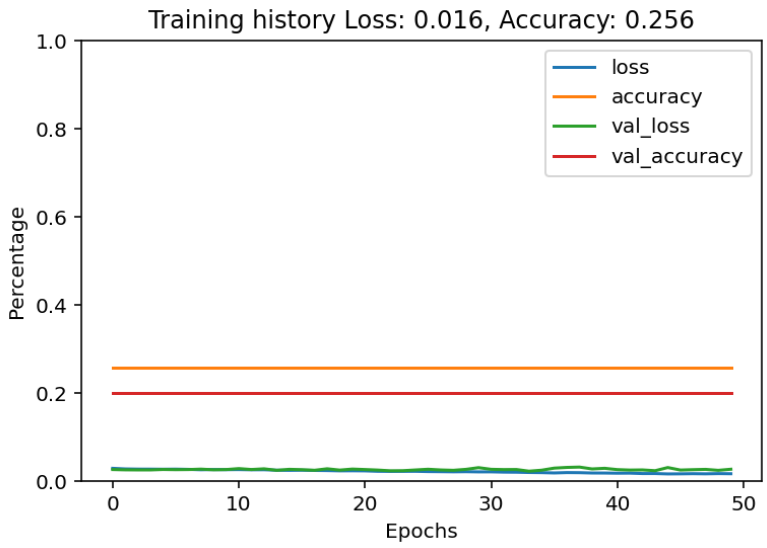


Figure 13 CNN LSTM model training loss and accuracy history

The following 3 figures are the model prediction performance on each of the three earthquake events in the unseen test dataset. Figure 14 is showing a peaking trend earlier than the actual event, then followed by a fast descent. It copied the trend of the curve but shifted to the left.

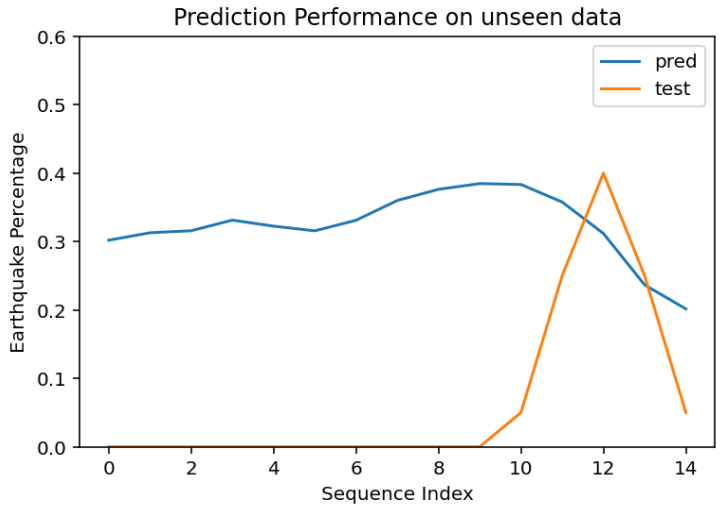


Figure 14 Model prediction performance on event 2013-12-16 (model unseen data)

Figure 15 performed better than the performance on the previous event, it predicted there was no chance to have an earthquake more confidently at the early stage, and the falling towards the end was sturdy, but the overall curve was still shifted to the left with about 1 sequence distance.

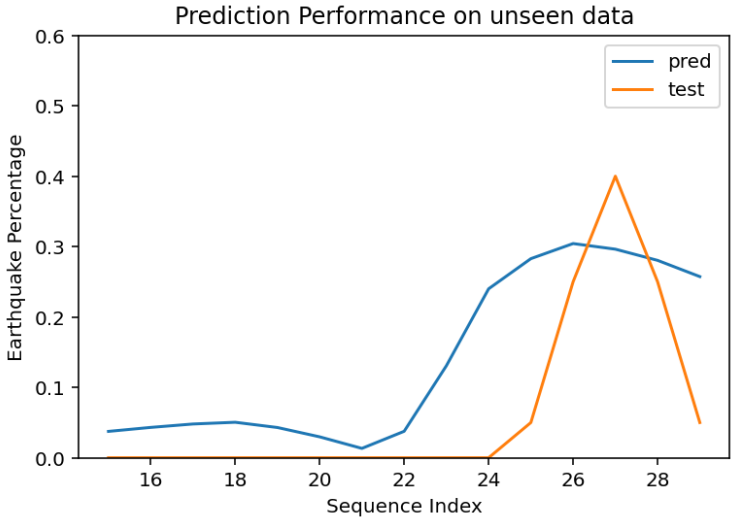


Figure 15 Prediction performance on event 2015-05-04 (model unseen data)

Figure 16 shows the prediction performance was unstable, the predicted value stayed around 0.1 without much distinction, and the curve shifted to the left about three and a half sequences.

