EXPLORATIONS INTO MACHINE LEARNING TECHNIQUES FOR PRECIPITATION NOWCASTING

A Thesis Outline Presented

by

ADITYA NAGARAJAN

Submitted to the Graduate School of the
University of Massachusetts Amherst in partial fulfillment
of the requirements for the degree of

MASTER OF SCIENCE

May 2016

Mechanical and Industrial Engineering

EXPLORATIONS INTO MACHINE LEARNING TECHNIQUES FOR PRECIPITATION NOWCASTING

A Thesis Outline Presented

by

ADITYA NAGARAJAN

Approved as to style and content by:

Michael Zink, Chair

David L. Pepyne, Member

Hari Balasubramanian, Member

Jonathan Rothstein, Graduate Program Director
Mechanical and Industrial Engineering

ACKNOWLEDGMENTS

This work was performed within the UMass Center for Collaborative Adaptive Sensing of the Atmosphere (CASA) under funding provided by the Jerome M. Paros Endowment for Measurement Science Research.

ABSTRACT

EXPLORATIONS INTO MACHINE LEARNING TECHNIQUES FOR PRECIPITATION NOWCASTING

MAY 2016

ADITYA NAGARAJAN

B.E, MADRAS INSTITUTE OF TECHNOLOGY

M.Sc., UNIVERSITY OF MASSACHUSETTS, AMHERST

Directed by: Dr. Michael Zink

Significant advances in cloud based big data technologies now makes data driven solutions feasible for increasing numbers of scientific computing applications. One such scientific computing application is machine learning where patterns in large data sets are brought to the surface by finding complex mathematical relationships within the data. Nowcasting or short term prediction of rain fall intensity in a given region is an important problem in meteorology. However there has not been much work on precipitation nowcasting using machine learning techniques and state of the art nowcasting systems today are based on an underlying model. We thus explore the nowcasting problem from a data science perspective by formulating it as a machine learning problem. We take advantage of a fundamental relationship between water vapor and rainfall in that water must first exist in the atmosphere as water vapor before it can fall to the ground as rain. Using the GPS-Meteorology technique to measure the water vapor in the atmosphere and weather radar reflectivity

to measure rainfall, this thesis proposes to explore the use of number of machine learning techniques to predict from the joint spatial-temporal evolution of these two complementary quantities to predict the rainfall field 0-2 hours into the future.

TABLE OF CONTENTS

LIST OF TABLES

Table **Page**

LIST OF FIGURES

Figure **Page**

CHAPTER 1

INTRODUCTION

We live in an age where large volumes of data generated each day can provide us with valuable insights to the underlying system the data is representing. Coupled with significant advances in big data technology, we are now able to extract complex patterns from high-dimensional data within reasonable time. Mastering the ability to processes large data sets to extract actionable information using sophisticated machine learning algorithms has aroused significant interest in the field of data science. In this thesis we explore the short-term weather prediction (aka nowcasting) problem from a data science perspective. Specifically this thesis will seek to apply data science and machine-learning techniques to the problem of nowcasting precipitation fields 1-3 hours in the future from time-sequences of past spatial fields of precipitable water vapor and weather radar reflectivity. Reflectivity being the measure of location and intensity of precipitation and precipitable water vapor (otherwise termed as Integrated Precipitable Water or IPW) a measure of the amount of water in the atmosphere that could potentially fall as precipitation. Thus, where IPW tells about the potential for precipitation, weather radar reflectivity tells us where precipitation is currently occurring. Using the complementary nature of these two measurements we seek to explore the spatiotemporal patterns and correlations between these fields to make 1-3 hour precipitation nowcasts through the use of various machine learning algorithms. In addition to the challenge of obtaining and pre-processing the input data for machine-learning, we also face the challenge of developing a machine learning algorithm that can handle spatiotemporal input data - essentially short video streams

of IPW and radar reflectivity fields to predict the precipitation field 1-3 hours in the future.

1.1 Precipitation Nowcasting

Because rain affects so many human activities, predicting rain has a long history. Whereas long-term rainfall forecasts (i.e., beyond 3 hours in the future) are based on models of atmospheric processes (cf. [32]), short-term 1-3 hour nowcasts are frequently based on weather radar data. This is because numerical weather models tend to be too computationally time-consuming for real-time predictions. Nowcasts can be "manual" as when a weather radar meteorologist plays a radar reflectivity loop to infer where and how fast a storm is moving. Nowcasts can be automated as with the Storm Cell Identification and Tracking (SCIT) algorithm [30] uses the history radar data to identify storm cells and estimate their speed and direction or the Dynamic Adaptive Radar Tracking of Storms (DARTS) algorithm [41] that uses the history of radar data to infer its rate of advection in order to project the reflectivity field 1-20 minutes into the future. Nowcasts based on weather radar data alone tend to quickly break down, so that a 20-minute DARTS nowcast of the reflectivity will often bear little resemblance to the actual reflectivity field that occurs 20 minutes in the future. This is because nowcast techniques based on weather radar data alone, while they can obtain a good estimate of storm advection, have very little skill in predicting storm growth and decay. They are thus very poor at predicting that a storm will pop-up at a given location when there is currently no radar data coming from that location, and they are very poor at recognizing that a storm will dissipate 20 minutes from now when it is currently growing in intensity. This has lead many to look for other information with which to augment weather radar data.

Since the discovery in the early 1990s that GPS signal propagation delays can be used to infer atmospheric water vapor content [10] [9] there has been significant

maturation in GPS-Meteorology (GPS-Met) technology and techniques. NOAA (National Oceanic and Atmospheric Administration), UCAR (University Collaboration of Atmospheric Research) and SOPAC (Scripps Orbital and Permanent Array Center) have contributed to the development, operation and maintenance of a nationwide realtime GPS-based water vapor monitoring system [55] [11]. This has lead to the availability of real-time water vapor products to the public and operational forecasters from over 500 GPS-Met stations distributed around the continental United States. In addition, it is now also possible to obtain mature, validated software for GPS-Met calculations (e.g., [26]) allowing for research deployments of GPS-Met stations [1] and repurposing of GIS Continuously Operating GPS Reference Stations (CORS http://geodesy.noaa.gov/CORS/) stations for the GPS-Met application.

Motivated by the fact that in the water cycle that describes the movement of water through the atmosphere, water must first exist in the atmosphere as water vapor before it becomes rain, a number of agencies and researchers have explored the potential of real-time observations of atmospheric water vapor content derived from GPS-Met stations for weather forecasting and precipitation nowcasting. Japan thus far has the highest density of GPS stations with an average spacing between stations of 17km [48]. Using this densely spaced GPS network, a number of studies have looked into the ability of high spatial-temporal resolution GPS-Met derived Integrated Precipitable Water Vapor (IPW) to nowcast thunderstorms and severe rain, e.g., [29], [43]. In a study of how IPW fields relate to the onset of convective weather [28] showed that the maximum IPW occurred 1-2 hours prior to thunderstorm activity where thunderstorms were measured using cloud-to-ground lightning and convective activity was measured by hourly accumulated rainfall. In another study looking at relationships between spatial variations in IPW and precipitation it was shown that rainfall intensity is related to IPW gradients and in particular that strong conver-

gence (concentration) of water vapor is generally present several hours in advance of convective precipitation [56].

Spatial variations in IPW and their correlations with thunderstorm activity have also been studied in Europe. De Haan [15] devised a method to construct IPW maps from a network of GPS stations using two dimensional variational techniques. The IPW maps were then studied with regard to lightning and thunderstorm events in the Netherlands. The level of convergence of water vapor evident in the IPW fields again correlated well with subsequent precipitation rates and thunderstorm activity. A similar analysis was conducted in Spain [53] using normalized IPW fields to take into account the seasonal variation. The interesting observation made in this paper is that the decay of IPW values coincided with storm direction and intensity.

Generation and validation of IPW spatial fields have also been carried in the US and the potential use of IPW fields for analyzing long and short term climatological activity have been studied (cf.[38]). These studies have shown that IPW fields generated from point measurements using GPS-Met systems prove to be an excellent tool to visualize convergence and build up of water vapor. They observed a strong build-up of water vapor 1-3 hours in advance of a convective event, leading to the conclusion that IPW holds the potential to accurately predict convective initiation.

A common indicator of rainfall is the steep increase in IPW seen from individual GPS-Met stations. [46] [2] showed that variations in IPW usually peaks a few hours prior to the onset of precipitation and that the variations in IPW and the variations of rainfall rate measured by rain gauges are also strongly correlated. However both papers also note that IPW alone is not a precise predictor of rainfall and there are other atmospheric parameters that play a role in the onset of rain. This observation is based on peak IPW values encountered at times which were not followed by rainfall.

Based on the above findings, we propose a precipitation nowcasting approach that uses both IPW and weather radar reflectivity. Rather than attempting a prediction

system that tries to model how these to fields jointly evolve, we propose instead a machine learning approach to learn the joint spatial-temporal patterns and correlations.

The use of machine learning for precipitation nowcasting is not new. An early application of machine learning to the precipitation nowcasting problem used artificial neural networks to predict rainfall fields 1 hour in advance based on the current rainfall field [20]. A single layer neural network was trained using a 100×100 km^2, 4 km resolution precipitation field to predict the corresponding precipitation field 1 hour in the future. The predictions were evaluated against mean areal index (MAI) and PAC (Percent Areal Coverage) and performed as well as the standard nowcasting algorithm of the time. A neural network approach was also applied to forecasting rain gauge readings with 0-6 hr lead times using moisture and updraft data from the NCEP (National Center for Environmental prediction) Nested Grid Point Model. A feature selection method was used to reduce the initial 528 feature input space to the 25 most important features [31].

A machine learning approach using random forests and logistic regression was used in [35] to make probabilistic 1-hour ahead predictions of convective storms. The inputs were GOES satellite data and NWP data and the problem was one of determining if a particular cloud would turn into a convective storm. As part of the analysis, this paper used the feature importance attribute of the random forest method to determine relative importance of each feature in predicting convective initiation. In an another study a modified version of the random forest called the Spatiotemporal Relational Random Forest (SRRF) [34] was used to detect turbulence areas for aircraft. This algorithm used archived NWP estimates of the weather and meteorological observations. This algorithm achieved a 0.80 area under the curve detection rate of turbulence with 100 trees in the forest.

The state-of-the-art machine learning algorithms of today fall under the category of "Deep Learning" [4]. The word "Deep" was coined to emphasize the fact that this type of learning algorithm tries to learn multiple layers of representations of the input data space. These multiple layers of representations take advantage of the spatial correlations in an input image or the temporal correlations of time series data. Some of the well known deep learning architectures include Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN) and LSTM (Long Short Term Memory). A very recent application of Deep Learning and one of the first in the meteorological domain to tackle the nowcasting problem used a CNN-LSTM, a combination of CNN and LSTM, to predict rainfall fields based on spatiotemporal sequences of radar reflectivity products [47]. Specifically, the goal was to predict the next 15 frames of reflectivity from the previous 5. The algorithm was trained and tested on a dataset of reflectivity echoes from a radar in Hong-Kong for rainfall days which occurred over a period of three years. The results showed that the CNN-LSTM performs better than the current model based state-of-the-art nowcasting algorithm called the Real-time Optical flow by Variational methods for Echoes of Radar (ROVER).

In this thesis we attempt to go a step beyond the works cited above (which use reflectivity only, IPW only, or use NWP data), to develop a machine-learning precipitation nowcasting system based on direct observations of both IPW (water vapor) and reflectivity (precipitation).

1.2 Proposal Organization

The remainder of this proposal is organized as follows. Chapter 2 will discuss the hydrologic cycle of how water moves through the atmosphere and the theory behind the two remote-sensing systems (GPS-Met and Weather Radars) that we will use for our studies. Chapter 2 also gives a brief description of the interpolation technique we use to build IPW fields from IPW point measurements. Chapter 3 will discuss the

region of Texas that will be the focus of our studies and how we collected, organized, and processed the data set we will use for our experiments. Chapter 4 will discuss our initial nowcasting results. The proposal ends in Chapter 5 with a brief outline of the additional work we propose to do for the thesis and a timeline for its completion.

CHAPTER 2

BACKGROUND

In the hydrologic cycle that describes the movement of water through the atmosphere, water must first exist as water vapor before it precipitates back to the earth as rain. After a simplified explanation of the how water vapor becomes precipitation, this chapter describes the instruments we will use to measure atmospheric water vapor and rain.

2.1 Mechanisms of Precipitation

In the hydrologic cycle shown in Figure 2.1, water enters the atmosphere as vapor through evaporation and transpiration. Humidity, measured in mass of water per mass of volume of atmosphere, is a measure of the amount of water vapor present in the atmosphere. The total amount of water vapor the atmosphere can hold is mainly a function of temperature. Relative humidity gives the saturation percentage - 0% implies no water vapor at all in the atmosphere, 100% implies the atmosphere is fully saturated and can hold no more. The dew point is the temperature at which the relative humidity becomes 100%.

