Project Report: Team 5
White Christmas Prediction

Ashwin Giridharan, Datt Goswami, Chaitanya Mallampati, and Jiemin Zeng

Department of Computer Science, Stony Brook University,
Stony Brook, NY 11794-4400
{ashwin.giridharan,datt.goswami,saichaitanya.mallampati,jiemin.zeng}@cs.stonybrook.edu
http://www.cs.stonybrook.edu/~skiena/591/projects

1 Challenge

Weather forecasting is a scientific and technological application in which the main goal is to predict the state of the atmosphere of a particular location at a particular instance of time.

As Patrick Young once said - *"The trouble with weather forecasting is that it is right too often for us to ignore it and wrong too often for us to rely on it."* [1]

If the above statement doesn't emphasize enough, weather forecasting is truly a complex problem. Here we discuss four important aspects that make accurate weather forecasting a challenging task.

- Firstly, for weather forecasting to be considerably accurate, many unpredictable atmospheric variables need to be considered and incorporated into the model.
- Secondly, for making accurate weather predictions we need to deal with large amount of data. The larger the dataset the better the chances of making a well informed prediction.
- Thirdly, weather prediction is computationally expensive. This is a direct result of the first two factors. Using a large number of atmospheric variables to build a model which predicts future patterns is obviously computationally expensive.
- Last but not the least, unfortunately we aren't that technologically advanced yet to carry out these computationally expensive calculations for long term predictions.

As a part of this project, we deal with one small aspect of weather forecasting which is predicting white christmas. So, what is white christmas ? Over the years many definitions of white christmas existed depending upon the era and depending upon the place. For example, in most countries, it simply means that the ground is covered by snow at Christmas, but some countries have more strict definitions. In Canada the official definition is that there has to be more than 2 cm (0.79 in) on the ground on Christmas Day.[2] As part of our project, we stick to the following definition of a white Christmas - there has to be a snow depth of at least 1 in or 2.5 cm on the morning of Christmas day.

2 History/Background

Weather forecasting existed since millennia. In 650 BC, the Babylonians predicted the weather from cloud patterns as well as astrology. In about 340 BC, Aristotle described weather patterns in Meteorologica. Later, Theophrastus compiled a book on weather forecasting, called the Book of Signs. Chinese weather prediction lore extends at least as far back as 300 BC, which was also around the same time ancient Indian astronomers developed weather-prediction methods. In 904 AD, Ibn Wahshiyya's Nabatean Agriculture discussed the weather forecasting of atmospheric changes and signs from the planetary astral alterations; signs of rain based on observation of the lunar phases; and weather forecasts based on the movement of winds. [3]

Eventually, we moved from these ancient methods to more modern scientific methods. Before moving on to the different existing weather prediction models, let us define the following terminology:

- **Short Term Weather Prediction Model :** A short term weather prediction model is a weather prediction model which is capable of making decently accurate weather predictions only for a few days into the future. Typically, this period ranges within 5-7 days from the present day.

- **Long Term Weather Prediction Model :** A long term weather prediction model is a weather prediction model which is capable of making decently accurate weather predictions longer into the future. Typically, this period ranges beyond 5-7 days from the present day.

There exist several different methods for weather forecasting. Based on the availability of input information, the forecasting problem, and the degree of accuracy or confidence required, suitable models are used. The following are the most common existing weather prediction models: [4]

- **Persistence Method :** This is one of the simplest and the most naive weather prediction model. The basic assumption of this model is that the conditions present at the time of making a forecast won't change. For example, if there is a snowfall of two inches today, this model predicts two inches of snowfall for tomorrow as well. This relatively short term model is effective in cases where the weather doesn't vary as much.

- **Trends Method :** This model has mathematics at its core. For example, if a tornado is approaching London at 1000 miles a day and is 2500 miles away now, this model predicts the tornado to reach London in two and a half days. The basic assumption in this model is that weather fronts remain constant through time. If that is not the case, this model may not be accurate. This is also a relatively short term weather prediction model.

- **Climatology Method :** This model is a simple model which uses statistics to predict future weather patterns. For example, we need to predict the

amount of snowfall on Christmas in London. This model takes an average of the amounts of snowfall on Christmas in London during the previous years and uses this as the basis for prediction. The basic assumption of this model is that during the period of consideration, the weather will be uniform across all the years.

- **Analog Method :** This model involves examining today's weather forecast scenario and finding a corresponding analog (a day in the past when the weather scenario matches to the present scenario). This model makes a weather prediction similar to the one observed in the past for the current scenario as well. One of the major drawbacks of this model is that it is not always possible to find a perfect analog and even if a suitable analog is found, the chances of the pattern repeating may not be very high in most of the cases.

- **Numerical Weather Prediction :** Numerical Weather Prediction is a more sophisticated methodology in which forecasting models for each of the environmental variables is run on supercomputers. Although slightly error prone, this model is capable of making long term weather predictions.

Each of the above discussed models have their own benefits and drawbacks and we intend to use a hybrid of these methods to come up with our white christmas prediction.

3 Literature Review

Predicting weather well ahead of time has many socio-economic benefits. For example, knowing about an approaching tornado in advance can contribute immensely towards saving lives and minimizing property damage. Due to many such practical benefits, weather forecasting as a research activity has been around for a long time. As a result of the knowledge gathered from these research activities, we were successfully able to increase our understanding of weather gradually over time. Now, research in weather has evolved immensely with its application now spanning from biology to aeronautics. Despite such progress, there is a lot of work to be done before we can confidently claim to have mastered weather and the impact it has. For this project as part of literature survey we studied some of these research works in recent years which focussed mainly on precipitation and snowfall patterns across the globe as well as the correlation between temperature and snowfall patterns.