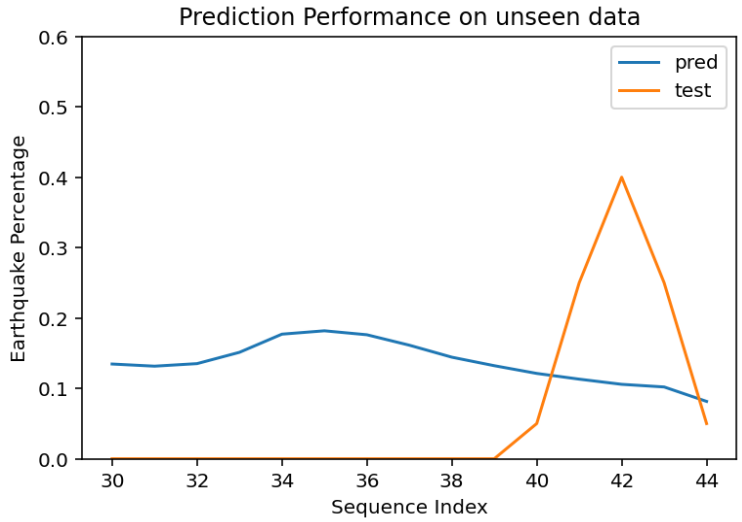


Figure 16 Prediction performance on event 2017-10-22 (model unseen data)

Figure 17 boxplot showed how the predicted values distributed against test data values in an overall viewing with all three events combine. The predicted data points distributed in a good range that covered most of the true range, however, the prediction data points are well spread over the range, with a medium value about 0.18, and there are more data points closely below to the medium. It is unlike test data points with outliers distance away from the bulk body and the main body with notches showing most of them are zeros.

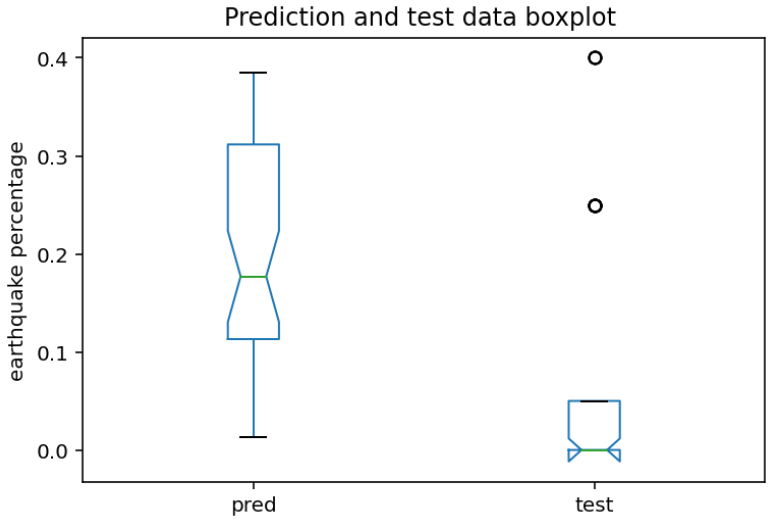


Figure 17

Figure 18 explained more about how far the predicted data points away from the test data. The 0 'class' well spread over the whole range of the predicted values. This explained most of the boxplot body in figure 17, that well covered the true range, was consisted of 'class' 0. The maximum of the prediction value, which should come from 'class' 0. 4, came from 'class' 0. The best value predicted on the day of the event, was approximately 0.3. 'class' 0.05 are all overpredicted with larger values, the largest value was close to 0.4. 0.25 'class' was spread between around 0.1 and 0.35. Combined with figure 14, 15 and 16, 'class' 0 and 0.05 have the largest values, might due to the curves being peaked in average about 2 to 3 sequences earlier.

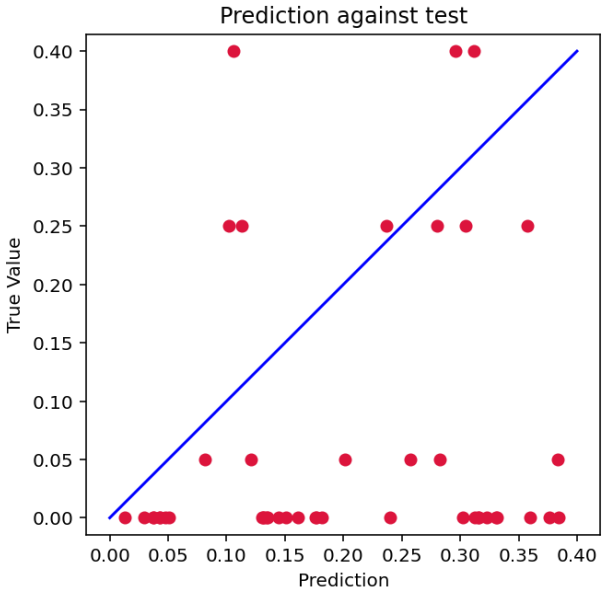


Figure 18

From figure 19 to figure 23 showed another example of how another CNN LSTM model performed while the same method is trained again. This model predicted higher value than 0.4 on a no event sequence when it is tested on the 2013 event. After the sixth sequence, it copied the trend of the original curve but shifted to the left again.

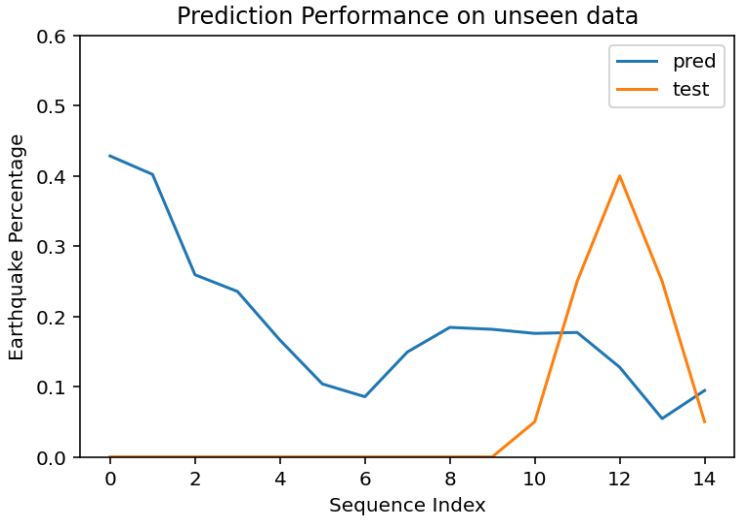


Figure 19 Model prediction performance on event 2013-12-16 (model unseen data)

The performance of this test showed in figure 20 was excellent. Even it was confusing at the beginning, but it captured the event with a great level.