Precipitation forms when there is uplift that forces warm moist air into increasingly colder air aloft. As the moist air is lifted, the relative humidity increases to 100% at which point the water vapor begins to condense into water droplets. Very small droplets (0.01 mm in size) remain suspended in the atmosphere to become clouds. The height of the cloud base being roughly the height where the temperature is equal to the dew point (a "cloud base" at ground level gives fog). For small

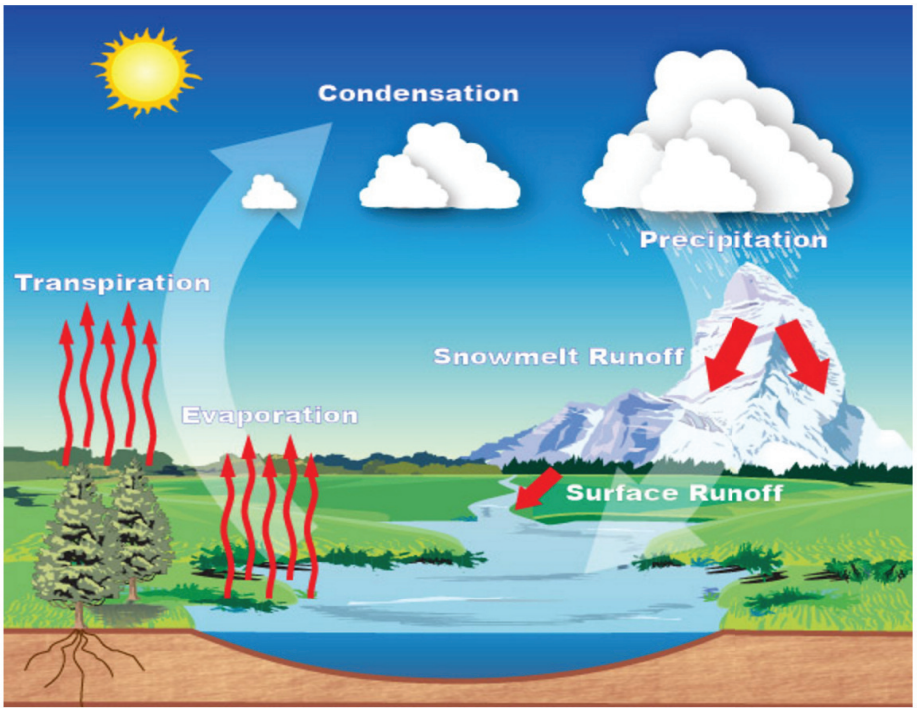

Figure 2.1: The basic hydrological cycle (www.srh.noaa.gov/jetstream/atmos/hydro.htm).

cloud droplets to become larger, they need a surface (condensation nuclei), in the form of dust, pollen or frozen ice crystals to coalesce onto. Once the water droplets become sufficiently large (e.g., greater than 0.1 mm in size), the vertical motion of the atmosphere can no longer hold them aloft and they begin to fall to earth as precipitation.

Figure 2.2 shows the two basic mechanisms leading to precipitation. The left of the figure shows a cold front, where convective uplift is caused by cold air being forced into warm moist air. This is the mechanism of thunderstorms and the supercells that can spawn tornadoes. The right of the figure shows a warm front where warm moist air is dynamically forced into cold air. This is the mechanism of less violent and more widespread stratiform rain. For more detailed explanations of the hydrologic cycle, atmospheric water vapor, and the mechanisms of precipitation, the reader is referred to [6] and the summaries in [45] and [16].

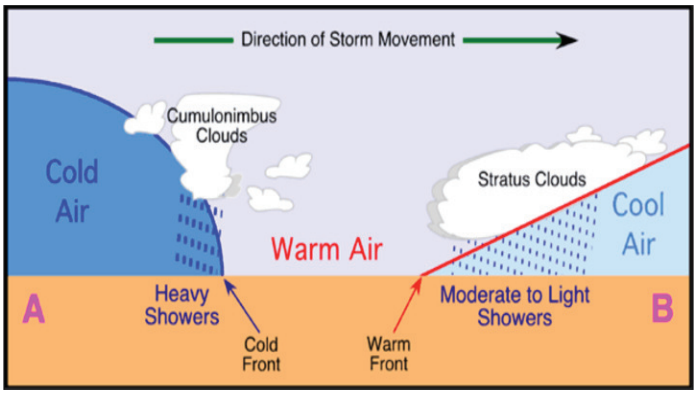

Figure 2.2: Basic mechanisms of precipitation - warm moist air uplifted by or into colder air. On the left is a cold front leading to convective lift and thunderstorms. On the right is a warm front leading to dynamic lift and stratiform rain. Figure from www.physicalgeography.net.

2.2 Measuring Atmospheric Water Vapor

There are a number of instruments for measuring the amount of water vapor in the atmosphere. These include:

1. Radiosondes: A radiosonde is a battery powered telemetry instrument package with sensors for sampling various atmospheric variables as it is carried up by a balloon from ground launch to between 20-30 km altitude[1]. While radiosondes have the advantage that they can measure the vertical distribution of water vapor, they have the distinct disadvantages that they are typically only launched two times a day (at 0000 and 1200 UTC) and from only a handful of locations (the entire CONUS is covered by a mere 90 radiosonde launch sites).

2. Radiometers. Ground-based water vapor radiometers measure the background microwave radiation emitted by atmospheric water vapor along a given line of site [7]. An advantage of these instruments is their ability to make continuous measurements of water vapor. Disadvantages are cost, calibration, and sparse

[1] http://www.wrh.noaa.gov/rev/tour/UA/introduction.php

spatial deployment. They are also limited in that they do not work when it is raining.

3. Satellites. The GOES (Geostationary Operational Environmental Satellite) system provides two sources of information about the water vapor [19]: imagery through its water vapor channel (at 4km spatial resolution, 15 min temporal resolution), and sounder retrievals (at 20km spatial resolution, 1hr temporal resolution). Both of these observations, however, are negatively impacted by cloud cover.

4. GPS-Meteorology. GPS-meteorology (GPS-Met) is a technique that allows GPS receivers to simultaneously perform the multiple functions of position estimation and precipitable water vapor estimation [10]. For a given GPS receiver, precipitable water vapor estimates can be made with 30-minute temporal resolution. In regions, such as the middle and western U.S., where there is a high density of Continuously Operated GPS Reference Stations (CORS), techniques have been developed to combine the water vapor measurements from multiple stations into 2D and 3D water vapor fields. The spatial resolution of the field depends on the density of GPS stations (spatial Nyquist). While GPS-Met currently cannot provide the spatial and temporal resolution of that of GOES satellite, it has the advantage that it is accurate in all weather conditions and not impacted by clouds or precipitation.

Based on our previous work [1], where we developed low-cost GPS-Met systems for near real-time Integrated Precipitable Water Vapor (IPW) estimation and an infrastructure for disseminating the IPW data on-line, we will use the GPS-Met technique as our source of atmospheric water vapor information.[2].

[2]UMass operates two low-cost GPS-Met stations in the Dallas-Fort-Worth, Texas metroplex region, one (site designation CNVL) at the Univ. of Texas at Arlington and another (site designation

2.2.1 GPS Meteorology Technique

The Global Positioning System (GPS) is a system of satellites operated by the U.S. Department of Defense (DoD). First launched in the 1970s for the purpose of military navigation, the system was later opened up for civilian use. The GPS system consists of a core of 24 satellites flying at 22,200 ft AGL orbiting in 6 different orbital planes inclined at 55° to each other [3]. For a GPS receiver located in the CONUS the number of satellites in view at any one time ranges between 8 and 12, though only 4 are required for an estimate of horizontal and vertical position.

The signal path from a given satellite to a GPS receiver is called a slant path. Since GPS satellites are not in geosynchronous orbit, the azimuth and elevation angles of the slant paths to the satellites in view change with time, as does the particular set of satellites in view[4]. Along each slant path a GPS receiver receives carrier signals at two distinct frequencies L_1 ($f_1 = 1575.42MHz$) and L_2 ($f_2 = 1227.60MHz$). These carrier signals are modulated as a sequence of bits called Pseudo Random Noise (PRN) and each satellite is identified by a unique PRN code [27]. From the carrier signals (code and carrier phase) the GPS receiver obtains measurements of the distance (pseudorange) between the satellite and the receiver. As the GPS carrier signals travel from a satellite to a receiver, they accumulate delays that cause the pseudoranges to accumulate errors,

$$P_r^s(t_r) = \rho_r^s - (\delta t_r - \delta t^s) * c + \delta_{r,ion}^s + \delta_{r,trop}^s + \xi \tag{2.1}$$

Here P_r^s is the code pseudorange measurement from satellite s to receiver r, ρ_r^s is the geometric distance as a function of the receiver and satellite coordinates, and

---

NWSD) at the NWS Dallas-Fort-Worth Weather Forecast Office (WFO). IPW observations from these two sites is published on-line at http://emmy9.casa.umass.edu/gpsmet/2015/)

[3] http://www.gps.gov/systems/gps/space/

[4] see the animation at wiki page for GPS https://en.wikipedia.org/wiki/Global_Positioning_System

the rest of the terms are range corrections - δt^s and δt_r the clock corrections for the satellite and receiver respectively, c is the speed of light in vacuum, $\delta^s_{r,ion}$ is the correction for the signal delay through the electrically charged ionosphere, $\delta^s_{r,trop}$ is the correction for the refractivity induced delays through the troposphere, and ξ are residual corrections for things like multipath delay and receiver and satellite hardware biases (cf. [45]).

For the geodest, position accuracy is limited by the accuracy of the clocks and the accuracies of the ionospheric and tropospheric corrections. For the meterologist, its the tropospheric correction that is of interest, since the tropospheric delay is the term that varies with atmospheric water vapor content [10] [40] [18]. Estimating the tropospheric delay from the observed code range and carrier phase requires estimating and subtracting the other correction terms. For high accuracy, such as required by the GPS-Met application, the so-called double differencing technique is often used [39], [3]. This involves taking the differences of the pseudorange equations between two receivers and two different satellites and then taking the difference of these differences. If the baseline distance between the two receivers is sufficiently large ($> 500km$) that the observables are uncorrelated, then the result of double differencing is the elimination of both satellite and receiver clock errors (the $(\delta t_r - \delta t^s) * c$ term in eq. 2.1). For the electrically charged ionosphere, which is the region of the atmosphere between 60 and 1000 km altitude, the ionospheric delay (the $\delta^s_{r,ion}$ term in eq. 2.1) is frequency dependent (dispersive) [50]. This delay, which amounts to between 1 and 15 meters of pseudorange error, can be estimated to millimeter precision via linear combinations of the GPS dual frequency observables [40].[5] Residual errors (ξ in eq.

[5]The lowest cost GPS receivers, such as those in cell-phones, are single frequency receivers that use L_1 to estimate position. It is because these receivers cannot correct for the large ionospheric delay as accurately as dual-frequency receivers that they are not generally used for the GPS-Met application.

2.1), such as due to multipath, are avoided by careful selection of GPS site to avoid obstructions such as cell phone towers and buildings.

What remains after applying the above corrections is the tropospheric delay. The troposphere is the lowest portion of the atmosphere from the earth's surface to about 17 km and is the site of all weather on earth. The excess path length that GPS signals travel in the troposphere is due to refraction and can reach up to 2.5 meters at sea-level. This excess path length is given by [10], [45],

$$\delta^s_{r,trop} = 10^{-6} \int N ds + (S - G) \tag{2.2}$$

where N is the refractivity, S is the actual signal path and G is the geometric signal path respectively along the slant path between satellite s and receiver r, and the integral is along the slant path. The refractivity, N, can be split into a dry part, N_h, (refractivity of dry air) and a wet part, N_w, (refractivity due to water vapor) [14], [42],

$$\delta^s_{r,trop} = 10^{-6} \int (N_h + N_w) ds + (S - G) \tag{2.3}$$

The refractivity's depend on pressure, temperature, and humidity according to,

$$N_h = 77.6(\frac{P_d}{T}), N_w = 64.8(\frac{P_w}{T}) + 3.776 \cdot 10^5(\frac{P_w}{T^2}) \tag{2.4}$$

where P_d and P_w are the partial pressures (in millibars) of dry air and water vapor respectively, and T is the surface temperature (in degrees Kelvin).

A GPS-Met estimation of the right hand side of eq. 2.3 proceeds as follows. The tropospheric delay along a slant path is called the Slant Total Delay (STD). This is broken into a Slant Hydrostatic Delay (SHD) term and a Slant Wet Delay (SWD) term,

$$STD = SHD + SWD \tag{2.5}$$

Of these, the SHD accounts for the majority of the excess path length, or about 2 meters, while the SWD accounts for only about 1-2 meters of excess path length. The STD is commonly written in terms of the hydrostatic and wet delays in the zenith direction as (cf.[18]),

$$STD = m_h(\theta)ZHD + m_w(\theta)ZWD \tag{2.6}$$

where $m_h(\theta)$ and $m_w(\theta)$ are mapping functions (inversely proportional to the sine of the slant path elevation angle θ) for the hydrostatic and wet components respectively. For elevation angles above 15° the hydrostatic and wet mapping functions are essentially equal allowing us to write,

$$STD = m_n(\theta)ZTD \tag{2.7}$$

where,

$$ZTD = ZHD + ZWD \tag{2.8}$$

is the Zenith Total Delay. The ZTD is determined from the measured STDs [54] and the ZWD is determined as,

$$ZWD = ZTD - ZHD \tag{2.9}$$

where the ZHD is a slowly varying quantity that can be estimated to a fraction of a millimeter via the so-called Saastamoinen model [42],

$$ZHD = \frac{0.00227768P_0}{1 - 0.00266cos(2\phi) - 0.00028h_{ref}} \tag{2.10}$$

where P_0 is the surface pressure (in millibars), h_{ref} is the geodic height of the station (in meters) and ϕ is the station latitude. Given the ZWD, the Integrated Precipitable Water (IPW), which represents the depth of water in mm per square meter that the column of atmosphere directly over the GPS receiver is holding in the vapor state, is given by [9],

$$IPW = \Pi ZWD \tag{2.11}$$

where,

$$\Pi = \frac{10^6}{461,525(\frac{373,900}{T_m} + 22.1)} \tag{2.12}$$

and

$$T_m = 70.2 + 0.72T_0 \tag{2.13}$$

with T_0 the surface temperature at the GPS receiver location.