The first paper we looked at was "Variability and trends of total precipitation and snowfall over the United States and Canada." [5] by Pavel Ya. Groisman and David R. Easterling. This paper aims to study the variability and trends of precipitation and snowfall over the United States and Canada. For this study the authors have used precipitation and snowfall data from various United States

and Canada stations. As part of the data pre-processing stage, many scale adjustments had to be done. For example, the Canadian rainfall data was adjusted in order to incorporate for the gauge change that occurred in the mid-1970s. On this data several statistical analyses were applied and the authors made the following observations:

- During the last century, annual precipitation has increased in southern Canada by 13% and in the contiguous United States by 4%.
- During the last four decades, a 20% increase has occurred in annual snowfall and rainfall in northern Canada.
- Analysis of the relationship between century-long precipitation time series over North America with Northern Hemisphere surface air temperature and the South Oscillation index showed that ENSO is usually accompanied by an increase of precipitation.

Next we went through the research paper "A global analysis of snow depth for numerical weather prediction." [6] by Bruce Brasnett. This paper uses a two-step numerical process to analyze and predict the depth of snow. Firstly, the amount of snowfall is estimated from Numerical Weather Prediction model. This model uses various weather variables which affect snowfall to come up with the prediction. Next, a temperature based melting algorithm is defined. Combining these two studies, the authors arrive at an estimate for snow depth.

Now that we have gained some insight into how the precipitation and snowfall varied across the years, we dig further deep into the domain by examining the variation of snowfall with other weather parameters . For this, we first read the research work "Recent variations of snow cover and snowfall in North America and their relation to precipitation and temperature variations." [7] by Thomas R. Karl, Pavel Ya. Groisman, Richard W. Knight and Richard R. Heim Jr. Then we also read "Snow cover in eastern Europe in relation to temperature, precipitation and circulation.", [8]. Both these papers assert the following -

- A strong positive correlation between the annual number of days with snow cover and the annual number of days with mean temperature less than 0 degree C was discovered for most parts of the study area.
- A negative correlation between the monthly number of days with snow cover and monthly mean temperature was found.
- A positive correlation between snow depth and precipitation appeared significant only in some areas.

In an attempt to complement these observations, we went ahead and used our large dataset to get a clear understanding of the correlation between mean temperature and snow cover.

Firstly, we obtained the correlation between mean monthly temperature [Oct - Mar] and mean monthly snow cover count. When we took values of temperature and snow cover for the months of October to March, we got a high correlation. For a linear dependence, we obtained the correlation coefficient R as -0.87 and when we looked at higher order dependence (quadratic and cubic) the correlation

coefficient R was obtained as -0.95. [Figure 1] One more interesting observation that can be visualized is that snow cover has an optimal temperature range i.e., too cold temperatures or too warm temperatures are not ideal for snow cover. We will further explore this in the observations section. [Figure 6 and 7]

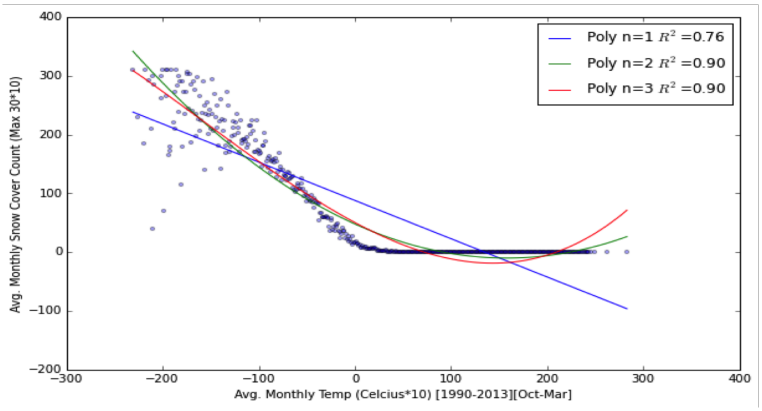


Fig. 1: Correlation - mean snow cover count & mean monthly temp [Oct-Mar]

Next, we went on and obtained the correlation between mean temperatures in the month of November and mean White Christmas snow depth (snow depth on December 25). For a linear dependence, we obtained the correlation coefficient R as -0.49 and when we looked at higher order dependence (quadratic and cubic) the correlation coefficient R was obtained as -0.59. This relation can be seen in figure 2.

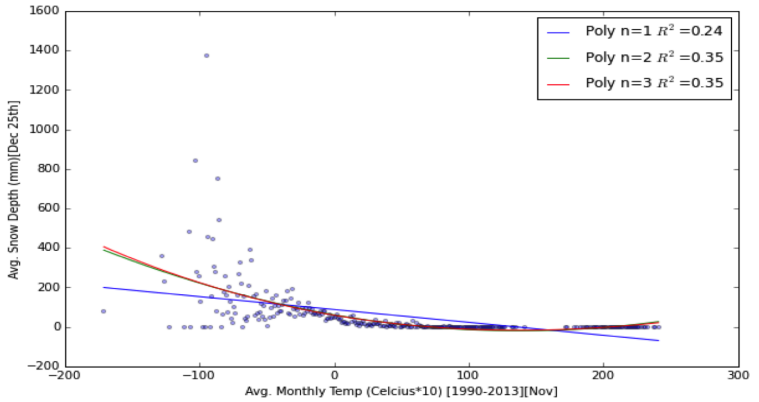


Fig. 2: Correlation - mean snow depth [Dec 25] & mean November monthly temperature

Next, we went on and obtained the correlation between mean temperatures in the month of November and mean snow depth from December 23 to 27. For a linear dependence, we obtained the correlation coefficient R as -0.63 and when we looked at higher order dependence (quadratic and cubic) the correlation coefficient R was obtained as -0.76. This relation can be seen in figure 3.