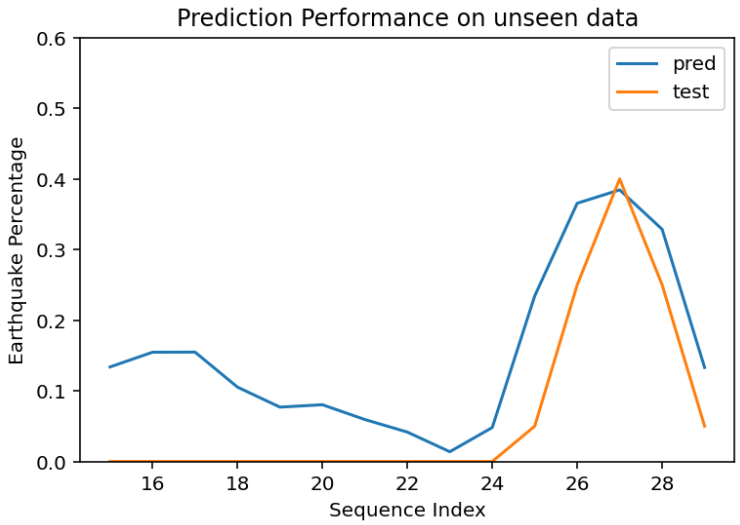


Figure 20 Prediction performance on event 2015-05-04 (model unseen data)

In figure 21, the entire predicted value was constant at about 0.15, and the value only changed at the end, but it went in the wrong direction.

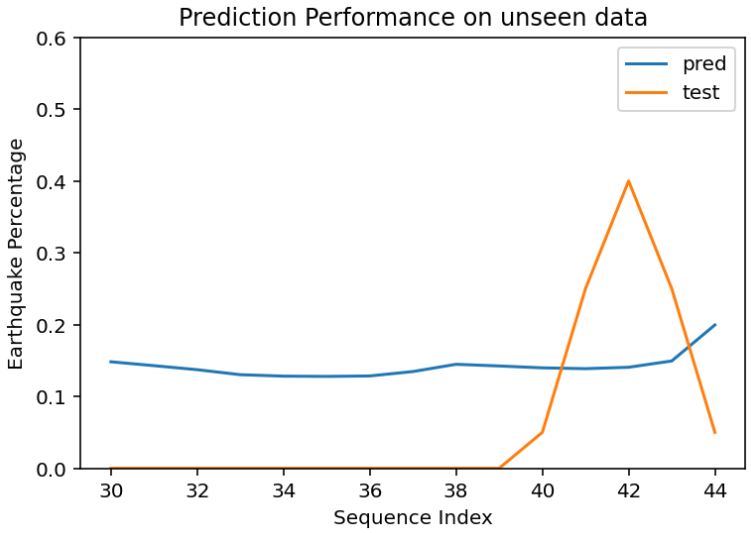


Figure 21 Prediction performance on event 2017-10-22 (model unseen data)

The overview boxplot in figure 22, different from the previous model, which more evenly spread out in the whole predicted range, these boxplots looked mode similar, while the predicted values packed together between 0.1 and 0.2, more points closer to 0.1, values greater than about 0.24 and smaller than approximately 0.06 are outliers.

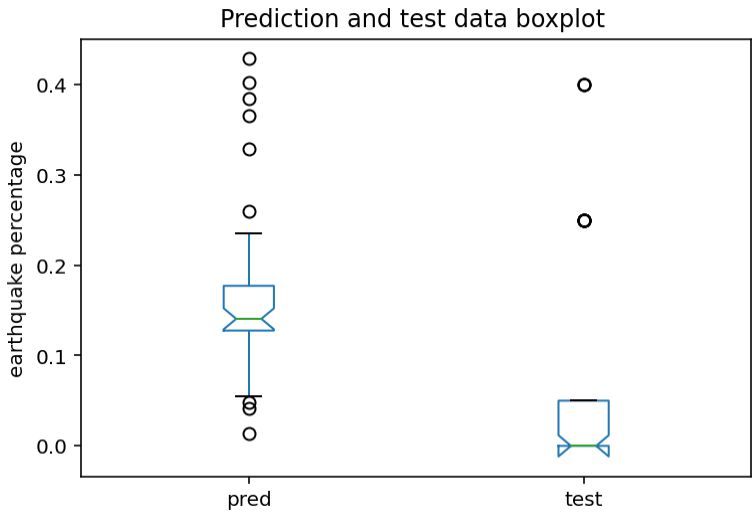


Figure 22

In figure 23, 'class' 0 is closer together than the previous model, most of the points were less than 0.2, others look like outliers. 0.05 'class' points were closer together and closer to the line. 0.25 'class' still spread between about 0.04 and 0.36. One of the 0.4 prediction points is much closer to the liner.