2.2.2 GAMIT Software

There are a number of available software packages that one can use for IPW estimation. The one we use in this work is the GAMIT software package developed at MIT [26]. In addition to providing high-precision position analysis, GAMIT has routines for GPS-Met IPW estimation. The routines use the double differencing technique and tropospheric delay models described previously.

Inputs to GAMIT are RINEX (Receiver Independent Exchange Format [25]) navigation files containing the receiver and satellite clock offsets, observation files containing the code and carrier phase measurements for each slant path at 30 second sample rate, and for the GPS-Met application, meteorological files (from a collocated or nearby weather station) containing the surface pressure, temperature, and relative humidity data at the GPS site location. The satellite orbital parameters giving precise orbit information for each satellite are also an input. These are generated

by the IGS (International GNSS services) [17] and are automatically downloaded by GAMIT for the analysis time period in order to correct for orbit errors. Reference stations with baselines of more than 500 km from the GPS stations at which IPW is desired are chosen to satisfy the double-differencing condition that the reference site is uncorrelated from the IPW sites [40]. GPS-Met software, such as GAMIT, has been validated to produce precipitable water vapor measurements of better than 2 mm RMS [18].

2.2.3 IPW Normalization

The GAMIT software produces point estimates of IPW. To obtain the spatial distribution of IPW, we need to combine IPW values from multiple GPS receivers. Before we can do so, however, we first need to normalize the IPW values from the different GPS stations. IPW is the amount of water vapor in a vertical column over the site. Because the amount of water vapor the atmosphere can hold is a function of temperature and temperature generally decreases with altitude, IPW will consequently also depend on the altitude of the station. Since IPW depends on temperature, IPW will also change with season so that an IPW value corresponding to low humidity at summer temperatures might be saturated at winter temperatures. Moreover, IPW by itself does not tell the level of saturation, since again that depends on the daily temperature. To account for station height differences, seasonal (monthly) variations, and to identify anomalously high or low IPW values, we normalize the IPW values by the monthly mean and standard deviation before we combine them as follows,

$$NIPW_{ij} = \frac{IPW - \mu_{ij}}{\sigma_{ij}} \; \forall \; i, j \tag{2.14}$$

Here the subscripts i and j are for the station and month respectively and μ_{ij} and σ_{ij} are the mean and standard deviations respectively. In the literature NIPW is

termed the standardized anomaly of precipitable water vapor and is a common way of presenting precipitable water information [23].

2.2.4 Multiquadric Interpolation

To obtain the spatial distribution NIPW we interpolate from a set of point measurements. The paper [52] describes a number of methods for interpolating geophysical data. The method we chose is the multiquadric method. This method was chosen because it has been shown to perform nearly as well as the more common Kringing technique [38] but without the need for historical data.

Similar to Kringing, the multiquadric method is a weighted linear interpolation method where the estimate h_0 for any grid point (x_0, y_0) is given by,

$$h_0 = \sum_{j=1}^{n} w_j \cdot h_j \tag{2.15}$$

where h_j is the observed NIPW at point (x_j, y_j) and w_j is the weight giving the influence of h_j in determining h_0.

In the multiquadric method the weights w_j are calculated from the matrix of distances between the observed points (x_j, y_j) and the distances between the interpolated point and each observed point as follows. We start by expressing the observed points as a weighted linear combination of the distances between the observation points,

$$h_j = \sum_{i=1}^{n} c_i \cdot d_{ji} \ \forall j = 1, 2..n \tag{2.16}$$

where d_{ij} is the distance between GPS site i and GPS site j. The coefficients c_i are then determined by,

$$c_i = \sum_{j=1}^{n} \delta_{ij} \cdot h_j \ \forall i = 1..n \tag{2.17}$$

where δ_{ij} is an element of the inverse of the $n \times n$ distance matrix $d_{ij}, j = 1..n \; and \; i = 1..n$. Given the c_i we can thus write,

$$\begin{aligned} h_0 &= \sum_{i=1}^{n} c_i \cdot d_{0i} \\ &= \sum_{i=1}^{n} [\sum_{j=1}^{n} \delta_{ij} \cdot h_j] \cdot d_{0i} \\ &= \sum_{j=1}^{n} [\sum_{i=1}^{n} \delta_{ij} \cdot d_{0i}] \cdot h_j \end{aligned}$$

or

$$w_j = \sum_{i=1}^{n} \delta_{ij} \cdot d_{0i} \tag{2.18}$$

2.2.5 IPW Field Generation

In light of the above discussion, we can summarize our method for obtaining fields of NIPW as follows. Every 30-minutes we do the following,

1. Obtain GPS navigation and observation files from N GPS sites distributed throughout the geographical region of interest;
2. Obtain Pressure(P), Temperature(T), Relative Humidity(RH) for each GPS station;
3. Put the P, T, RH data in the required RINEX format;
4. Feed the GPS and meterological RINEX files into GAMIT to get IPW values for each station for the current 30-minute interval;
5. Normalize the IPW values based on the average and standard deviation for the given station and given month;

6. Apply multiquadric interpolation to obtain the NIPW field for the current 30-minute interval.

2.3 Measuring Rainfall

The primary instrument for obtaining spatial-temporal fields of precipitation is the ground-based weather radar. There are a number of different weather radar systems in operation in the U.S. These include the 160 long-range WSR-88D, Next Generation Weather Radars (NEXRAD) operated by the U.S. National Weather Service (NWS) for real-time weather monitoring and short-term forecasting and the 45 Terminal Doppler Weather Radars (TDWR) operated by the NWS to provide high-update rate, high-resolution weather and wind data at key major airports. In addition to these, increasing numbers of television stations have their own weather radars, and there are a variety of research weather radars, such as the network of small X-band radars operated by the University of Massachusetts, Center for Collaborative Adaptive Sensing of the Atmosphere (CASA) in the Dallas-Fort-Worth metroplex region.

As a weather radar scans in azimuth, it sends out a narrow, pencil beam of microwave pulse energy and then samples the return echo. This partitions the space around the radar into resolution volumes or voxels (volume elements). Voxels have the shape of a disk on its side: the diameter of the disk determined by the radar's beamwidth; the thickness by the radar's gate spacing. The size V of a voxel in cubic meters thus varies with the square of the range from the radar according to,

$$V = G \cdot \pi \cdot tan^2(\theta/2) \cdot R^2 \tag{2.19}$$

where G is the gate spacing in meters, θ is the beam width in radians, and R is the range from the radar to the voxel in meters. For NEXRAD, with its 1 degree beam and 250 meter gate spacing, this leads to voxel with volumes that are roughly 6

million cubic meters at 10 km from the radar to 3 billion cubic meters at the radar's maximum (Doppler) range of 230 km.

Under assumptions that the voxels are completely filled with liquid hydrometeors (raindrops) and that the hydrometeors are small relative to the radar's wavelength (Rayleigh scattering), weather radars measure the echo from each voxel to infer properties of the hydrometers in the voxel (their density, size, radial (Doppler) motion, (Polarimetric) shape asymmetry, and so on) [13] [16]. To sample the complete volume around the radar, multiple 360 degree sweeps in azimuth are performed, each at a different elevation tilt angle. Each such 360 degree sweep is termed a Plan Position Indication (PPI) scan, and the sequence PPIs starting at the radar's lowest elevation tilt angle and working up to the radar's highest tilt angle that are performed to cover the volume is called the Volume Coverage Pattern (VCP). The NEXRAD lowest elevation tilt angle is 0.5 degrees, and the standard convective weather VCP has a temporal update rate of approximately 5-minutes between revisits of the 0.5 degree lowest elevation PPI.

The main weather radar product, reflectivity (Z in mm^6/m^3), is defined as the 6-th moment of the drop size distribution [16].

$$Z = \int_0^\infty N(D) \cdot D^6 dD \tag{2.20}$$

where D is the (volume equivalent spherical) drop size (diameter in mm) and $N(D)$ is the drop size distribution given by (Marshall-Palmer model) [16].

$$N(D) = 8000 \exp(-4.1 R^{-0.21} \Lambda D) \tag{2.21}$$

where R is the rain rate (in mm/hr).

Assuming the drop size distribution in (2.21), (2.20) can be solved to obtain Z in terms of the median drop size D_0 [16],

$$Z = 642D_0^7 \tag{2.22}$$

or in terms of rain rate,

$$Z = 297R^{1.47} \tag{2.23}$$

Given the above relationships, Table 2.1 relates reflectivity (in dBZ $= 10log_{10}Z$) to median drop size D_0 (in mm) and rain rate R (in mm/hr)[6]. Noting that cloud droplets

Table 2.1: Reflectivity as a function of median drop size diameter and rainfall rate.

dBZ	D_0 (mm)	R (mm/hr)
-42	0.1	-
0	0.4	-
10	0.6	0.1
20	0.8	0.5
30	1.1	2.3
40	1.5	11.0
50	2.1	52.0

average $12\mu m = 0.012mm$, and that even the high-powered NEXRAD only has -21dBZ sensitivity at 10km (estimated from the WSR-88D specification of -7.5 dBZ at 50 km[7] and the fact that minimum detectable reflectivity is proportional to range squared), we see that weather radars generally cannot detect clouds or uncondensed water vapor, but can only detect active precipitation. We also remark that although dual-polarimetric weather radar, such as the upgraded NEXRAD, do provide a rainfall rate product[8], in this thesis we will use reflectivity as a proxy for rainfall rate with 30dBZ and above indicating rainfall and the dBZ value reflecting rainfall intensity.

[6]We remark that the Z-R relationship in equation 2.23 is only one of many such relationships. Others include the famous 1948 $Z = 200R^{1.6}$ Marshall-Palmer relationship [33] and the $Z = 300R^{1.4}$ relationship used by the NWS as the default Z-R relationship for the WSR-88D radar network http://www.srh.noaa.gov/tlh/?n=research-zrpaper

[7]https://www.roc.noaa.gov/WSR88D/PublicDocs/NTR96.pdf

[8]https://www.ncdc.noaa.gov/data-access/radar-data/nexrad-products

Specifically, we will use the NEXRAD 0.5 degree elevation reflectivity PPI as an indicator of precipitation activity and intensity at ground level.[9]

[9]To say that the 0.5 degree NEXRAD reflectivity product gives the rainfall at ground level ignores the fact that the earth is curved while radar beams travel in essentially straight lines leading to an increase in radar beam height above ground level with distance from the radar. In particular, we are ignoring the fact that even at its lowest tilt of 0.5 degrees, the bottom of a NEXRAD beam is some 3000 meters (10,000 feet) above ground level at the radar's maximum (Doppler) range of 230 km.

CHAPTER 3

DATA SET

The data set we will use for our nowcasting experiments comes from the Dallas-Fort-Worth (DFW) region. Part of the infamous U.S. "tornado alley", DFW spring and summer weather is dominated by convective thunderstorms that move in lines generally from west to east through the region. We choose the DFW region because we understand its climatology (through CASA's more than 15 years of operating networks of weather radars in tornado alley, first in Oklahoma and now in DFW) and because the DFW region has a high density of GPS receivers and weather stations whose data are publicly available on-line for our use.

We analyze 2 years of storm data (2014 - 2015) during the spring summer storm season in the DFW region which consists of days between May to August. We take as the center of our region the NWS KFWS NEXRAD radar in Fort-Worth Texas.[1] Within the 230 km coverage range of the radar we identified 44 Regional Reference Points, i.e., high performance dual-frequency GPS receivers. These GPS receivers, which are operated by the Texas Dept. of Transportation (TxDOT)[2], were deployed to provide precise position information for Geodetic studies. As such these GPS receivers do not have collocated weather stations. For the weather data (surface temperature, pressure, and relative humidity) required for IPW estimation, we used data from the network of Automated Surface Observation Stations (ASOS) operated by

[1] http://radar.weather.gov/radar.php?rid=fws

[2] http://www.txdot.gov/inside-txdot/division/information-technology/gps.html

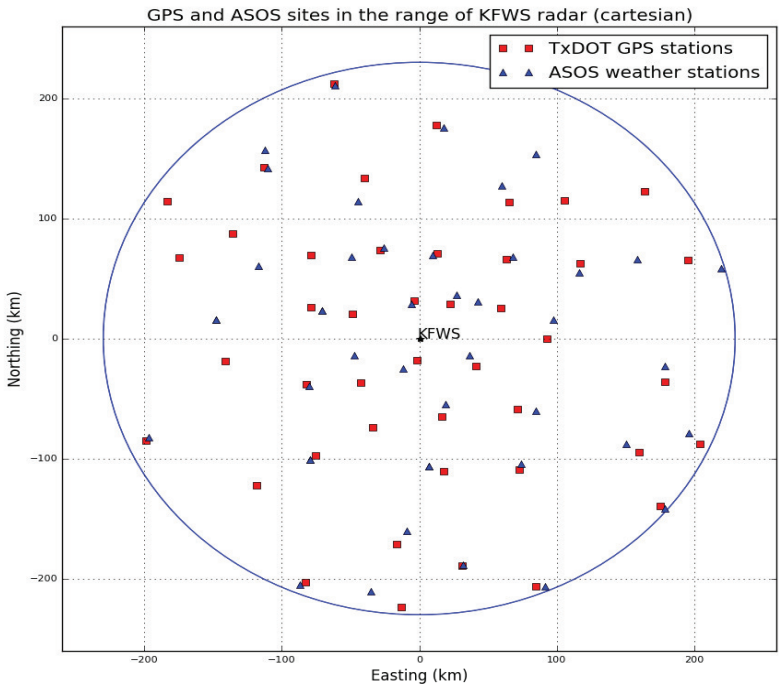

Figure 3.1: GPS and ASOS stations within the KFWS 230 km coverage range.