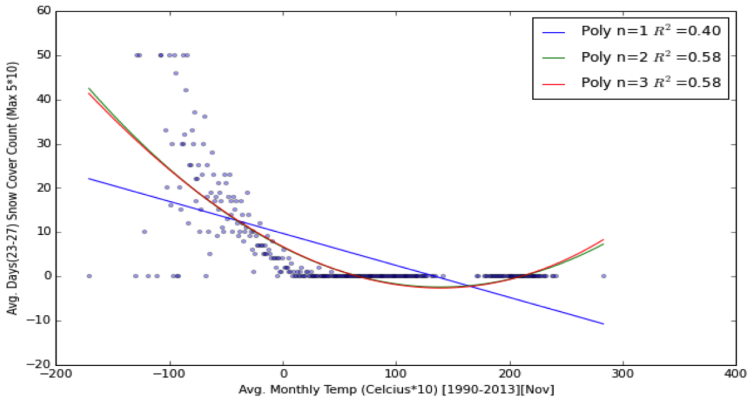


Fig. 3: Correlation - mean snow depth [Dec 23-27] & mean November monthly temperature

The above study clearly indicates a strong correlation between temperature and snowfall.

Before building any weather forecasting model, a basic understanding of what a good weather forecast is and what a bad weather forecast is, is absolutely necessary. In an attempt to understand what a good weather forecast is and how it differs from a bad and inaccurate weather forecast, we concluded our literature survey by reading "what is a good forecast? An essay on the nature of goodness in weather forecasting." [9] by Allan H. Murphy. This paper proposes three measures to define goodness/badness of a weather forecasting model:

- **Consistency :** Goodness/Badness is measured based on the correspondence between the forecasters' judgment and their forecasts
- **Quality :** Goodness/Badness is measured based on the correspondence between the forecasts and the observation
- **Value :** Goodness/Badness is measured based on the incremental economic and/or other benefits realized by the decision makers through the use of the forecasts

The paper also studies the relationship among these three measures. It shows that the level of consistency in a weather prediction has a direct impact on the

quality and value of the prediction.

To further our domain knowledge, apart from literature surveys, we also met with Prof. Edmund Cheng from Atmospheric and Marine Sciences Department at Stony Brook University.

Few of the keys inputs the professor provided are -

- White Christmas prediction is a easy task when it comes to prediction in areas like Texas (where the probability of white Christmas is 0 every year) and Wisconsin (where the probability of white Christmas is 1 every year). It is a much more complicated job when predicting in areas where the chances of snowfall don't belong to either extreme.
- One basic way to predict whether or not it is going to be a white Christmas is to make use of climatological models. These models make a statistical forecast based on the past occurrences of white Christmas.
- A much more advanced method to predict white Christmas is to build a correlation model between snow-depth and other weather parameters like temperature, precipitation, sea surface temperatures, air moisture, atmospheric pressure etc. For the correlation model to work, the parameters used in the model must have a considerable correlation with the actual snow-depth parameter.

4 Data Sets

4.1 Source of the dataset

The US National Oceanic and Atmospheric Administration (NOAA) is an immense and helpful resource for our project [10]. As a US government agency, the NOAA has three main goals; to obtain and provide information about the oceans and atmosphere, to act as a steward of the US coastal and marine environments, and to promote scientific research in ecosystems, climate, weather and water, and commerce and transportation. As part of the first mission, the NOAA provides global historical climate data free of charge to the public with the Global Historical Climatology Network (GHCN) [11]. The GHCN is a large database of historical daily climate data from over 75,000 stations in 180 countries around the world. An extensive resource widely used in climatology, the GHCN has many records going back more than 100 years.

To obtain the data, the NOAA provides a user friendly GUI interface as well as ftp access. There are advantages to both. When obtaining data from the GUI interface, weather stations can be plotted on a map and searched by location. This has been useful in determining which weather stations are in a particular city. In addition, there are options to specify specific fields of data and the specific stations to retrieve the data from. The data requested is also formatted in a easy to parse csv format. The downside to retrieving data from the GUI interface is that there is a 1 GB limit on the amount of data for each request and the one request can take a whole day to fulfill.

The main advantage of retrieving data through ftp access is that is very quick as the NOAA servers do not have to process the specific details of the request. However, the are several downsides as well. The first downside is that while the data from each station is in it's own file, we have to determine which stations are within each city. The NOAA also provides a file of all of the stations and their names. While most station names contain the name of the city it is in, some do not. As a result, we used the GUI tool to manually target the stations we need to collect data. The second downside is that the data is in a compact form and requires additional reformatting not needed in the GUI approach. Each line in a data file contains one month of data for a single parameter.

4.2 Dataset Description

The dataset is a collection of .dly files. Each .dly file contains data for one station. Each .dly file contains numerous variables, the key ones among which are the following -

- ID - Station Identification Code
- Year - Year of the record
- Month - Month of the record
- Element - type of element

There are five core elements that each of the stations recorded -

- PRCP - precipitation (tenths of mm)
- SNOW - snowfall (in mm)
- SNWD - snow depth (in mm)
- TMAX - maximum temperature (tenths of degree C),
- TMIN - minimum temperature (tenths of degree C).

In addition to these five, other data elements include details about the cloudiness, evaporation, details about the frozen ground layer, details about the wind, the soil temperature, and many many more. Most stations do not collect all of the data elements. In fact most stations do not even collect all five core data elements.

5 Observations

Now that we have obtained our dataset and have some understanding of how the different weather parameters are related, we played around with the data and did some exploratory data analysis.