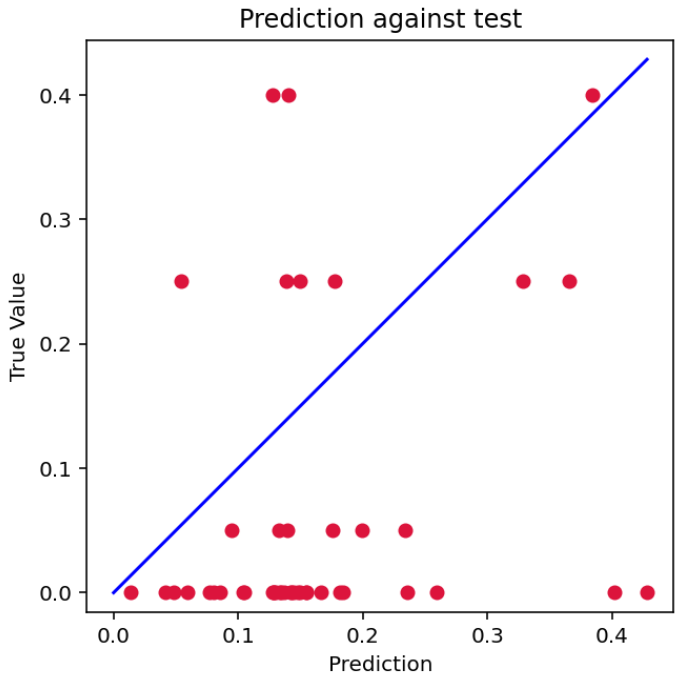


Figure 23

3.2 CNN LSTM model with 5 ConvLSTM2D layers

Figure 24 showed that both the accuracy and the beginning of both of the loss figures are slightly higher, and they are still constant during the training.

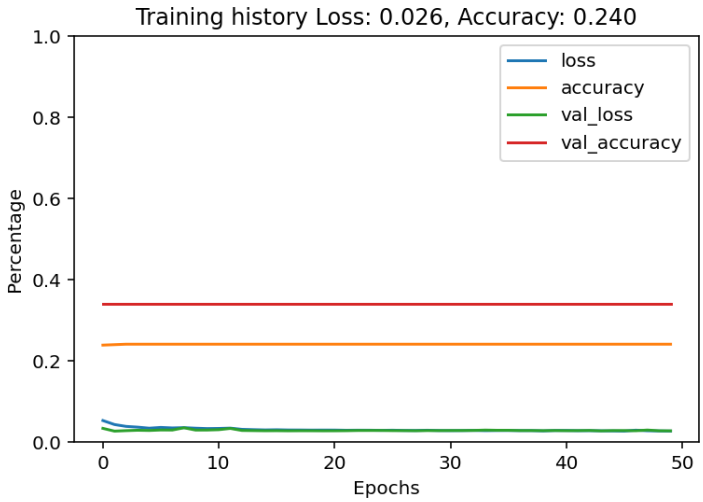


Figure 24 CNN LSTM model training loss and accuracy history

From figure 25 to 27, the model performed the best on the 2015 event, it confidently captured all the zeros and peaked at the same time with the value of approximate 2.22. But it started to show positive figures one sequence later from the original case. Figure 25 showed the curve shifted to the left and started with a large value. Figure 27 showed the model is confused in this case, the value started rising from the sixth sequence and the rising slope is kept until the end.

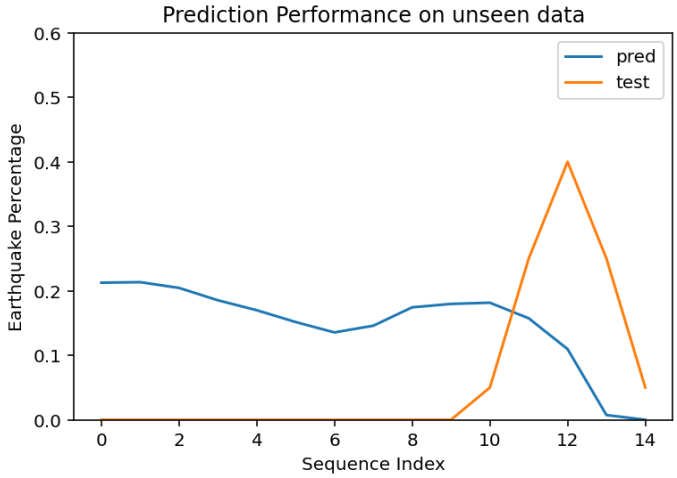


Figure 25 Model prediction performance on event 2013-12-16 (model unseen data)

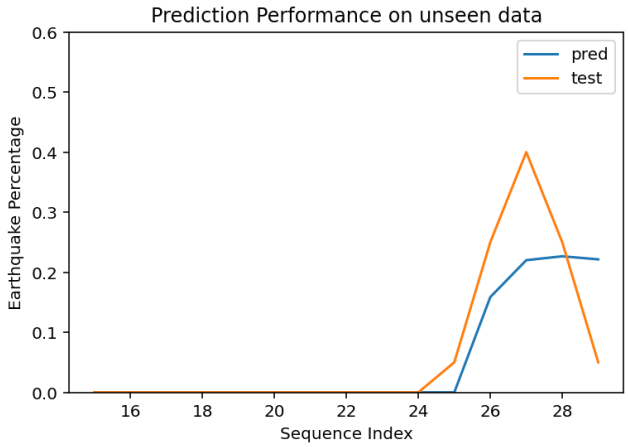


Figure 26 Prediction performance on event 2015-05-04 (model unseen data)