NOAA NWS[3]. The ASOS network of weather stations provide surface meteorological variables at 5 minute resolutions. Figure 3.1 shows the relative locations of the GPS receivers and ASOS stations within the 230 km coverage range of the KFWS radar. Table A.1 in the Appendix gives the locations and heights of the the GPS receivers and ASOS weather stations. GPS RINEX files of code range and carrier phase for the storm periods of 2014 - 2016 were downloaded for each of the 44 stations from databases maintained by SOPAC[4] [11] and CORS[5][49]. The meteorological data from the ASOS stations were obtained at 30 minute intervals for the year of 2014

[3]http://www.nws.noaa.gov/asos/

[4]http://sopac.ucsd.edu/

[5]http://www.ngs.noaa.gov/

from Teresa Vanhove of UCAR[6]. The KFWS 0.5 degree reflectivity product for the entirety of 2014 was downloaded from NCDC[7]. For the long baseline stations needed for double differencing, we chose the four stations: AC20 in Girdwood Alaska, CONZ in Concepcion Chile, P019 in Fairfield, Idaho, and UNBJ at the University of New Brunswick, Canada. The closest of these stations, P019 is approximately 1500 km from the KFWS radar defining the center of the DFW GPS network. These stations were chosen to ensure that there was always at least one satellite that had a view of all of the TxDOT GPS stations and one of the baseline stations.

To get the meteorological data for a particular TxDOT GPS station, we found the closest ASOS site and used the equations from [5] to interpolate the surface temperature (T), pressure (P), and relative humidity (RH) data from the MSL height of ASOS site to the MSL height of the GPS station,

$$P_{SL} = P_{MSL} \cdot (1 - 2.26 \cdot 10^{-5} \cdot H)^{5.225} \tag{3.1}$$

$$T_{SL} = T_{MSL} - 0.0065 \cdot H \tag{3.2}$$

$$RH_{SL} = \frac{RH_{MSL}}{e^{-0.0006396 \cdot H}} \tag{3.3}$$

In the above, P_{SL} is the pressure at the station level, P_{MSL} is the pressure at mean sea level, and H is the height of the station (in meters MSL). We then wrote a Python script to mine the large database of met values from the various ASOS sites to generate the met RINEX files for each GPS station in 30-minute intervals.

The GPS and met RINEX files for each station were run through GAMIT to produce 30-minute estimates of IPW for each station for the duration of our entire data set (spring summer storm season 2014-2016). The mean and standard deviation of IPW for each GPS station were calculated for each month, and the IPW values

[6]http://www.suominet.ucar.edu/

[7]https://www.ncdc.noaa.gov/data-access/radar-data

normalized to obtain the standardized anomaly of precipitable water, which we denote NIPW. The NIPW values were then mapped to a 300 km by 300 km, 3km resolution grid centered on the KFWS radar using multiquadric interpolation method.

The data processing was done on a Ubuntu server where GAMIT is installed. The server has 16 CPUs with 4 cores each. The data processing was performed in a manner that optimized the computation resources of the server thus allowing many days to be processed at the same time. Specifically, the 44 GPS stations were broken down into 4 different sub networks where the data processing for the networks could run in parallel.

Regarding the NEXRAD reflectivity data, this was first down-sampled from its native 5-minute update rate to the 30-minute NIPW update interval. We then performed a polar to Cartesian coordinate conversion to map the reflectivity field to the same 3 km grid resolution as used for NIPW. The result is two fields of 10,000 pixels each updated every 30-minutes, one of NIPW (a measure of atmospheric water vapor) the other of reflectivity (a measure of rainfall).

3.1 Experimental Data Set

Machine learning involves first training the machine learning system on a training data set and then testing it on a separate validation data set. To determine our training and validation set we selected days within the storm periods which satisfied a certain criteria and these days were termed as "weather anomaly" days. The criteria for selecting these dates were to check to see if 30 or more IPW values for a particular day showed a standard deviation of greater than 2 from the mean amongst any of the stations. The list of days was further padded by taking into consideration one day before and one day after the weather anomaly days to account for storm build up and decay. Table 3.1 lists the weather anomaly days for the years 2014 and 2015.

Table 3.1: Weather anomaly days

Month	2014	2015
May	8,12,23,25,31	9,10,19,20,23,24,29
June	13,18,19,22,25	15,16,17,18,21
July	15,16,17,18,24,28,29,30,31	3,7,8,22,31
August	11,16,17,18,19,29	1,19,20,21,25

Interestingly enough almost of of the weather anomaly dates had a severe storm in them.

The training and validation for the dataset was performed in a K-fold cross validation manner where each month was considered a fold. In other words one month from the dataset was left out for validation and the other months were considered as the training data set.

3.2 Preliminary Data Analysis

To give a sense of typical IPW values in the DFW region and how these IPW values vary with height, Figure 3.2 plots the monthly means for the 4 highest and 4 lowest GPS sites for the year of 2014. Clear in the plot is the height dependence, with mean IPW inversely proportional to station height.

To see how IPW varies with season in the DFW region, Figure 3.3 plots histograms of the IPW values observed at the lowest (TXBX) and highest (TXC3) GPS sites. From the figure we can see the seasonal dependence with higher mean IPW values during the warmer months (April-September) and lower mean IPW values during the cooler months (October-March). This is due to the fact that during the summer of the warmer months the atmosphere can hold more water vapor than the colder months. Again we see the height dependence with the IPW of the highest station having generally lower IPW values than the lowest station.

These first two plots illustrate why IPW values are typically normalized before they are combined to produce a field. Our approach of normalizing to obtain the number of standard deviations from the mean removes height and seasonal differences and also allows us to infer something about saturation level as many standard deviations above (below) the mean will typically indicate a saturated (dry) atmosphere respectively without a need to estimate relative humidities or dew points.

3.3 Preprocessing

Once we compute the IPW values we find several unusual drops and peaks in the IPW values which ideally does not make any sense. For example we a case as shown in the plot we see the IPW values falls from 60mm to 20mm which is not possible because energy can not be transferred at such a rate. We suspect that the cause of this is due to the lack of GPS signal for that particular time step where the averaging occurs. Instead of averaging over 8-12 different satellites we only fond 3-5 satellites to average from which causes the unusual drop.

To address this issue before we compute the monthly means and standard deviations, we look at all places where there is a decrease of greater than 50% or increase greater than 100% between consecutive time steps. We set these values to NaN and linearly interpolate all of the NaN values.

3.4 Atmospheric Rivers

From generating the the NIPW fields and the reflectivity fields we generally notice that the NIPW field creates a river like pattern in the domain. Further from overlapping the reflectivity field over the NIPW field we find that the the storm is flowing with the river or is being lea by the river. We also notice that the storm is slightly on the bank of the river. We will analyze several cases from our weather anomaly days in this section.

First let us plot the histogram of IPW values for certain stations for each month in the spring summer storm season. Figure ... shows the histogram for each month taken over both the years. Finally, to see visually if there are any spatial and temporal correlations between NIPW and reflectivity that might allow a machine learning algorithm to predict rain, Figure 3.4 shows the reflectivity fields superimposed atop the NIPW fields for three different storm cases from our experimental data set. These storm cases include May 8th (left hand column), July 17th (middle column), and August 29th (right hand column). Going from the top to the bottom of each column, the fields show the joint evolution of reflectivity and NIPW in 30 minute time intervals.

The May 8th and August 29th cases show lines of thunderstorms moving through the DFW region. In both these cases, we can clearly see that the reflectivity seems to ride just slightly behind the moving ridge of high NIPW. For the July 17th case, which is more of a slow moving stratiform event, the reflectivity surrounds a very slow moving peak of NIPW. Thus, whereas it might be difficult to predict the rainfall at a given location from reflectivity alone, noting that rainfall tends to track just behind large peaks in NIPW can be expected to aid rainfall nowcast ability. We explore this conjecture in the next chapter.

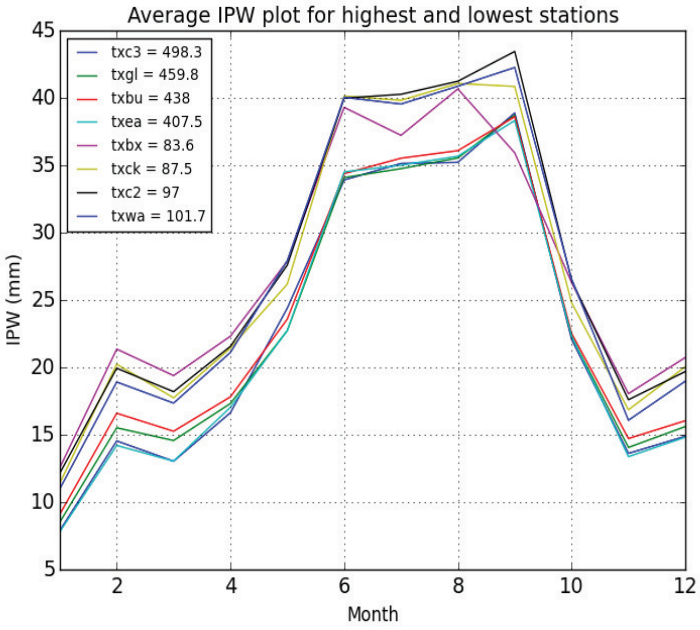

Figure 3.2: Mean IPW for each month for 4 highest and 4 lowest stations. The height is measured as the Geodic height in meters.

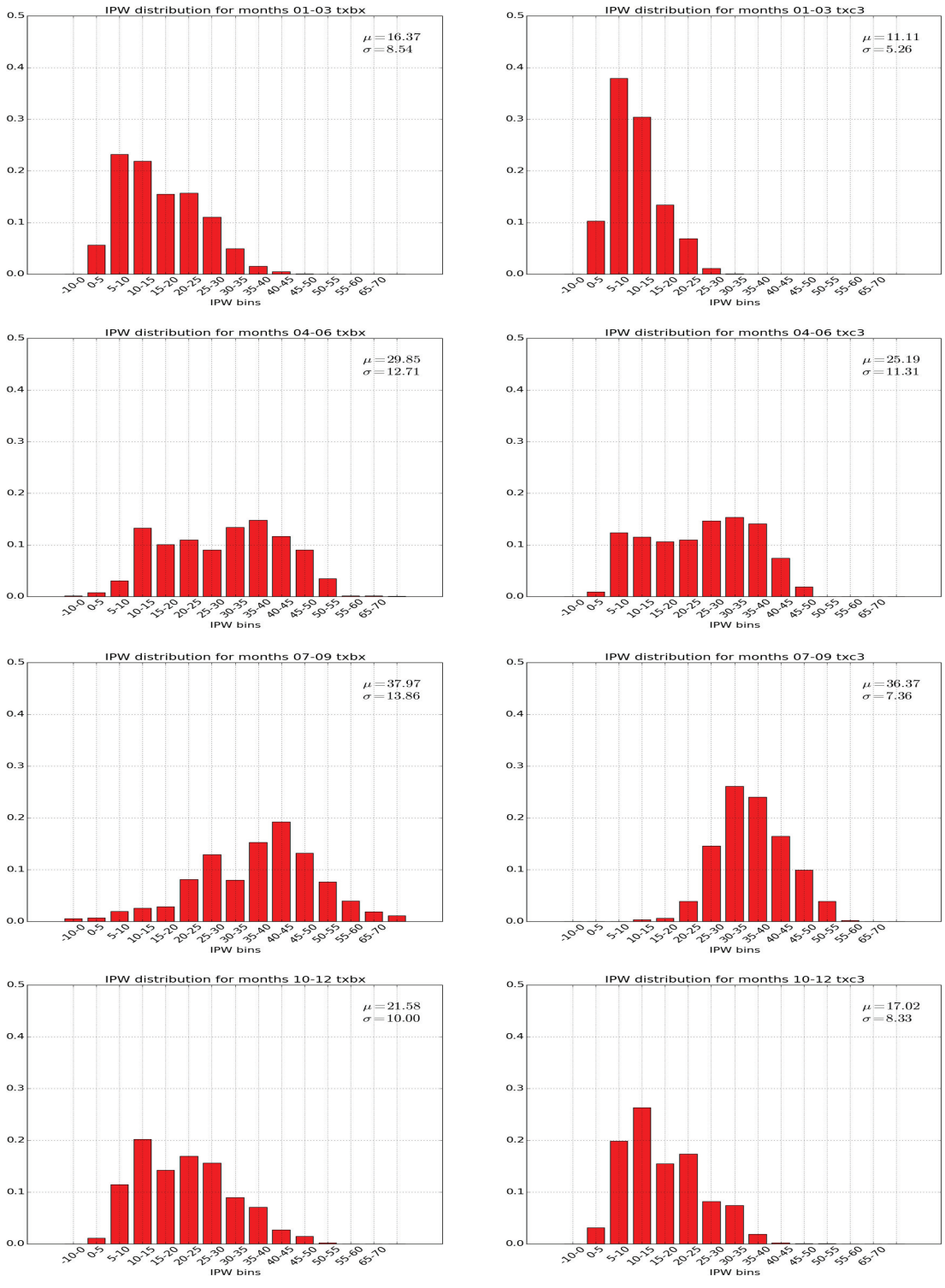

Figure 3.3: IPW histograms for each season for the left: lowest station TXBX and right: highest station TXC3 (right).