Figure 4 shows the time span of the data we retrieved for all the capital cities of the states in US

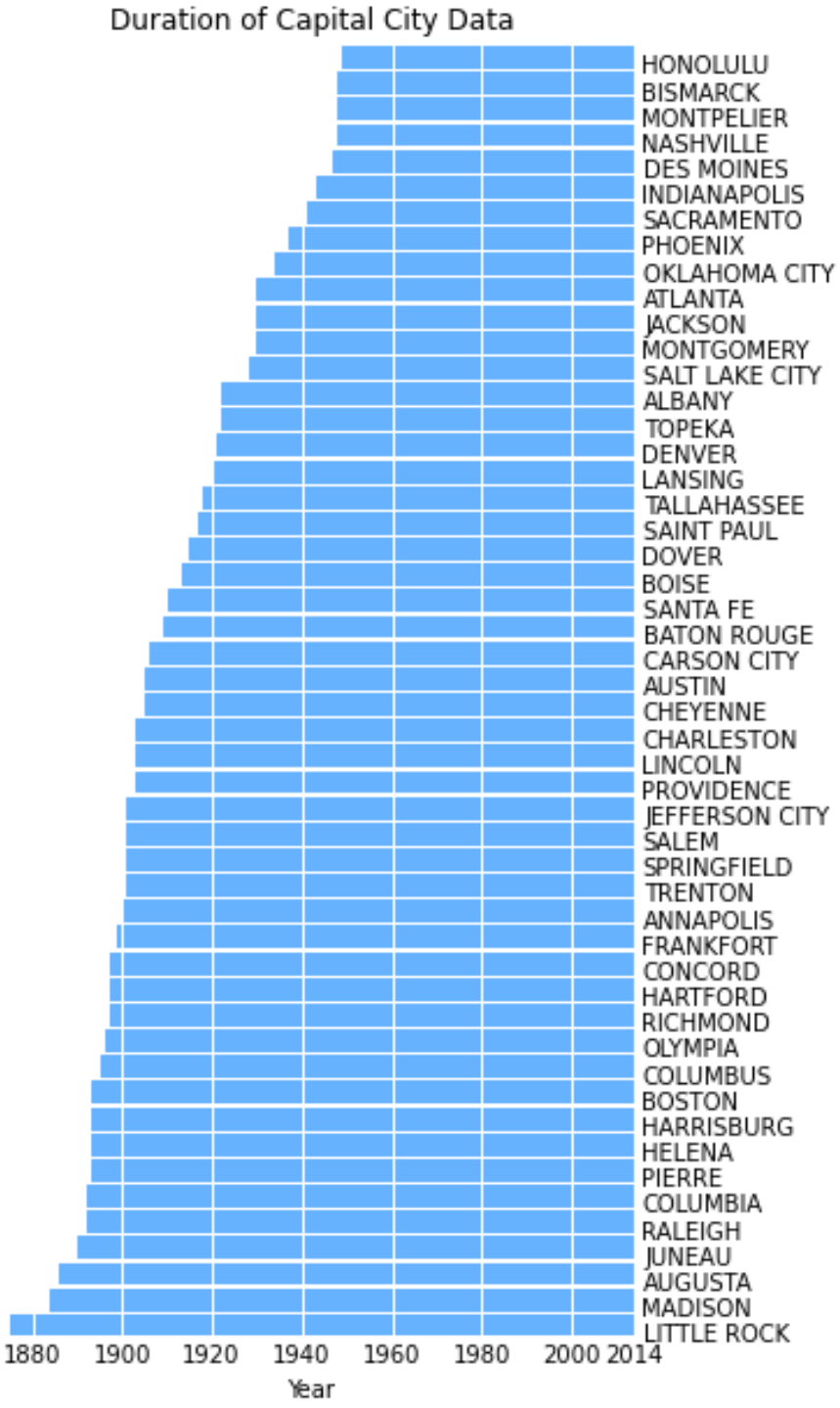
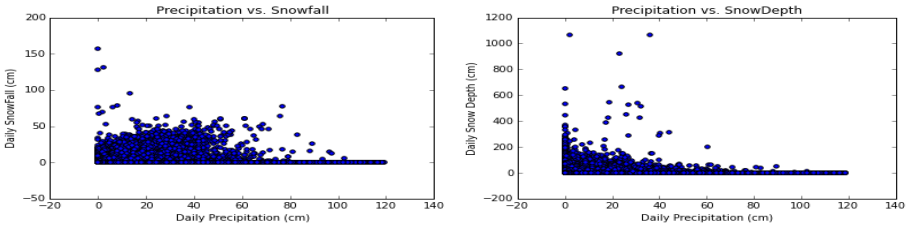


Fig. 4: Time span of data available for each capital city in US

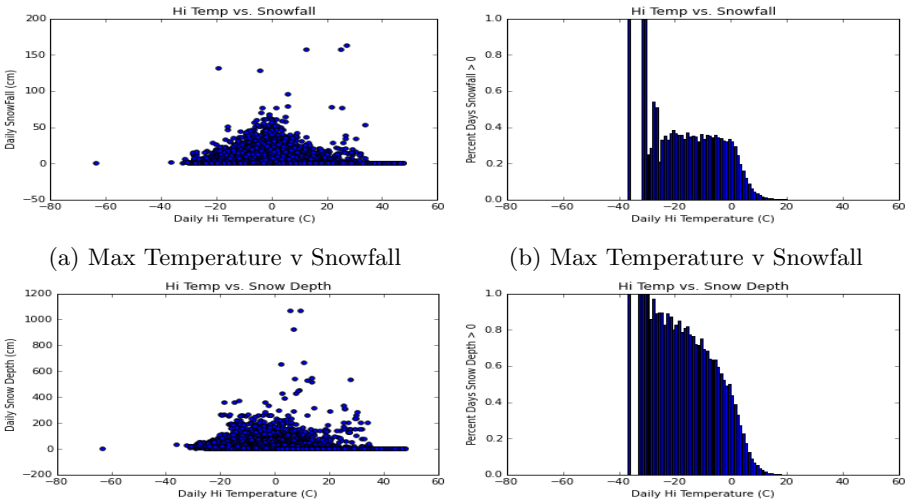
Now, we analyzed how weather variables like precipitation, minimum temperature and maximum temperature varied with snowfall and snow depth.



(a) Precipitation v Snowfall (b) Precipitation v Snow Depth

Fig. 5: Relation between Precipitation and Snow

From figure 5b we can see that as the daily precipitation increases, the snow depth decreases. This is because precipitation indicates that the temperature is not suitable for snow. Precipitation causes snow to melt faster as well.