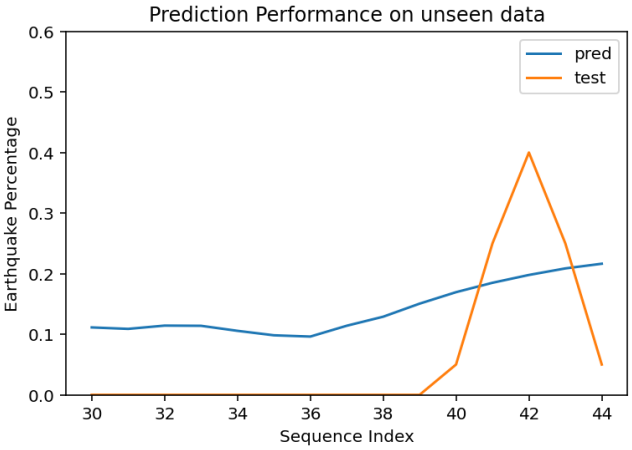


Figure 27 Prediction performance on event 2017-10-22 (model unseen data)

The boxplot showed most values are around about 0.15, with a smaller value range from 0 to about 0.23.

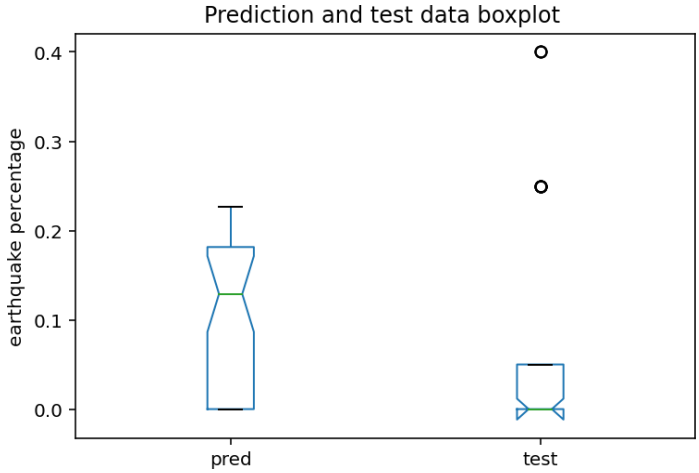


Figure 28

From figure 29, the model predicted values in a smaller range, it predicted some zero values correctly, but overall, it still lacked accuracy on the other value predictions.

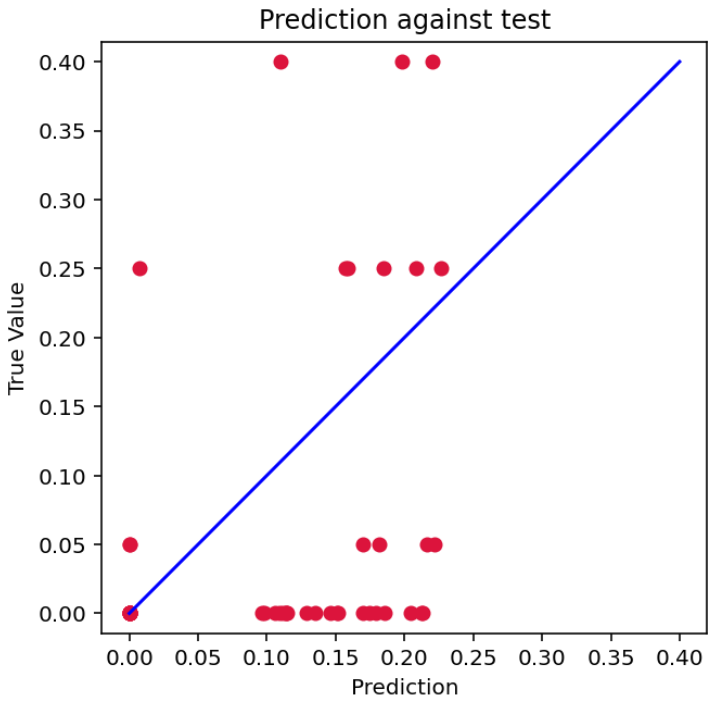


Figure 29

4 CONCLUSIONS

In conclusion, this CNN LSTM deep learning method didn't learn from the training data well, by dropping the loss and increasing the accuracy significantly. The model also didn't capture the underlying relationships between earthquakes and pre-earthquake ionospheric TEC anomalies overall. Statistically, it was not meaningful, and would not fit-for-purpose. From the unseen data model performance test, even the model could capture the sketchy trend of the curves, but it still performed poorly on the unseen data statistically in this regression problem. It could not be used in reality to predict earthquakes with a statistical confidence level.

However, this was only the first time deep learning methods were adopted in PETA research in New Zealand, with limited data, limited time and limited resources. In future works, data collection has a priority. Once all the data was collected, deep learning can be trained with all available data. Thrifty data would help the algorithms to learn and may perform better on the unseen data. Python generators should be embedded in the module to improve training efficiency.