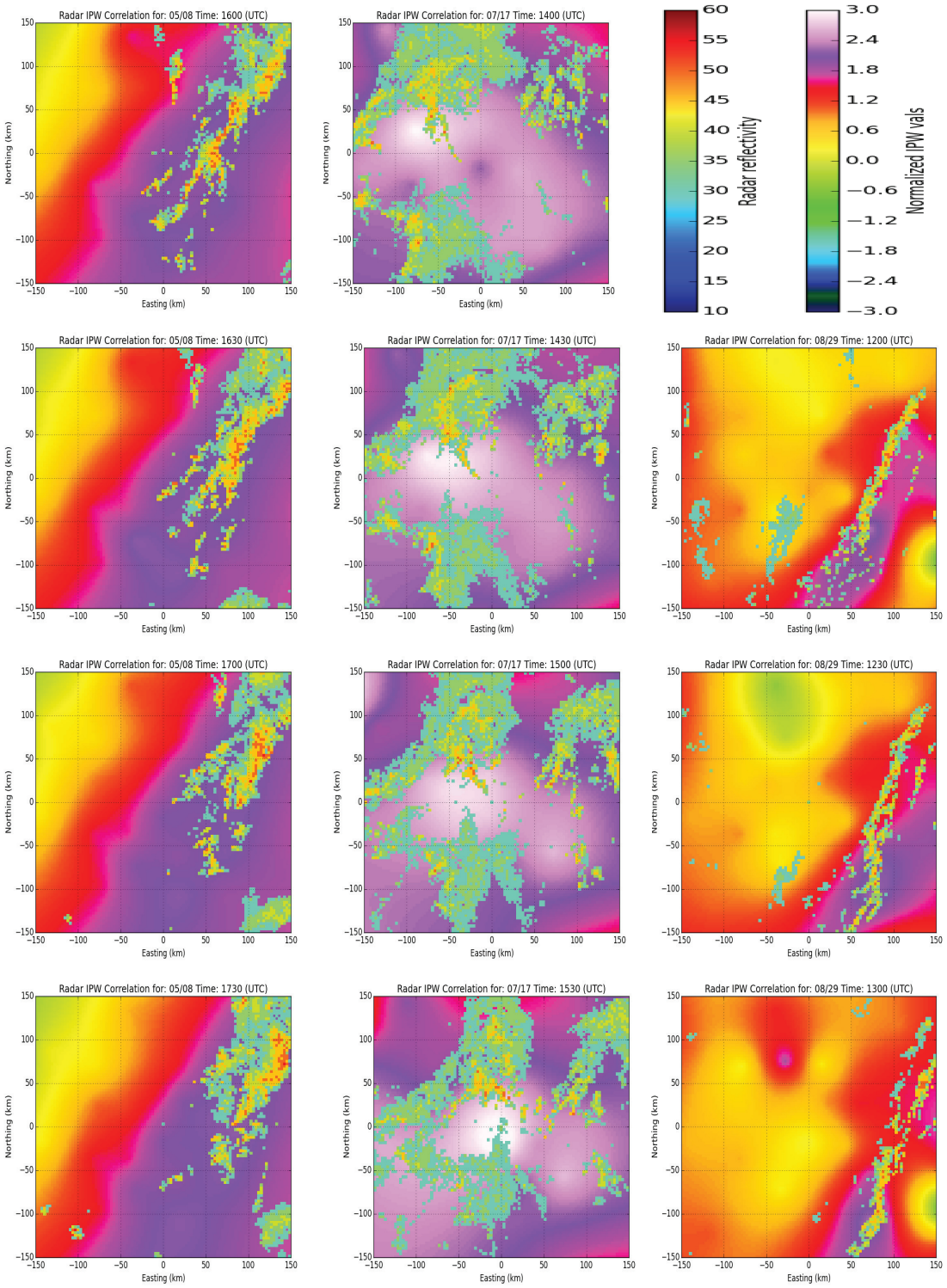

Figure 3.4: Reflectivity fields overlayd onto IPW fields for three storms left: May 8th, center: July 17th and right: August 29th).

CHAPTER 4

THEORY OF THE MACHINE LEARNING ALGORITHMS USED IN THIS THESIS

The goal of this Thesis is to determine if Integrated Presipitable Waver Vapor (IPW) measured from a network of GPS stations provides any additional value to nowcasting systems/applications. As such it is very difficult for us to model the underline generative process $P(X, Y)$ where X is the IPW field and Y is the reflectivity field. Thus a data driven approach is taken in order for us to learn from a training data L a function which maps spatial and temporal evolutions of precipitable water and precipitation fields and In order for us to analyze this we use machine learning techniques to determine the additional value.

This chapter gives a background information on the machine learning algorithms that we will use to model and predict the future precipitation fields. A description of why these algorithms will become clear in the next chapter where we define the general formalisms for this problem.

We thus focus on discriminative methods which model the probability $P(Y|X = x)$. Specifically we train machine learning algorithms explained in this chapter to our data set to find out if nowcasting performance is improved if the moisture information is provided along with the the time series of the spatial reflectivity fields. We focus on three main machine learning algorithms which are well suited for this problem. The Random forest, Naive Bayes and the convolutional neural network with dropout. For our initial machine learning results we chose to apply machine learning techniques similar to those used in [35]. The problem they looked at was similar to ours in

that they were trying to predict convective initiation 0-60 minutes in advance using GOES-R satellite data and NWP data. For their predictions they used two different machine learning classifiers, logistic regression and random forests. The accuracy of predicting convective initiation one hour in advance was reported as 84 % and 71% for logistic regression and random forest respectively.

4.1 Introduction

Supervised machine learning ideally involves a training data set where each instance or example has a label which can either take a finite value from a set $c_1, c_2, c_3...c_n$ or can take a real value where $y \in \mathbb{R}$. Let us define an input vector $\mathbf{x}$ to have values $x_1, x_2, x_3...x_d$ where $x_i \in X_i$ (i = 1..d) where d is the number of dimensions of the vector. If we thus have N different vectors of $\mathbf{x}$ we can thus represent the inputs in the form of a matrix $\mathbf{X} \in \mathbb{R}^{N \times d}$. Our training data set contains N labeled instances $(\mathbf{x}, y)$. From this data set we will seek to find a model $f : \mathbf{x} \rightarrow y$. The can exists many such models f within the set of hypothesis $f \in H$. In order to find the best model $f*$, we seek to minimuze some form of convex error function.

4.2 Random Forest

The random forest algorithm first introduced by [12] is a bagging technique where many different models with high bias but low variance are fit to patches of the data set which are averaged in the case of a regression task or a majority vote is taken in the classification task. Ideally the many different models which are fit to the data are decision trees or CART (Classification and Regression Trees). The random forest algorithm is an ensemble technique which ideally consists of an ensemble of CART (Classification and Regression Trees) where the decision on deciding the response variable is made via majority vote or averaging of an ensemble of decorrelated decision trees in the forest [12]. Trees are ideal candidates for random forests as they can

capture complex interactions in the data. Bagging or bootstrap aggregation is a method by which many different models are fit to the data and a majority vote or average is computed from all of the models put together. The main idea behind bagging is to improve on the bias variance trade off. We can assume that each model will have a high bias or under fit the data specific to it but have a low variance in general. In order to reduce the variance the dependent variable is decided by taking the majority vote or average of all of the models put together. In the case of regression task the output of the model is given by

$$f_{RF}(x) = \frac{1}{B}\sum_{i=1}^{B} T_b(x) \tag{4.1}$$

where $f_{RF}(x)$ is the predicted value of the random forest and in the case of a classification task the majority vote of all the models are taken given by,

$$f_{RF}(x) = argmax\ T_b(x) \tag{4.2}$$

4.2.1 Classification and Regression Trees

Decision trees are the most ideal single predictors for a random forest and thus the forest has many trees and hence its name. Given a data set of $(\mathbf{x}, y)$ pairs a simple decision tree works on the principle of finding a best split amongst all variables $d \in D$ given a set of training vector/output pairs $(\mathbf{x}, y)$ where $\mathbf{x}$ is a vector with dimension D. Classification is based on finding a threshold t such that a variable in a data case x_d is split based in $x_d < t\ or\ x_d > tor\ x_d = t$. The data case is assigned to the left or right branch of the tree based on x_d and threshold t. The data cases are traversed through the tree from the root node to the leaf of the tree and the class is determined at the leaf.

As a decision tree learns, it finds both the optimal variable d to split on and the optimal threshold value t using a greedy heuristic that tries to maximize the

training accuracy. In a geometrical sense the decision tree breaks the D dimensional feature space into smaller regions such that each region contains the maximum number of instances that belong to a single class. Hence a fundamental metric p_{km} is the proportion of observations in the m-th region that are in the k-th class. This leads to the so-called "Genie Index" criterion,

$$C_{GI} = \sum_{k=1}^{K} p_{km}(1 - p_{km}) \tag{4.3}$$

which decision tree learning seeks to maximize by adjusting the tree parameters d and t. Hyper-parameters to the algorithm include the number of trees in the forest B and the maximum tree depth. The number of trees (200) was selected by starting with a small number of trees and increasing the number until the f1 score obtained on the cross-validation sets stopped improving. For maximum tree depth we used the scikit learn default in which the tree will grow until all its leaves are pure[36].

The random forest algorithm thus can be summarized by the following algorithm from [21],

1. Initialize the number of trees B in the forest.
2. for b = 1:B
   (a) Draw a bootstrap sample Z of size N examples from the training data
   (b) Grow a random forest tree T_b to the bootstraped data by recursively repeating the following steps
      i. select m at random from D (ideally $m = \sqrt{|D|}$)
      ii. pick the best variable split among the m variables based on the Genie criteria
      iii. split the node to two daughter nodes
   (c) The classification is based on the majority votes from B trees.

The random forest thus forms its "decorrelated" trees by picking a set of random variables m at each iteration of each tree. In this fashion an ensemble of weak learners are built.

One of the key advantages of the Random Forest classifier is its interpretability through variable importance. The variable importance is an attribute of the random forest classifier which measures the relative importance of each variable. This is done by measuring the prediction performance of the "OOB (Out Of Box Samples)". When the b^{th} tree is being constructed, examples apart from its bootstrapped samples are used to measure its predictive performance at a particular split. Different permutations of the variable split are tried and the performance is measured on the OOB samples. The results are accumulated over all trees and the variables which are most important are ranked.

4.3 Convolutional Neural Network

Deep learning has shown promises in many different tasks such as image classification and speech recognition. Fundamentally it seeks to learn multiple layers of representation or extract features from the raw input which it finds useful. Deep neural networks are trained on a convex loss function depending on the task and the training is done using back propagation [citation]. These approaches have slowly started to surface in the meteorological domain where atmospheric variables are involved. The interesting thing about deep learning approaches is that the problem is mostly solver as long as we have an end to end trainable network. What sets CNN apart from conventional neural networks is the convolutional layer.

CHAPTER 5

PROBLEM FORMULATION

5.1 Problem Setup

The classifiers we choose for our initial machine learning experiments were the Naive Bayes classifier and the Random Forest classifier. Our goal is to make a binary 0, 1 (no rain, rain) prediction 1 hour in the future for a random subset of the 10,000 pixels that make up the 300 square km grid around the KFWS radar, where rain is taken as a pixel reflectivity value exceeding 24 dBZ.

With $y \in [0, 1]$ as the dependent binary variable indicating "rain" and "no rain" at a pixel point, the feature set for determining the dependent variable y is chosen to be the most recent four frames (2 hours) of NIPW and reflectivity fields. Specifically, the feature set for predicting rain at a given pixel was determined from the four most recent fields of NIPW and reflectivity in the 33 by 33 pixel subgrid centered on the pixel. As a simplification, we averaged the NIPW and reflectivities in each 33 by 33 grid. Our initial learning problem is thus one of making a binary rain, no-rain prediction of the state of the center pixel 1-hour in the future from a feature set consisting of the averages of the NIPW values and reflectivities in the most recent 4 frames (past 2 hours) in the 33 by 33 (100 by 100 km) region surrounding the pixel. As shown in 5.1, we do this for a total of 1500 pixel locations uniformly distributed throughout the 10,000 pixel DFW domain.