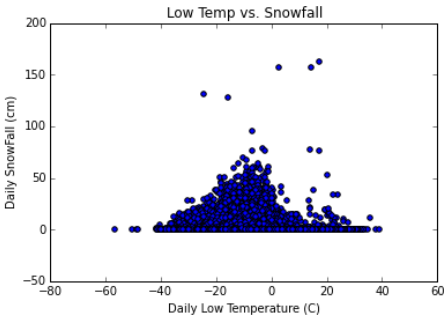


(a) Max Temperature v Snowfall (b) Max Temperature v Snowfall

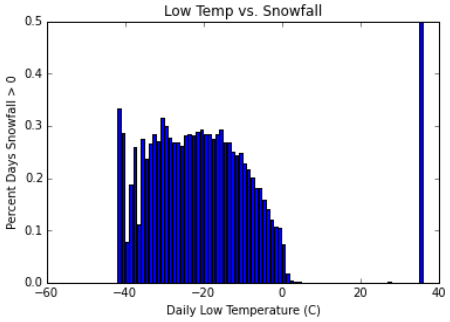
(c) Max Temperature v Snow Depth (d) Max Temperature v Snow Depth

Fig. 6: Relation between Maximum Temperature and Snow

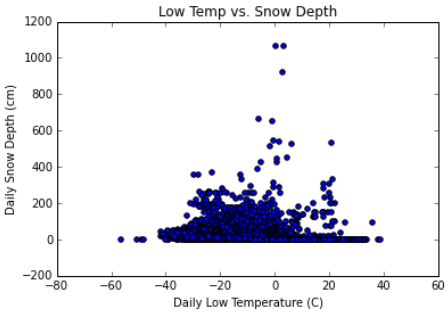
From figure 6, we can see that snow has an optimal maximum temperature range. As the temperatures increase beyond this range and fall below this range, snowfall and snow depth decrease.



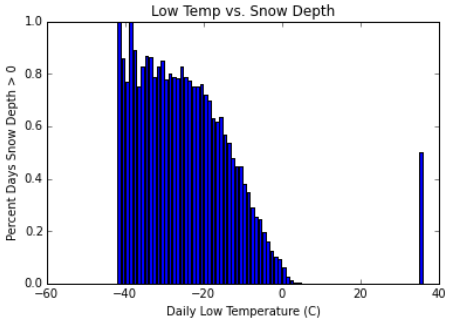
(a) Min Temperature v Snowfall



(b) Min Temperature v Snowfall



(c) Min Temperature v Snow Depth



(d) Min Temperature v Snow Depth

Fig. 7: Relation between Minimum Temperature and Snow

As is the case of maximum temperature, from figure 7, we can see that snow has an optimal minimum temperature range. As the temperatures increase beyond this range and fall below this range, snowfall and snow depth decrease.

6 Evaluation Environment

To evaluate the models we developed for White Christmas Prediction, we used a bunch of statistical measures. Before proceeding further, let us see what these terms mean.

– **Mean Squared Error :** Mean Squared Error is the mean of the squares of the difference between the values of the actual observations and the values predicted by a model. [12] The smaller the Mean Squared Error, the better the predictions of a model.

- **Root Mean Squared Error :** Root Mean Squared Error is the square root of the mean of the squares of the difference between the values of the actual observations and the values predicted by a model. [13] The smaller the Root Mean Squared Error, the better the predictions of a model.
- **True Positive :** In the case of a White Christmas prediction, a true positive is when we predict a White Christmas and a White Christmas occurs.
- **True Negative :** In the case of a White Christmas prediction, a true negative is when we don't predict a White Christmas and a White Christmas doesn't occur.
- **False Positive :** In the case of a White Christmas prediction, a false positive is when we predict a White Christmas and a White Christmas doesn't occur.
- **False Negative :** In the case of a White Christmas prediction, a false negative is when we don't predict a White Christmas and a White Christmas occurs.
- **Precision :** In the case of our White Christmas prediction model, precision is the ratio of true positives and White Christmas predictions.
- **Recall :** In the case of our White Christmas prediction model, recall is the ratio of true positives and White Christmas occurrences.
- **F-score :** The F-score is the weighted average of the precision and recall. [14] A F-score of 1 means the prediction model is doing a perfect job.

We develop a reusable evaluation environment that evaluates the predictions our models make and outputs values for the above defined statistical measures. Based on these values, we can see how well our model predicts.

7 Baseline Model and Extended Baseline Model

In this section, we start off by explaining our baseline model along with its evaluation and prediction results. Then, we explain our extended baseline model along with its evaluation and prediction results.

7.1 Baseline Model

For our baseline model, we first categorized the past 100 Christmases as a white Christmas or non-snowy Christmas. Then we averaged the past 25, 50, 75, and 100 years to come up with four preliminary values. Keeping in mind the fact that climate trends may have changed in recent decades due to factors like global warming, we weighed the recent data more heavily than the older data. As a result, we created a weighted average of these four values. Finally, to eliminate predictions of zero percent, we added a small fraction (one over the number of non-na values in the past 100 years) to come to our final prediction.

7.2 Evaluation of Baseline Model

On evaluating our baseline model with our evaluation technique, we get the following values for our statistical terms [Figure 8] -

Measure	Value
Mean Squared Error	0.0969320712109
Root Mean Squared Error	0.31133915785
True Positives	7
True Negatives	142
False Positives	8
False Negatives	17
Precision	0.466666666667
Recall	0.291666666667
F-score	0.358974358974

Fig. 8: Baseline Model Evaluation

We obtained a F-score of 0.35 which is far away from the ideal value of 1. This means that there is scope for making better predictions.

For better visualization of where we went wrong, we calculated the average absolute prediction error and plotted the data on a data-map. [Figure 9] As the color gets darker, the average absolute prediction error gets higher. Our baseline model does well in the case where the probability of white Christmas is 0 most of the times and where the amount of snowfall is large. It struggles a bit in places where snowfall doesn't belong to either extremes.