In general, there were a lot of noises in the dataset. As has been seen in the temporal visualization data analysis, PETA varied in different events. These variations could due to solar activities or could be other undiscovered factors. For instance, there might be certain types of earthquakes, and different types of earthquakes might be caused by different geology

factors. For example, different geological structures and different geological layers in different regions, land and ocean, and other geological factors could vary the affection on ionospheric TEC. Correlations between earthquake depth, magnitude and TEC, would also be the variables that the research would include, to identify the difference between the earthquakes, might help the deep learning to recognize the correlation between earthquakes and the TEC variations in a better place.

Not just PETA, research on pre-earthquake radon and thoron emission was another popular topic in recent (Lim and Leong, 2428). These observations were also the key factors in earthquake forewarning research. Fetching the observations of these two elements, combined with TEC anomalies, will help the deep learning algorithms narrow down the truth in a great depth.

Finally, the quality, completeness and updating speed of RINEX observation GPS data of New Zealand is the foundation for this research. There were missing data in early 2020, RINEX observation data on 2018-10-28 might have quality issues (unable to be processed), missing data of some GPS ranging signal receivers on certain days, and the data for 2021 is not available until early February. Keeping a high standard of the data will assist the research primarily. Personally strong recommend GNS Science to continue the PETA research, it may have a great potential to discover.

5 REFERENCES

"About Us." GNS Science, https://www.gns.cri.nz/Home/About-Us. Accessed 25 Nov 2020.

"Ionosphere." Wikipedia, Wikimedia Foundation, 31 Dec. 2019, https://en.wikipedia.org/wiki/Ionosphere.

"free electron." Dictionary.com, https://www.dictionary.com/browse/free-electron.

Rhoades, D.A.; Mueller, C.; Buxton, R.; Gerstenberger, M.C. 2015. "Ionospheric Earthquake Precursors." GNS Science Consultancy Report. 2015/06. 26 p.

"RINEX." Wikipedia, Wikimedia Foundation, 14 Jan. 2021, https://en.wikipedia.org/wiki/RINEX.

"RINEX Data Archive." Land Information New Zealand, New Zealand Government, https://www.linz.govt.nz/data/geodetic-services/positionz/rinex-data-archive. Accessed 30 Nov 2020.

"PositionNZ." Land Information New Zealand, New Zealand Government, https://www.linz.govt.nz/data/geodetic-services/positionz/rinex-data-archive. Accessed 30 Nov 2020.

"Hatanaka Format Information at UNAVCO." Hatanaka, UNAVCO, https://www.unavco.org/data/gps-gnss/hatanaka/hatanaka.html. Accessed 10 Dec 2020.

"Data Policy." Geological hazard information for New Zealand, GeoNet, https://www.geonet.org.nz/policy. Accessed 18 Jan 2021.

"Description." Processing and archiving high-precision GPS data for the study of earthquake hazards, tectonic plate motion, crustal deformation and meteorology, SOPAC, http://sopac.ucsd.edu/sopacDescription.shtml. Accessed 18 Jan 2021.

"GPSTk Post-Processing." postprocessing, UNACO, https://www.unavco.org/software/data-processing/postprocessing/gpstk/gpstk.html. Accessed 20 Jan 2021.

Gaussiran, T.; Munton, D.; Harris, B., Tolman, B. 2004. "An Open Source Toolkit for GPS Processing, Total Electron Content Effects, Measurements and Modeling." Applied Research Laboratories, The University of Texas at Austin, International Beacon Satellite Symposium, October 2004, Trieste, Italy.

Lim, B. J. M. and Leong, E. C. “Challenges in the Detection of Ionospheric Pre-Earthquake Total Electron Content Anomalies (PETA) for Earthquake Forewarning.” Pure and Applied Geophysics. Springer Nature Switzerland AG, 2019, pp.2425-2449, https://doi.org/10.1007/s00024-018-2083-7.

Saha, Sumit. “A Comprehensive Guide to Convolutional Neural Networks — the ELI5 way.” Towards data science. https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53. December 16, 2018. Accessed 10 Jan 2021.

Brownlee, Jason. “CNN Long Short-Term Memory Networks.” Machine Learning Mastery. https://machinelearningmastery.com/cnn-long-short-term-memory-networks/. August 21, 2017, Vermont Victoria 3133, Australia. Accessed 10 Jan 2021.

“Sunspot Number”. SILSO, Sunspot Index and Long-term Solar Observations. http://www.sidc.be/silso/datafiles. Accessed on 08 Jan 2021.

“Keras vs Tensorflow.” Guru99. https://www.guru99.com/tensorflow-vs-keras.html. Accessed on 21 Jan 2021.

Ioffe, Sergey and Szegedy, Christian. “Batch Normalization: Accelerating Deep Network Training byReducing Internal Covariate Shift.” Machine Learning. chrome-extension://dagcmkpagjlhakfdhnbomgmjdpkdklff/enhanced-reader.html?openApp&pdf=https%3A%2F%2Farxiv.org%2Fpdf%2F1502.03167.pdf. March 2, 2015. Accessed 04 Jan 2021.

“Batch Normalization.” Wikipedia, Wikimedia Foundation, 14 Jan. 2021, https://en.wikipedia.org/wiki/Batch_normalization.

“Max Pooling.” Deep AI. https://deepai.org/machine-learning-glossary-and-terms/max-pooling. Accessed 04 Jan 2021.