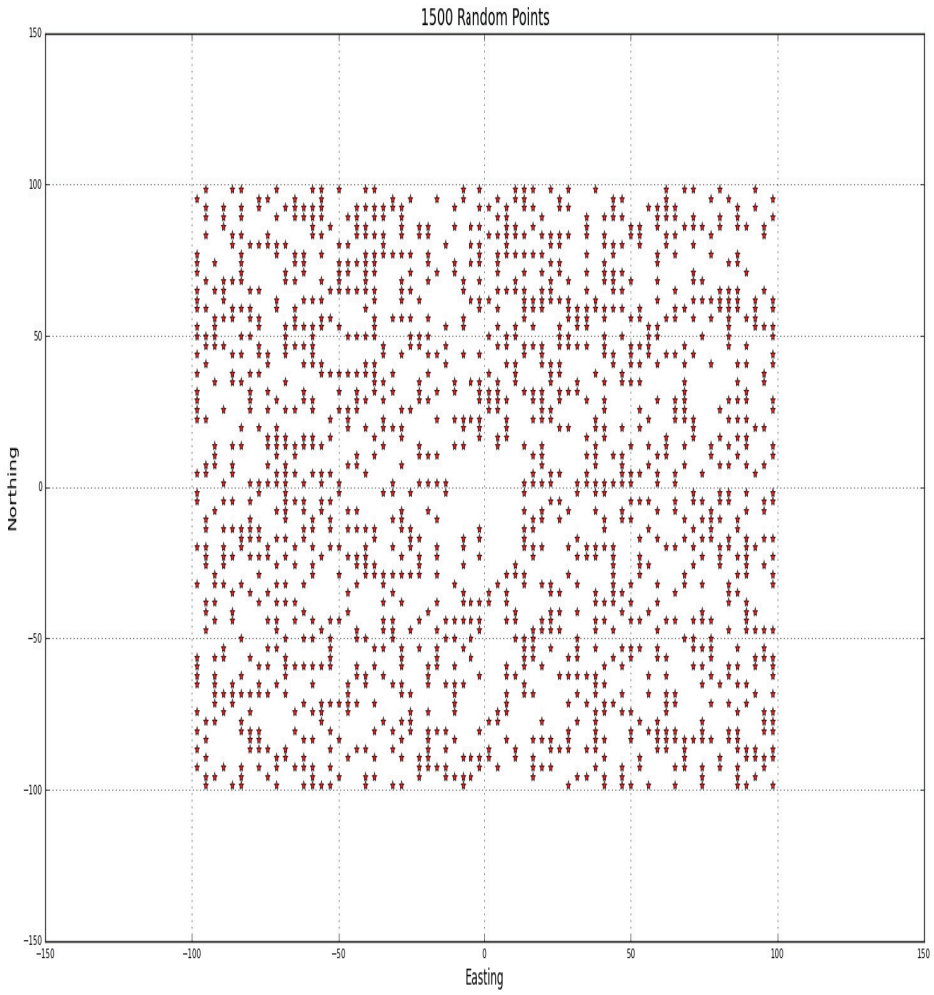

Figure 5.1: 1500 Uniformly sampled pixels which can have a 33 x 33 grid around it. The center points surrounding the radar were left out as this area is prone to clutter and noise

5.2 Training and Validation

The experimental set (see Section 3) was broken down into a series of training set and validation sets using the K-Fold cross validation technique with $k = 7$ [21]. The algorithm for cross validation technique is explained as follows:

1. Randomly shuffle the experimental data set.

2. Divide the experimental data set into K equal parts.

3. Train on K-1 of the blocks and test on the remaining block.

4. Repeat until all K blocks have been tested on using the other K-1 blocks as training examples.

5.3 The Naive Bayes Classifier

The Naive Bayes or Bayes optimal classifier [57] is based on the assumption that each variable behaves independently with respect to the other variables and disregards any correlation one variable may have on any other. Given a training example $x_i \in \mathbb{R}^D$ and its corresponding output $Y_i = c$ where $c \in G$ where G is the set of all possible classes, we can compute the prior probability $P(Y = c) = \pi_c$ and the true probability $\phi_c(x) = p(X = x|Y = c)$. We can then apply Bayes rule to calculate the posterior probabilities,

$$P(Y = c|X = x) = \frac{\phi_c(x)\pi_c}{\sum_{c' \in Y} \phi_{c'}(x)\pi_{c'}} \tag{5.1}$$

making the "Naive" assumption that all variables are independent we can say that,

$$\phi_c(x) = p(X = x|Y = c) = \prod_{d=1}^{D} p(X_d = x_d|Y = c) = \prod_{d=1}^{D} \phi_{cd}(x_d) \tag{5.2}$$

Thus we can write the general classification function as,

$$f_{NB}(x) = \underset{c \in Y}{\operatorname{argmax}}\, \pi_c \prod_{d=1}^{D} \phi_{cd}(x_d) \tag{5.3}$$

The Naive Bayes classifier despite making its fundamental assumption of independent variables, which is never true for real-world problems, is capable of forming complex decision boundaries.

5.4 Nowcasting Performance Results

Given the mechanisms of precipitation discussed in Chapter 2 and our initial analysis in Chapter 3, we believe that the combination of NIPW and reflectivity should be more predictive of precipitation than either NIPW or reflectivity alone. To test this hypothesis, three kinds of predictions were made using the two classifiers on each of the 7 training and validation set - predictions using only NIPW fields, predictions using only reflectivity fields, and predictions using both NIPW and reflectivity fields.

The output of the Bayes and Random Forest classifiers are real numbers between 0 and 1. These can be interpreted as beliefs that it is raining or not raining at a selected pixel location. The actual rain, no-rain decision is determined by a decision threshold - rain if the output exceeds the decision threshold, no-rain otherwise.

To evaluate the performance of the two classifiers we chose a set of metrics which are tailored to our problem where we only have 5% of the experimental set with rain cases. So a measure on how well we predict rain is needed. The first metric is the precision-recall curve [37]. A precision-recall curve starts with the determination of the number of Hits (H), Misses (M), False alarms (F) and Correct negatives (C). From these the precision (P) and recall (R) are defined as,

$$P = \frac{H}{H + F} \tag{5.4}$$

$$R = \frac{H}{H + M} \tag{5.5}$$

Precision, which is the proportion of cases that are predicted positive that are actually positive, is a measure of confidence. Recall, which is the proportion of real positive cases that are predicted positive, is a measure of sensitivity. A precipitation nowcast system with high recall but low precision may indicate that that the predictor is returning a lot of cases of rain but that only few are actually correct. A system with low recall but high precision indicates that the predictor is returning very few positive

results but most of these results are accurate. Ideally a good balance between the two scores is desirable. A measure of this balance is the f_1 score, defined as,

$$f_1 = 2 \cdot \frac{P \cdot R}{P + R} \tag{5.6}$$

A high value of f_1 would indicate a good balance between precision and recall and would thereby indicate that the classifier is making accurate predictions.

Figure 5.2 shows the precision-recall (P-R) curves for the Bayes and Random Forest classifiers. These are plots of precision vs. recall as a function of the decision threshold. For a P-R curve, best performance is when the curve bends towards the point (1,1). The area under the P-R curve is called the average precision score. Similar to the area under the Receiver Operating Characteristic ROC curve, a large area under the curve, or large average precision score signifies a good classifier. From Figure 5.2 we see that for both classifiers best performance is obtained when both IPW and reflectivity features are used to nowcast rainfall 1 hour into the future. We also see the intuitively appealing result that precipitation nowcasting using reflectivity performs better than precipitation nowcasting using NIPW. This is intuitively appealing since reflectivity is a more direct and immediate measure of precipitation than NIPW.

The poor performance of the Naive-Bayes classifier is due to the assumption that all of the variables are independent of each other, something that is not true with weather fields which change in smooth continuous ways and for NIPW and reflectivity fields where high NIPW is often followed by high reflectivity (recall Chapter 3). The random forest classifier, which does not make such an assumption performs much better.

The poor performance of the Bayes nowcaster and good performance of the Random Forest nowcaster are further verified by Figure 5.3 which plots the f_1 scores as a function of the 7 cross-validation training/validation sets. Here the decision threshold was set of 0.5. The random forest classifier has an average f_1 score of 0.68 and does

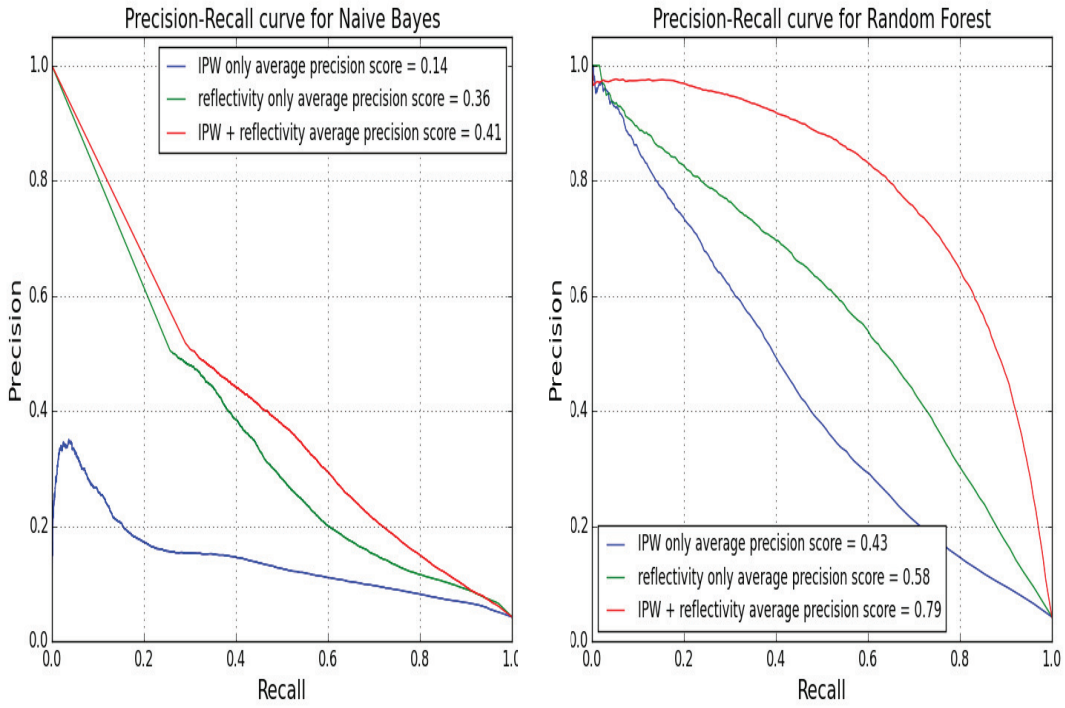

Figure 5.2: Average precision score for (a) Gaussian Naive Bayes Classifier (b) Random Forest Classifier.

not vary much between validation sets. This suggests that all of the misclassified examples are of a very specific type and they are also equally distributed between the validation sets. The Bayes classifier performs poorly ($f_1 < 0.4$) in all cases.

Table 5.1 compares the two classifiers in terms of Probability of Detection (POD), False Alarm Rate (FAR), and Critical Success Index (CSI) for a decision threshold of 0.5. The formula for these three metrics can be found in Table 5.1. As seen earlier it is clear that the probability of detection is highest when both IPW and reflectivity variables are used for nowcasting. It is interesting to note that the POD for the NB is higher than the RF classifier for all three cases but this comes at a cost of much higher FAR. We can thus conclude that the RF classifier is better able to use the spatiotemporal evolution of the IPW and the reflectivity fields to nowcast rainfall 1 hour ahead.

Table 5.1: Performance metrics on a cross-validation set using the two classifiers.

Metrics	NB Classifier			RF Classifier			Formula
	IPW	Refl.	IPW + Refl.	IPW	Refl	IPW + Refl.	
POD	0.45	0.54	0.61	0.17	0.38	0.56	$H/(H+M)$
FAR	0.86	0.75	0.72	0.23	0.29	0.15	$F/(F+H)$
CSI	0.12	0.21	0.24	0.16	0.33	0.51	$H/(H+M+F)$

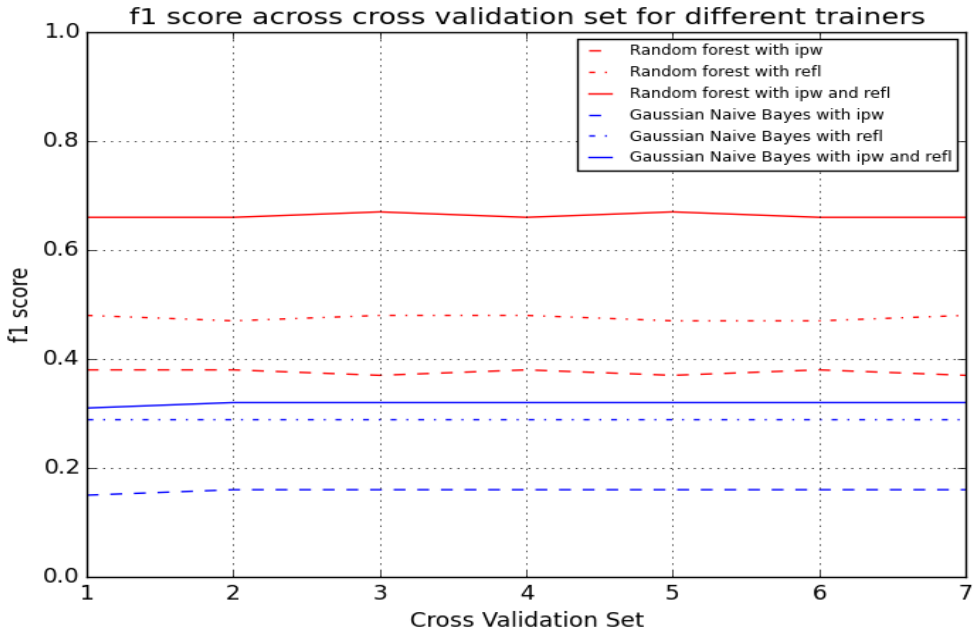

Figure 5.3: f_1 score for each train validation set.