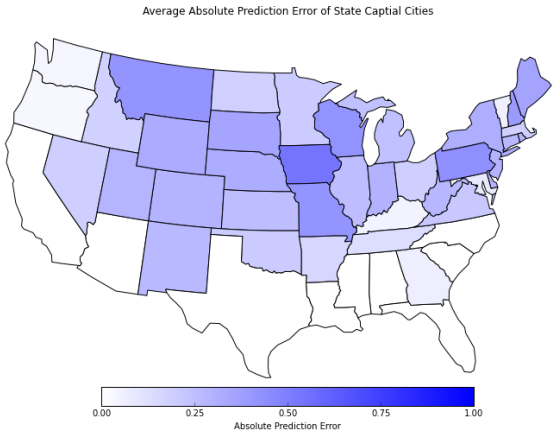


Fig. 9: Average Absolute Prediction Error of State Capitals - Baseline Model

7.3 Prediction and Analysis of Baseline Model

See Table 1 for a table of our baseline predictions.

Capital City	State	Baseline	Avg 25 Yrs	Avg 50 Yrs	Avg 75 Yrs	Avg 100 Yrs
MONTGOMERY	AL	0.012	0.000	0.000	0.000	0.000
JUNEAU	AK	0.548	0.440	0.540	0.544	0.551
PHOENIX	AZ	0.015	0.000	0.000	0.000	0.000
LITTLE ROCK	AR	0.054	0.080	0.040	0.040	0.040
SACRAMENTO	CA	0.017	0.000	0.000	0.000	0.000
DENVER	CO	0.291	0.400	0.340	0.253	0.237
HARTFORD	CT	0.296	0.143	0.261	0.314	0.306
DOVER	DE	0.110	0.095	0.087	0.097	0.095
TALLAHASSEE	FL	0.018	0.000	0.000	0.000	0.000
ATLANTA	GA	0.012	0.000	0.000	0.000	0.000
HONOLULU	HI	0.016	0.000	0.000	0.000	0.000
BOISE	ID	0.255	0.280	0.280	0.203	0.244
SPRINGFIELD	IL	0.234	0.280	0.220	0.213	0.220
INDIANAPOLIS	IN	0.211	0.280	0.160	0.194	0.194
DES MOINES	IA	0.377	0.400	0.340	0.364	0.364
TOPEKA	KS	0.190	0.240	0.160	0.189	0.165
FRANKFORT	KY	0.060	0.043	0.022	0.058	0.057
BATON ROUGE	LA	0.012	0.000	0.000	0.000	0.000
AUGUSTA	ME	0.653	0.520	0.660	0.652	0.652
ANNAPOLIS	MD	0.096	0.100	0.073	0.070	0.085
BOSTON	MA	0.273	0.222	0.279	0.279	0.250
LANSING	MI	0.620	0.625	0.633	0.591	0.591
SAINT PAUL	MN	0.678	0.760	0.680	0.640	0.649
JACKSON	MS	0.012	0.000	0.000	0.000	0.000
JEFFERSON CITY	MO	0.075	0.050	0.024	0.060	0.089
HELENA	MT	0.439	0.364	0.404	0.431	0.454
LINCOLN	NE	0.249	0.200	0.220	0.267	0.237
CARSON CITY	NV	0.165	0.095	0.154	0.155	0.155
CONCORD	NH	0.591	0.458	0.592	0.592	0.592
TRENTON	NJ	0.162	0.120	0.120	0.173	0.160
SANTA FE	NM	0.228	0.217	0.233	0.206	0.203
ALBANY	NY	0.350	0.200	0.380	0.360	0.337
RALEIGH	NC	0.011	0.000	0.000	0.000	0.000
BISMARCK	ND	0.629	0.560	0.660	0.606	0.606
COLUMBUS	OH	0.158	0.200	0.140	0.160	0.130
OKLAHOMA CITY	OK	0.054	0.080	0.040	0.030	0.030
SALEM	OR	0.045	0.040	0.040	0.027	0.034
HARRISBURG	PA	0.214	0.200	0.180	0.227	0.200
PROVIDENCE	RI	0.152	0.095	0.130	0.145	0.145
COLUMBIA	SC	0.012	0.000	0.000	0.000	0.000
PIERRE	SD	0.336	0.286	0.283	0.333	0.347
NASHVILLE	TN	0.070	0.042	0.041	0.062	0.062
AUSTIN	TX	0.011	0.000	0.000	0.000	0.000
SALT LAKE CITY	UT	0.423	0.480	0.460	0.400	0.372
MONTPELIER	VT	0.796	0.810	0.783	0.774	0.774
RICHMOND	VA	0.057	0.045	0.043	0.042	0.052
OLYMPIA	WA	0.050	0.059	0.024	0.017	0.049
CHARLESTON	WV	0.139	0.120	0.120	0.123	0.136
MADISON	WI	0.562	0.600	0.580	0.533	0.540
CHEYENNE	WY	0.247	0.320	0.240	0.212	0.212

Table 1: White Christmas Prediction by our Baseline Model.

To present a proper visualization of our findings, we plotted the above predictions by our baseline model in a data-map. [Figure 10] We assigned a color to each state based on the probability of a white Christmas in its capital city. As the color keeps getting darker and darker, the chances of a white Christmas are higher and higher. As expected, most of the capital cities in the northern part of United States have high probability of a White Christmas when compared to the southern part of the country.

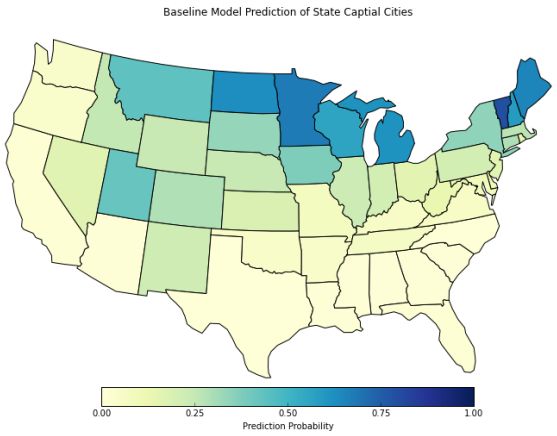


Fig. 10: Data-map showing White Christmas probability - Baseline Model

7.4 Extended Baseline Model

One major change we did from our baseline model to the extended baseline model is the use of the stations to collect data corresponding to a particular city.