Brownlee, Jason. “A Gentle Introduction to Dropout for Regularizing Deep Neural Networks.” Machine Learning Mastery. https://machinelearningmastery.com/dropout-for-regularizing-deep-neural-networks/. December 3, 2018, Vermont Victoria 3133, Australia. Accessed 05 Jan 2021.

6 APPENDICES

6.1 Appendix 1

30 largest earthquake events from 2003 in New Zealand (not a aftershock within 3 days)

Public ID	Event type	Origintime	Longitude	Latitude	Magnitude	Depth	
5342	2020p092995	earthquake	2020-02-04T09:47:26.309Z	176.54582	-37.62954	5.52740	256.93402
12149	2019p738432	earthquake	2019-10-01T05:16:41.168Z	177.77242	-39.15775	5.65211	22.00418
19617	2019p428761	earthquake	2019-06-08T15:24:08.722Z	168.15244	-44.34042	5.52366	5.00000
33644	2018p816466	earthquake	2018-10-30T02:13:41.648Z	175.00746	-39.03259	6.15735	207.14906 Taranaki
4 55070	2017p795065	earthquake	2017-10-22T04:10:59.674Z	173.45894	-42.35814	5.58879	8.27472
5 80252	2016p981371	earthquake	2016-12-29T02:34:33.905Z	174.46187	-41.54913	5.55976	16.01562
6 84549	2016p913880	earthquake	2016-12-04T03:15:47.108Z	174.23567	-41.69030	5.51542	10.65643
8 88454	2016p881118	earthquake	2016-11-22T00:19:42.525Z	177.22881	-40.67358	5.54864	29.11194
93724	2016p858000	earthquake	2016-11-13T11:02:56.346Z	173.02214	-42.69254	7.82038	15.11445 Kaikoura
110550	2016p118944	earthquake	2016-02-14T00:13:43.988Z	172.75459	-43.49732	5.70180	8.00000
28 110840	2016p105478	earthquake	2016-02-09T00:39:00.115Z	173.07483	-42.06830	5.74514	48.12500
29 117481	2015p768477	earthquake	2015-10-12T08:05:01.717Z	176.32572	-40.57807	5.69113	23.28125
125824	2015p332712	earthquake	2015-05-04T02:29:10.430Z	168.83693	-44.53661	5.79217	4.29440
31 126373	2015p305812	earthquake	2015-04-24T03:36:42.412Z	173.06804	-42.09022	6.23312	51.52344
133334	2015p012816	earthquake	2015-01-05T17:48:41.752Z	171.25201	-43.05787	5.95127	5.11719
154656	2014p051675	earthquake	2014-01-20T02:52:45.984Z	175.86497	-40.62058	6.22976	34.17969
156290	2013p944608	earthquake	2013-12-16T12:07:24.965Z	166.06284	-46.16135	6.13638	5.00000
38 164563	2013p613797	earthquake	2013-08-16T02:31:05.926Z	174.16166	-41.70171	6.46877	7.46094
169016	2013p537512	earthquake	2013-07-18T21:06:39.068Z	174.40459	-41.54557	5.69714	20.64453 Seddon

42 179897 2012p923684 earthquake 2012-12-07T18:19:07.412Z 176.06412 -38.33440 5.59734 167.53906
187964 2012p498491 earthquake 2012-07-03T10:36:14.209Z 173.76487 -40.04852 6.18403 241.01562
45 198959 3631359 earthquake 2011-12-23T00:58:38.147Z 172.79951 -43.48138 5.80000 10.01470
46 209392 3528839 earthquake 2011-06-13T02:20:49.260Z 172.73773 -43.57064 6.00000 6.91600
48 218825 3468575 earthquake 2011-02-21T23:51:42.320Z 172.67848 -43.58213 6.20000 5.41430 Christchurch
220452 3456063 earthquake 2011-01-27T15:02:46.965Z 176.00592 -38.41254 5.60000 151.02300 Christchurch no pattern
232142 3366146 earthquake 2010-09-03T16:35:41.836Z 172.16794 -43.52731 7.20000 11.04260 Darfield
257184 3128974 earthquake 2009-07-15T09:25:20.223Z 166.18684 -46.01908 5.62500 12.00000 only one pattern showing on 17-08-2010
290542 2839343 earthquake 2007-12-20T07:55:16.822Z 178.53700 -38.89009 6.70000 33.00000
328908 2484105 earthquake 2005-11-08T22:51:53.517Z 176.29129 -37.58774 5.57900 291.60270
366947 2150155 earthquake 2003-11-24T07:43:48.334Z 176.65959 -37.51037 5.51600 195.79770