The Random Forest Classifier seems to show promising results with an average precision score of 0.73. Mentioned earlier was the interpretability of the results learned by a Random Forest classifier through its ability to rank the importance of the different input variables in its decision making. Figure 5.4 shows the relative variable importance for the classification done given the input of average NIPW and reflectivity fields. As seen by the figure the the top three features for identifying the rain case are the reflectivities at $t-1, t-1.5$ and $t-2$ hours prior to the prediction time t (the prediction for time t is made at time t-1). This is followed by the 4 NIPW

averages in the order $t - 1.5$, $t - 1$, $t - 2$, $t - 2.5$. It is interesting to note that the IPW features play a significant role in the improvement of the classifier as shown by the P-R curves and f_1 scores. Again we see the intuitive result that reflectivity is more predictive of precipitation in the short-term than NIPW.

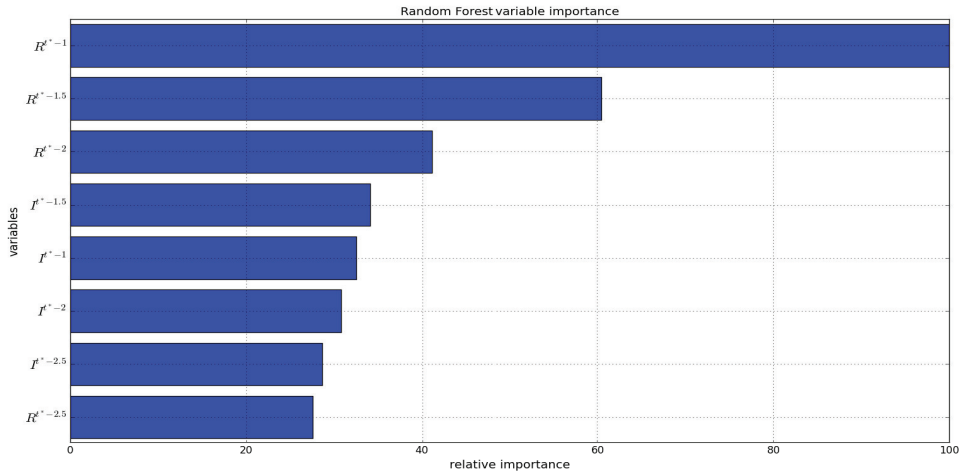

Figure 5.4: Relative variable importance measured by random forest classifier.

CHAPTER 6

DISCUSSION, PROPOSED WORK, THESIS TIMELINE

Out of an interest in data science and machine learning, this thesis tackles the challenging problem of nowcasting precipitation from the spatiotemporal evolution of atmospheric water vapor fields (measured by normalized IPW) and rainfall fields (measured by weather radar reflectivity). Following work by [12] [35], our initial results applied the Naive Bayes and Random Forest classifiers to the problem with mixed results - the Bayes classier performed poorly, the Random Forest classifier reasonably well.

6.1 Proposed Work

To complete this thesis, we propose to look into other machine learning techniques for finding spatiotemporal patterns in "video" sequences of NIPW and reflectivity.

On reviewing literature for machine learning algorithms that incorporate video sequences as inputs, the paper [44] bears similarity to our problem. In that work the problem addressed was that of recognizing human emotions from two separate video sequences, one focused on the face and the other recording body gestures. In their solution, the authors incorporated a technique from multi-view learning called Canonical Correlation Analysis (CCA). CCA is a method of finding a representation through a basis vector that seeks to maximize the correlation between 2 sets of features [51]. In [44] the facial expressions and hand movements were fused using CCA to obtain a set of features which performed better at emotion recognition than when the two features were concatenated. The CCA technique has also found success in other

problems involving time series data. The paper [24], for example, examined a stock market data set consisting of historic stock values along with more general economic factors. These two features being correlated and complementary were fused together using the CCA technique to predict the closing price for the stock market the next day using a Support Vector Machine Regressor with a Radial Basis Function Kernel. As one task for our final thesis, we propose to investigate the CCA technique as a way to fuse NIPW and reflectivity fields and to investigate the use of the resulting correlated field of NIPW and reflectivity as an input to a machine learning nowcast algorithm. The CCA technique has been implemented in the Python machine learning package scikit-learn [36], and can be readily applied to our problem using the two fields.

In addition to the above, we are particularly interested in investigating an application of the so-called deep learning techniques to our problem. Deep learning is a new area of machine learning research at the cutting-edge of artificial intelligence (cf. http://deeplearning.net) that is just recently starting to be applied to the problem of weather forecasting [47], [22]. We propose to investigate the use of deep learning to not only make a no-rain, rain binary nowcasts, but to try to nowcast rainfall intensity as well. For this effort we propose to use the extensive library of deep learning tools from the Python package Theano [8]. Our plan is to use a GPU instance from Amazon Web Services (AWS) and run our nowcasting experiments in the cloud. This will enable us to try new things and speed up computation significantly.

6.2 Expected Contributions

The expected contributions of this thesis work are as follows:

1. A hardware/software design for a low-cost GPS-Met station capable of near-real-time IPW estimates. This work is documented in the publication [1].

2. Deployment of two of the low-cost GPS-Met stations in the DFW metroplex that include high-resolution barometers for the DFW Weather Forecast Office and other researchers to use for IPW and fine resolution barometric pressure analysis during events such the passing of nearby tornados and other weather anomalies.

3. A Python-based suite of tools for near-real-time estimates of IPW fields from the network of TxDOT GPS-receivers and NWS ASOS weather stations in the DFW region. This set of tools is capable of automatically downloading the necessary GPS and ASOS data from on-line databases, putting the data in appropriate RINEX formats and directories, executing GAMIT for IPW estimation, and then calling a multiquadric interpolation algorithm to map the results to a field.

4. An improved understanding of the relationships between IPW and precipitation, including the ability to make movies to show visually the joint spatial-temporal evolution of IPW and weather radar reflectivity, and through analysis of machine-learning techniques such as the Random Forest classifier that can rank the importance of the input variables.

5. Development of Python-based tools analyzing and conducting machine-learning experiments with spatial-temporal fields of weather data that others will be able to use for future weather forecasting and analysis studies.

6. Development, analysis, and comparison of different machine learning approaches for precipitation nowcasting from spatial-temporal sequences of normalized IPW and weather radar reflectivity, including the Naive Bayes classifer, Random Forest classifier, CCA based feature extraction, and deep learning neural network approaches.

7. Construction of a data set of IPW, pressure, temperature, relative humidity, and reflectivity data for the 230 km region surrounding Dallas-Fort-Worth for the entire year of 2014 that others can use as a "challenge data set" for developing and testing big data analytics techniques and machine learning weather prediction algorithms.

6.3 Timeline

The tentative timeline for completing this thesis is summarized in Table 6.1.

Table 6.1: Timeline

	Activity	Completion Date
1	Implementation of a deep learning neural network precipitation nowcast algorithm similar to the one explored in [47].	01/05/16
2	Exploration into the use of the CCA technique for dimensionality reduction, feature extraction and precipitation nowcasting.	01/05/16
3	Oral presentation of the precipitation nowcasting work at the 2016 Annual Meeting of the American Meteorological Society (AMS) in New Orleans, LA.	01/12/16
4	Completion of thesis writeup for distribution to committee members. We are also considering the submission of a journal version of the thesis work.	01/30/16
5	Thesis defense.	Feb 2016

APPENDIX

DFW PRECIPITABLE WATER VAPOR, REFLECTIVITY NETWORK

This appendix gives the locations of the sensor assets used for the studies in this thesis.

Weather radar reflectivity data came from the KFWS WSR-88D weather radar located in Fort-Worth, TX (32.569 Deg Lat, -97.299 Deg Lon). This radar defined the center point of our 300 km by 300 km region of study. The table in Figure A.1 gives the locations of the GPS stations used along with the location of the closest ASOS weather station from which the met variables needed for IPW estimation were obtained. Table A.1 gives the location of the long baseline stations used for the double differencing processes.

Table A.1: Long baseline stations.

Site ID	Location	Latitude(d)	Longitude(d)	Height(m)
AC20	Girdwood Alaska	60.92	-149.35	43.73
CONZ	Concepcion Chile	-36.84	-73.07	176.22
P019	Fairfield, Idaho	43.30	-115.31	1682.39
UNBJ	University of New Brunswick, Canada	45.95	-66.64	22.8

GPSid	GPSlat	GPSlong	GPSheight	ASOSid	ASOSlat	ASOSlong	ASOSheight	GPSgeoid	GPSMSL	ASOS_GPS	Network
okar	34.1684639	-97.169244	235.8	K1F0	34.15	-97.11	257	-26.045	261.845	-4.845	net1
okdn	34.4793111	-97.966553	314.3	KDUC	34.466	-97.96	339	-25.651	339.951	-0.951	net1
txbn	33.6067278	-96.175331	160.5	KDUA	33.95	-96.4	213	-26.224	186.724	26.276	net1
txbt	31.03265	-97.479008	177.1	KTPL	31.133	-97.4	208	-27.053	204.153	3.847	net1
txbu	30.7504611	-98.184386	438	KBMQ	30.733	-98.23	389	-24.668	462.668	-73.668	net1
txbx	30.7178389	-96.396622	83.6	KCFD	30.716	-96.33	112	-26.409	110.009	1.991	net1
txc2	30.8765028	-96.972342	97	KT35	30.883	-96.96	123	-26.074	123.074	-0.074	net1
txc3	31.8098139	-99.422094	498.3	KCOM	31.833	-99.4	517	-25.751	524.051	-7.051	net1
txck	31.3226333	-95.435908	87.5	KDKR	31.3	-95.4	106	-26.947	114.447	-8.447	net1
txco	33.1652639	-96.627944	161.9	KTKI	33.183	-96.58	179	-25.697	187.597	-8.597	net1
txda	32.7999861	-96.672914	160.6	KDAL	32.85	-96.85	158	-26.043	186.643	-28.643	net1
txdc	33.2362222	-97.608672	255.3	KLUD	33.25	-97.58	319	-27.449	282.749	36.251	net2
txde	33.2104528	-97.162778	178.8	KDTO	33.2	-97.2	196	-26.683	205.483	-9.483	net2
txea	32.4027694	-98.808933	407.5	KBKD	32.716	-98.88	392	-28.408	435.908	-43.908	net2
txes	32.3696972	-96.862775	163.7	KJWY	32.45	-96.91	217	-26.595	190.295	26.705	net2
txge	33.1319667	-96.055539	134.9	KGVT	33.066	-96.06	163	-25.57	160.47	2.53	net2
txgl	31.4721722	-98.567969	459.8	KMNZ	31.666	-98.15	396	-25.601	485.401	-89.401	net2
txgr	32.2403861	-97.754444	177.5	KGDJ	32.45	-97.81	237	-27.896	205.396	31.604	net2
txhi	31.9892111	-97.129797	153.8	KINJ	32.083	-97.1	209	-27.093	180.893	28.107	net2
txhm	31.6994889	-98.106747	345.1	KMNZ	31.666	-98.15	396	-26.766	371.866	24.134	net2
txja	33.1948222	-98.145625	326	KXBP	33.183	-97.83	260	-28.045	354.045	-94.045	net2
txka	32.5717556	-96.314278	115.9	KTRL	32.716	-96.26	145	-25.319	141.219	3.781	net2
txke	32.4097083	-97.323236	227.9	KCPT	32.35	-97.43	260	-27.506	255.406	4.594	net3
txmn	31.9100833	-97.661911	210.6	KACT	31.616	-97.23	151	-27.51	238.11	-87.11	net3
txmv	33.1618361	-95.221089	132.5	KOSA	33.1	-94.96	111	-26.649	159.149	-48.149	net3
txmw	32.8041694	-98.142889	246.4	KMWL	32.783	-98.06	284	-29.173	275.573	8.427	net3
txmx	31.5951222	-96.524375	119.7	KLXY	31.633	-96.51	166	-25.356	145.056	20.944	net3
txna	32.0417833	-96.538744	105.4	KCRS	32.033	-96.4	133	-25.57	130.97	2.03	net3
txno	33.7756556	-97.726028	258.3	K0F2	33.6	-97.78	336	-26.319	284.619	51.381	net3
txol	33.3560139	-98.7497	331.8	KRPH	33.116	-98.55	342	-28.487	360.287	-18.287	net3
txpa	33.6742361	-95.557025	145	KSLR	33.166	-95.61	149	-27.029	172.029	-23.029	net3
txpi	31.7244944	-95.594925	124.6	KPSN	31.783	-95.7	129	-26.423	151.023	-22.023	net3
txru	31.7849028	-95.126172	146	KJSO	31.866	-95.21	206	-26.621	172.621	33.379	net3
txsg	32.8557167	-97.344172	181.7	KFTW	32.833	-97.36	214	-27.801	209.501	4.499	net4
txsr	33.5915722	-96.607003	194.3	KGYI	33.716	-96.66	228	-25.829	220.129	7.871	net4
txst	32.2325917	-98.182194	376.6	KSEP	32.216	-98.16	402	-27.857	404.457	-2.457	net4
txsy	33.6024333	-99.258442	367.3	KCWC	33.85	-98.48	305	-28.398	395.698	-90.698	net4
txta	30.564225	-97.445042	147.7	KGTU	30.683	-97.68	240	-26.159	173.859	66.141	net4
txth	33.1789694	-99.1679	371.9	KBKD	32.716	-98.88	392	-29.062	400.962	-8.962	net4
txty	32.2496167	-95.393611	120.1	KTYR	32.366	-95.4	165	-26.258	146.358	18.642	net4
txwa	31.5777167	-97.110511	101.7	KACT	31.616	-97.23	151	-26.844	128.544	22.456	net4
txwe	32.7588972	-97.823478	337.4	KMWL	32.783	-98.06	284	-28.658	366.058	-82.058	net4
txwf	33.853925	-98.505556	280.2	KSPS	33.983	-98.5	308	-27.701	307.901	0.099	net4
zfw1	32.83065	-97.066469	155.2	KDFW	32.9	-97.01	174	-27.252	182.452	-8.452	net4

Figure A.1: GPS stations and ASOS ststions table.