In the baseline model we just used a single station per city to come up with the predictions. Unlike the baseline model, for the extended baseline model, we used the latitude and longitude data of the cities and stations and for each city we picked up stations which are within the diameter of the city. As a result of this approach, we now have around 30 stations associated with each city.

While making predictions, we weighed in factors like station with max probability, number of stations having probability greater than 0.5 and average probability of stations and came up with the actual chance of a White Christmas.

To better visualize this city and the corresponding station information, we plotted a data-map as shown in Figure 11. This data-map uses the latitude and longitude data and shows the city marked in blue and the corresponding stations

for that particular marked in red.

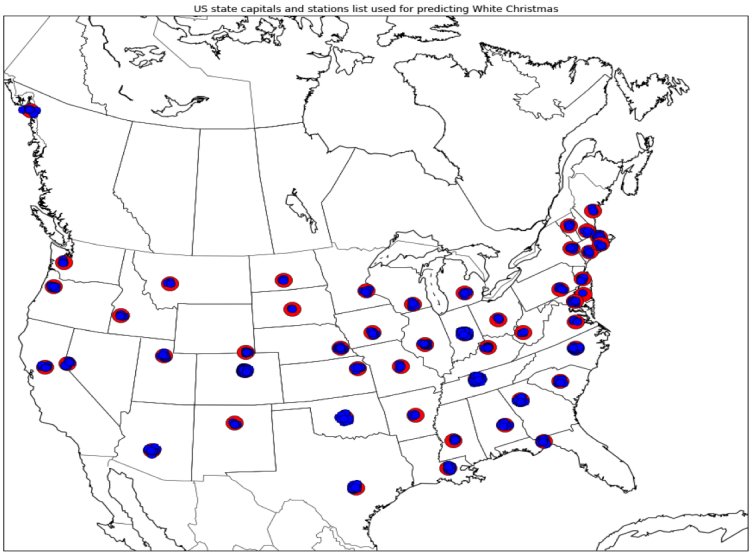


Fig. 11: Data-map showing capital cities and the stations used

7.5 Evaluation of Extended Baseline Model

We did a similar evaluation as above on our extended baseline model and got the following values for our statistical terms [Figure 12] -

Measure	Value
Mean Squared Error	0.0907195134143
Root Mean Squared Error	0.3011968018
True Positives	22
True Negatives	428
False Positives	12
False Negatives	48
Precision	0.647058823529
Recall	0.314285714286
F-score	0.423076923077

Fig. 12: Extended Baseline Model Evaluation

From the above statistics we can see that our extended baseline model is doing a better job at predicting than our baseline model. The F-score has increased from 0.35 to 0.42. Also, we can see that the precision has gone up from 0.46 to 0.64 which means that the extended baseline model is making more relevant predictions when compared to the baseline model. Given the better performance of the extended baseline model over the baseline model, we built our advanced model over the extended baseline model.

Just like in the case of the baseline model, we plot the data-map with the average absolute errors for the extended baseline model as well. [Figure 13]

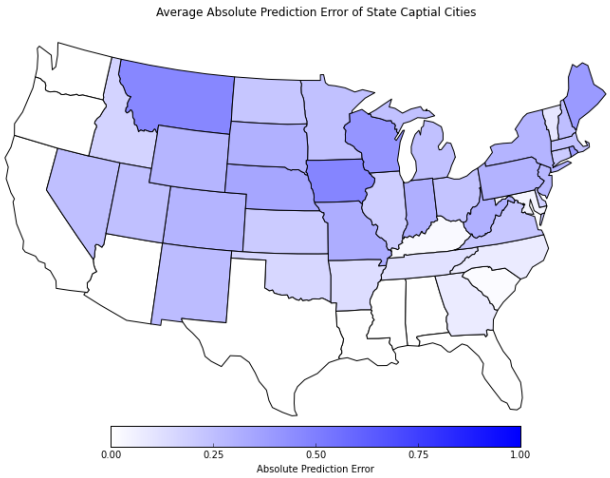


Fig. 13: Average Absolute Prediction Error of State Capitals - Extended Baseline Model

7.6 Predictions of Extended Baseline Model

Table 2 shows the chance of a White Christmas for capital city of every state in the United States -