6.2 Appendix 2

18 major earthquakes TEC anomaly correlation images anomaly patterns

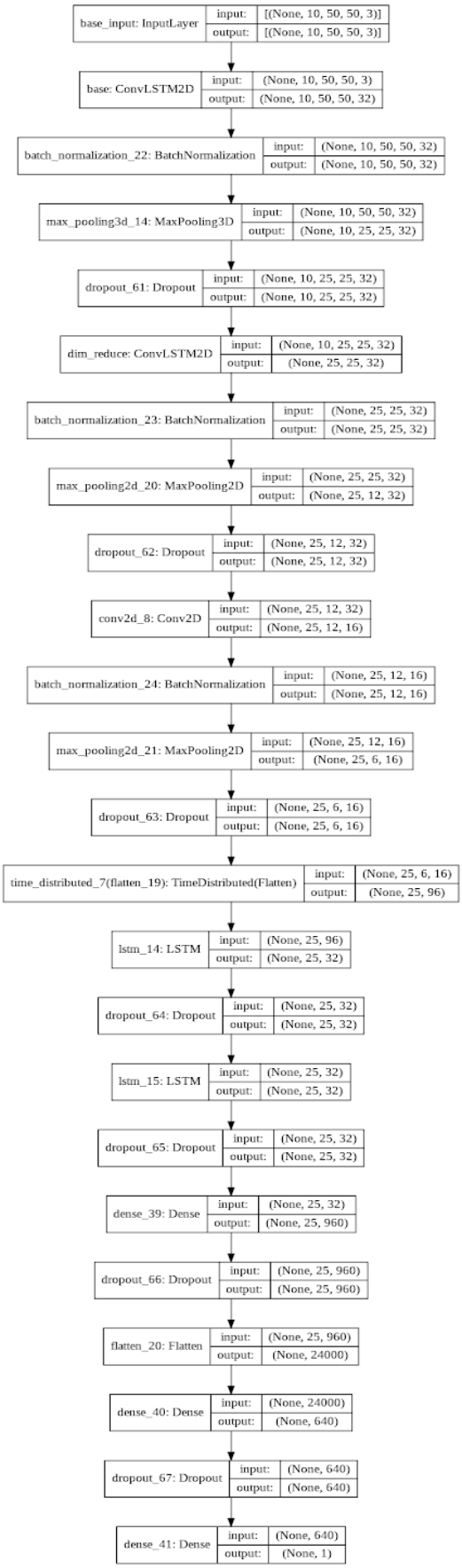
(n = no pattern, w = weak pattern, s = strong pattern)

Event time (UTC)	Location (Lat., Long)	Depth (km)	Magnitude (M_L)	Number of days to the event / Visualized anomaly pattern																							
2019-10-01 05:16:41	-39.16, 177.77	22.00	5.65	21	20	19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	0	1	2
				n	n	n	w	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n
2019-06-08 15:24:08	-44.34, 168.15	5.00	5.52	21	20	19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	0	1	2
				w	s	w	s	s	s	s	s	w	s	s	w	w	s	s	s	s	s	s	s	w	w	n	w
2018-10-30 02:13:41	-39.03, 175.00	207.15	6.16	21	20	19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	0	1	2
				n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	s	s	w	w	w	w	n	w
2017-10-22 04:10:59	-42.36, 173.46	8.27	5.59	21	20	19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	0	1	2
				n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	w	n
2016-11-13		15.11	7.80	21	20	19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	0	1	2

11:02:56	-42.69, 173.02			n	n	n	w	n	n	n	n	w	n	n	n	n	n	w	n	w	w	w	n	w	n	n	n
2016-02-14	-43.50,	8.00	5.70	21	20	19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	0	1	2
0:13:43	172.75			w	n	n	n	w	w	w	w	w	w	n	s	w	w	w	w	w	w	w	n	w	s	n	w
2015-05-04	-44.54,	4.29	5.80	21	20	19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	0	1	2
2:29:10	168.84			w	n	w	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n
2015-01-05	-43.06,	5.12	5.95	21	20	19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	0	1	2
17:48:41	171.25			n	w	s	n	w	n	w	n	w	n	w	n	w	n	n	n	w	w	w	w	w	w	n	n
2014-01-20	-40.62,	34.18	6.23	21	20	19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	0	1	2
2:52:45	175.86			w	w	n	n	s	n	n	n	w	n	w	n	w	w	w	n	w	s	s	w	s	s	w	w
2013-12-16	-46.16,	5.00	6.14	21	20	19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	0	1	2
12:07:24	166.06			n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	s	n	n	n	n	n	n	n	n
2013-07-18	-41.55,	20.64	5.70	21	20	19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	0	1	2
21:06:39	174.40			n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n
2012-07-03	-40.05,	241.02	6.18	21	20	19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	0	1	2
10:36:14	173.76			n	n	n	n	n	n	n	n	n	n	n	n	n	w	n	n	n	n	n	w	n	n	n	n
2011-02-21	-43.58,	5.41	6.2	21	20	19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	0	1	2
23:51:42	172.68			w	n	n	n	n	n	n	n	n	n	w	n	w	w	w	n	w	w	s	w	w	w	n	n
2011-01-27	-38.41,	151.02	5.60	21	20	19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	0	1	2
15:02:46	176.01			n	n	w	w	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	w	n	n	n
2010-09-03	-43.53,	11.04	7.20	21	20	19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	0	1	2
16:35:41	172.17			n	n	n	n	w	w	w	n	w	n	n	n	n	n	w	n	n	n	n	n	n	n	w	n
2009-07-15	-46.02,	12.00	5.63	21	20	19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	0	1	2
9:25:20	166.19			w	s	s	w	s	w	s	w	w	w	s	s	s	w	w	w	s	s	s	s	w	w	w	w
2007-12-20	-38.89,	33.00	6.70	21	20	19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	0	1	2
7:55:16	178.54			n	n	n	w	w	n	n	n	n	n	n	n	n	n	w	n	w	n	w	w	w	w	n	n
2003-11-24	-37.51,	195.8	5.52	21	20	19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	0	1	2
7:43:48	176.66			n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n

6.3 Appendix 3

CNN LSTM model with Conv2D and LSTM layers



6.4 Appendix 4

CNN LSTM model with 5 ConvLSTM2D layers