BIBLIOGRAPHY

[1] A, Nagarajan, MC, Jacques, A, Lagace, DL, Pepyne, M, Zink, DJ, and McLaughlin. Lower-cost gps met station design for use in dense network slant path gps-met estimates of tropospheric wet delay and precipitable water vapor. In *Proceedings of 19th Conference on Integrated Observing and Assimilation Systems for the Atmosphere, Oceans, and Land Surface (IOAS-AOLS)* (2015), AMS.

[2] Akilan, A, Azeez, KK Abdul, Balaji, S, Schuh, H, and Srinivas, Y. Gps derived zenith total delay (ztd) observed at tropical locations in south india during atmospheric storms and depressions. *Journal of Atmospheric and Solar-Terrestrial Physics 125* (2015), 1–7.

[3] Alber, Chris, Ware, Randolph, Rocken, Christian, and Braun, John. Obtaining single path phase delays from gps double differences. *Geophysical Research Letters 27*, 17 (2000), 2661–2664.

[4] Arel, Itamar, Rose, Derek C, and Karnowski, Thomas P. Deep machine learning-a new frontier in artificial intelligence research [research frontier]. *Computational Intelligence Magazine, IEEE 5*, 4 (2010), 13–18.

[5] Bai, Zhengdong, and Feng, Yanming. Gps water vapor estimation using interpolated surface meteorological data from australian automatic weather stations. *Journal of Global Positioning Systems 2*, 2 (2003), 83–89.

[6] Barry, Roger G, and Chorley, Richard J. *Atmosphere, weather and climate*. Routledge, 2009.

[7] Batelaan, PD, Sato, T, Slobin, SD, and Reilly, H. Development of a water vapor radiometer to correct for tropospheric range delay in dsn applications. *DSN Progress Report 42 33* (1976), 77–84.

[8] Bergstra, James, Breuleux, Olivier, Bastien, Frédéric, Lamblin, Pascal, Pascanu, Razvan, Desjardins, Guillaume, Turian, Joseph, Warde-Farley, David, and Bengio, Yoshua. Theano: a CPU and GPU math expression compiler. In *Proceedings of the Python for Scientific Computing Conference (SciPy)* (June 2010). Oral Presentation.

[9] Bevis, Michael, Businger, Steven, Chiswell, Steven, Herring, Thomas A, Anthes, Richard A, Rocken, Christian, and Ware, Randolph H. Gps meteorology: Mapping zenith wet delays onto precipitable water. *Journal of applied meteorology 33*, 3 (1994), 379–386.

[10] Bevis, Michael, Businger, Steven, HERRING, THOMASA, Rocken, Christian, ANTHES, RICHARDA, and WARE, RANDOLPHH. Gps meteorology- remote sensing of atmospheric water vapor using the global positioning system. *Journal of Geophysical Research 97*, D14 (1992), 15787–15801.

[11] Bock, Y, Behr, J, Fang, P, Dean, J, and Leigh, R. Scripps orbit and permanent array center (sopac) and southern californian permanent gps geodetic array (pgga). *The Global Positioning System for the Geosciences* (1997), 55–61.

[12] Breiman, Leo. Random forests. *Machine learning 45*, 1 (2001), 5–32.

[13] Bringi, VN, and Chandrasekar, V. *Polarimetric Doppler weather radar: principles and applications*. Cambridge University Press, 2001.

[14] Davis, JL, Herring, TA, Shapiro, II, Rogers, AEE, and Elgered, Gunnar. Geodesy by radio interferometry: Effects of atmospheric modeling errors on estimates of baseline length. *Radio science 20*, 6 (1985), 1593–1607.

[15] de Haan, Siebren, Holleman, Iwan, and Holtslag, Albert AM. Real-time water vapor maps from a gps surface network: Construction, validation, and applications. *Journal of Applied Meteorology and Climatology 48*, 7 (2009), 1302–1316.

[16] Doviak, Richard J, and Zrnic, Dusan S. *Doppler Radar & Weather Observations*. Dover, 1993.

[17] Dow, John M, Neilan, RE, and Rizos, C. The international gnss service in a changing landscape of global navigation satellite systems. *Journal of Geodesy 83*, 3-4 (2009), 191–198.

[18] Duan, Jingping, Bevis, Michael, Fang, Peng, Bock, Yehuda, Chiswell, Steven, Businger, Steven, Rocken, Christian, Solheim, Frederick, van Hove, Terasa, Ware, Randolph, et al. Gps meteorology: Direct estimation of the absolute value of precipitable water. *Journal of Applied Meteorology 35*, 6 (1996), 830–838.

[19] FORSYTHE, JOHN M, KIDDER, STANLEY Q, FUELL, KEVIN K, LEROY, ANITA, JEDLOVEC, GARY J, and JONES, ANDREW S. A multisensor, blended, layered water vapor product for weather analysis and forecasting. *Journal of Operational Meteorology 3*, 5 (2015).

[20] French, Mark N, Krajewski, Witold F, and Cuykendall, Robert R. Rainfall forecasting in space and time using a neural network. *Journal of hydrology 137*, 1 (1992), 1–31.

[21] Friedman, Jerome, Hastie, Trevor, and Tibshirani, Robert. *The elements of statistical learning*, vol. 1. Springer series in statistics Springer, Berlin, 2001.

[22] Grover, Aditya, Kapoor, Ashish, and Horvitz, Eric. A deep hybrid model for weather forecasting. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2015), ACM, pp. 379–386.

[23] Grumm, Richard H, and Hart, Robert. Standardized anomalies applied to significant cold season weather events: Preliminary findings. *Weather and forecasting 16*, 6 (2001), 736–754.

[24] Guo, Zhiqiang, Wang, Huaiqing, Liu, Quan, and Yang, Jie. A feature fusion based forecasting model for financial time series.

[25] Gurtner, Werner, and Estey, Lou. Rinex-the receiver independent exchange format-version 3.00. *Astronomical Institute, University of Bern and UNAVCO, Bolulder, Colorado.* (2007).

[26] Herring, TA, King, RW, and McClusky, SC. Gamit reference manual. *GPS Analysis at MIT, release 10* (2015), 36.

[27] Hoffman-Wellenhof, B, Lichtenegger, Herbert, and Wasle, Elmar. Gnss?global navigation satellite systems. *GPS, GLONASS, Galileo and more. Wien: Springer-Verlag* (2008).

[28] Inoue, Hanako Y, and Inoue, Toshiro. Characteristics of the water-vapor field over the kanto district associated with summer thunderstorm activities. *SOLA 3* (2007), 101–104.

[29] Iwasaki, Hiroyuki, and Miki, Takahiro. Diurnal variation of convective activity and precipitable water over the" semi-basin". preliminary study on the mechanism responsible for the evening convective activity maximum. *????. ? 2 ? 80*, 3 (2002), 439–450.

[30] Johnson, JT, MacKeen, Pamela L, Witt, Arthur, Mitchell, E De Wayne, Stumpf, Gregory J, Eilts, Michael D, and Thomas, Kevin W. The storm cell identification and tracking algorithm: An enhanced wsr-88d algorithm. *Weather and Forecasting 13*, 2 (1998), 263–276.

[31] Kuligowski, Robert J, and Barros, Ana P. Localized precipitation forecasts from a numerical weather prediction model using artificial neural networks. *Weather and Forecasting 13*, 4 (1998), 1194–1204.

[32] Lynch, Peter. The origins of computer weather prediction and climate modeling. *Journal of Computational Physics 227*, 7 (2008), 3431–3444.

[33] Marshall, John S, and Palmer, W Mc K. The distribution of raindrops with size. *Journal of meteorology 5*, 4 (1948), 165–166.

[34] McGovern, Amy, Supinie, TIMOTHY, Gagne, II, Collier, M, Brown, RA, Basara, J, and Williams, J. Understanding severe weather processes through spatiotemporal relational random forests. In *2010 NASA conference on intelligent data understanding (to appear)* (2010).

[35] Mecikalski, John R, Williams, John K, Jewett, Christopher P, Ahijevych, David, LeRoy, Anita, and Walker, John R. Probabilistic 0–1-h convective initiation nowcasts that combine geostationary satellite observations and numerical weather prediction model data. *Journal of Applied Meteorology and Climatology 54*, 5 (2015), 1039–1059.

[36] Pedregosa, Fabian, Varoquaux, Gaël, Gramfort, Alexandre, Michel, Vincent, Thirion, Bertrand, Grisel, Olivier, Blondel, Mathieu, Prettenhofer, Peter, Weiss, Ron, Dubourg, Vincent, et al. Scikit-learn: Machine learning in python. *The Journal of Machine Learning Research 12* (2011), 2825–2830.

[37] Powers, David Martin. Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation.

[38] Radhakrishna, Basivi, Fabry, Frédéric, Braun, John J, and Van Hove, Teresa. Precipitable water from gps over the continental united states: Diurnal cycle, intercomparisons with narr, and link with convective initiation. *Journal of Climate 28*, 7 (2015), 2584–2599.

[39] Remondi, Benjamin W. Using the global positioning system(gps) phase observable for relative geodesy: Modeling, processing, and results[ph. d. thesis].

[40] Rocken, Christian, Hove, Teresa Van, Johnson, James, Solheim, Fred, Ware, Randolph, Bevis, Mike, Chiswell, Steve, and Businger, Steve. Gps/storm-gps sensing of atmospheric water vapor for meteorology. *Journal of Atmospheric and Oceanic Technology 12*, 3 (1995), 468–478.

[41] Ruzanski, Evan, Chandrasekar, V, and Wang, Yanting. The casa nowcasting system. *Journal of Atmospheric and Oceanic Technology 28*, 5 (2011), 640–655.

[42] Saastamoinen, J. Atmospheric correction for the troposphere and stratosphere in radio ranging satellites. *The use of artificial satellites for geodesy* (1972), 247–251.

[43] Seko, Hiromu, Nakamura, Hajime, Shoji, Yoshinori, and Iwabuchi, Tetsuya. The meso-. gamma. scale water vapor distribution associated with a thunderstorm calculated from a dense network of gps receivers. *????. ? 2 ? 82*, 1B (2004), 569–586.

[44] Shan, Caifeng, Gong, Shaogang, and McOwan, Peter W. Beyond facial expressions: Learning human emotion from body gestures. In *BMVC* (2007), Citeseer, pp. 1–10.

[45] Shangguan, Ming. *Analysis and derivation of spatial and temporal distribution of water vapor from GNSS observations*. PhD thesis, Technische Universität Berlin, 2014.

[46] Shi, Junbo, Xu, Chaoqian, Guo, Jiming, and Gao, Yang. Real-time gps precise point positioning-based precipitable water vapor estimation for rainfall monitoring and forecasting. *Geoscience and Remote Sensing, IEEE Transactions on 53*, 6 (2015), 3452–3459.

[47] Shi, Xingjian, Chen, Zhourong, Wang, Hao, Yeung, Dit-Yan, Wong, Wai-Kin, and Woo, Wang-chun. Convolutional lstm network: A machine learning approach for precipitation nowcasting. *arXiv preprint arXiv:1506.04214* (2015).

[48] Shoji, Yoshinori, Yamauchi, Hiroshi, Mashiko, Wataru, and Sato, Eiichi. Estimation of local-scale precipitable water vapor distribution around each gnss station using slant path delay. *SOLA 10*, 0 (2014), 29–33.

[49] Snay, Richard A, and Soler, Tomás. Continuously operating reference station (cors): history, applications, and future enhancements. *Journal of Surveying Engineering 134*, 4 (2008), 95–104.

[50] Spilker, JJ. Gps signal structure and performance characteristics. *Global Positioning System 1* (1980), 29–54.

[51] Sun, Shiliang. A survey of multi-view machine learning. *Neural Computing and Applications 23*, 7-8 (2013), 2031–2038.

[52] Tabios, Guillermo Q, and Salas, Jose D. A comparative analysis of techniques for spatial interpolation of precipitation1, 1985.

[53] Terradellas, E, and Téllez, B. The use of products from ground-based gnss observations in meteorological nowcasting. *Advances in Geosciences 26*, 26 (2010), 77–82.

[54] Tralli, David M, and Lichten, Stephen M. Stochastic estimation of tropospheric path delays in global positioning system geodetic measurements. *Bulletin géodésique 64*, 2 (1990), 127–159.

[55] Wolfe, Daniel E, and Gutman, Seth I. Developing an operational, surface-based, gps, water vapor observing system for noaa: Network design and results. *Journal of Atmospheric and Oceanic Technology 17*, 4 (2000), 426–440.

[56] Yoshinori, SHOJI. Retrieval of water vapor inhomogeneity using the japanese nationwide gps array and its potential for prediction of convective precipitation. *????. ? 2 ? 91*, 1 (2013), 43–62.

[57] Zhang, Harry. The optimality of naive bayes. *AA 1*, 2 (2004), 3.