Capital City	State	maxProbStation	avgProbOfStations	0.5StationsCount	totalStations	Chance
Montgomery	AL	0	0	0	11	0
Juneau	AK	0.8888889	0.5043556	7	12	70.70151
Phoenix	AZ	0	0	0	12	0
Little Rock	AR	0.04347826	0.02173913	0	4	3.434783
Sacramento	CA	0	0	0	12	0
Denver	CO	0.75	0.2954146	29	138	58.61332
Hartford	CT	0.2307692	0.08093645	0	5	17.95518
Dover	DE	0.04347826	0.04347826	0	1	3.608696
Tallahassee	FL	0	0	0	18	0
Atlanta	GA	0.04347826	0.002717391	0	16	3.282609
Honolulu	HI	0	0	0	38	0
Boise	ID	0.6086957	0.2235193	1	7	47.44033
Springfield	IL	0.2727273	0.1578923	0	13	21.71768
Indianapolis	IN	0.3571429	0.1319969	0	35	27.84169
Des Moines	IA	0.8571429	0.5231111	4	11	68.4706
Topeka	KS	0.4285714	0.147692	0	8	33.32439
Frankfort	KY	0.1666667	0.109127	0	12	13.37302
Baton Rouge	LA	0	0	0	25	0
Augusta	ME	0.5652174	0.5101449	2	3	46.47246
Annapolis	MD	0.1428571	0.05793651	0	8	11.17778
Boston	MA	0.4782609	0.2988651	0	9	38.26049
Lansing	MI	0.6666667	0.4890191	5	9	53.91215
Saint Paul	MN	0.8695652	0.552015	6	10	69.63351
Jackson	MS	0	0	0	9	0
Jefferson City	MO	0.1428571	0.1245342	0	5	11.71056
Helena	MT	0.4666667	0.1927536	0	4	36.54203
Lincoln	NE	0.5714286	0.1834052	1	16	44.32438
Carson City	NV	0.6521739	0.2038043	1	14	50.54348
Concord	NH	0.6	0.4167984	2	4	48.33439
Trenton	NJ	0.1666667	0.04227053	0	12	12.83816
Santa Fe	NM	0.5384615	0.1491035	1	7	41.57744
Albany	NY	0.5714286	0.3578667	1	7	45.72008
Raleigh	NC	0.05555556	0.001501502	0	37	4.178679
Bismarck	ND	0.8571429	0.4748102	3	5	68.0842
Columbus	OH	0.2608696	0.1531621	0	4	20.79051
Oklahoma City	OK	0.1666667	0.08308941	0	37	13.16472
Salem	OR	0	0	0	20	0
Harrisburg	PA	0.3333333	0.2021517	0	7	26.61721
Providence	RI	0.4347826	0.1043478	0	5	33.44348
Columbia	SC	0	0	0	13	0
Pierre	SD	0.2173913	0.1086957	0	2	17.17391
Nashville	TN	0.1666667	0.04788015	0	60	12.88304
Austin	TX	0	0	0	38	0
Salt Lake City	UT	0.9565217	0.4628635	8	18	75.44204
Montpelier	VT	1	0.680176	4	6	80.44141
Richmond	VA	0.1666667	0.1020531	0	4	13.31643
Olympia	WA	0.1666667	0.1	0	5	13.3
Charleston	WV	0.2222222	0.1516563	0	4	17.87992
Madison	WI	0.8571429	0.4423199	4	11	67.82427
Cheyenne	WY	0.5714286	0.2110686	3	27	44.54569

Table 2: Chance of White Christmas as predicted by the extended baseline model.

To present a proper visualization of our findings, we plotted the above baseline predictions in a data-map. [Figure 14] The location of each capital city is associated with a star sign and the size of the star denotes the chance of White Christmas i.e., the bigger the star, the higher the chance of White Christmas. As expected, most of the capital cities in the northern part of United States have high probability of a White Christmas when compared to the southern part of the country.

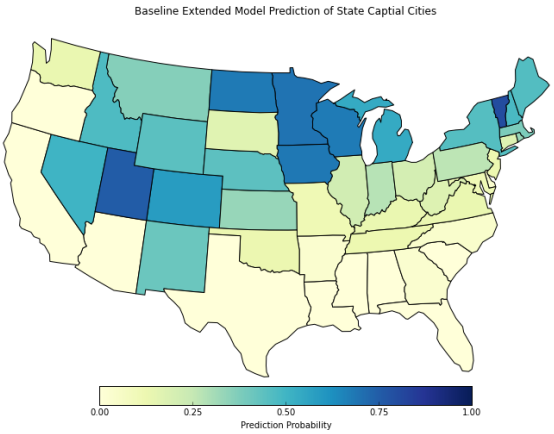


Fig. 14: Data-map showing White Christmas probability - Extended Baseline Model

8 Advanced Model

For our advanced model, we studied the correlation between various weather variables like maximum temperature, minimum temperature, precipitation, sea water temperature etc. and snowfall. Out of all these correlations, we observed that the heavy correlation between temperature and the snow depth in a given region gave the most optimal results (the highest F-score). This observation also complements our earlier observation of the high correlation between temperature and the snow depth in a given region.

With the regression models discussed earlier, we can predict the snow depth on Christmas day given the average temperature of November. We integrate this regression model with our extended baseline model. This way we gather data from the multiple stations that span across the diameter of the city and make predictions using temperature correlation model.

8.1 Evaluation of Advanced Model

We did a similar evaluation as above on our extended baseline model and got the following values for our statistical terms [Figure 15] -

Measure	Value
Mean Squared Error	0.19156715476
Root Mean Squared Error	0.437683852524
True Positives	176
True Negatives	393
False Positives	87
False Negatives	70
Precision	0.669201
Recall	0.71544
F-score	0.6915552

Fig. 15: Advanced Model Evaluation

From the above statistics we can see that our advanced model is doing a better job at predicting than our baseline model and extended baseline model. The F-score has increased to almost 0.7. Also, we can see a slight increase in the precision and recall values.

Just like in the case of the baseline model, we plot the data-map with the average absolute errors for the extended baseline model as well. [Figure 16]

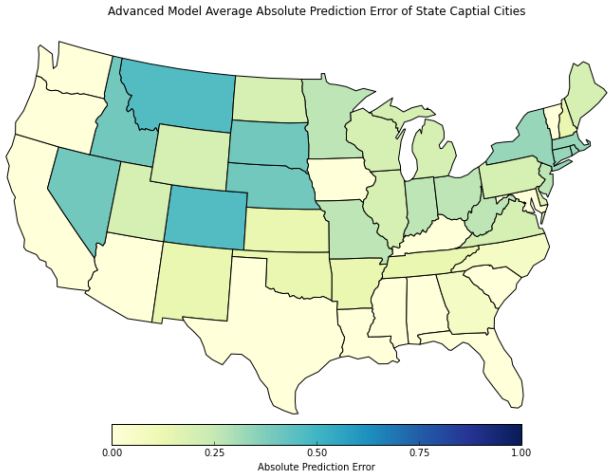


Fig. 16: Average Absolute Prediction Error of State Capitals - Advanced Model

8.2 Predictions of Advanced Model

Using our advanced model, we made predictions for some global cities as well of the state capitals of United States of America.

While choosing the global cities, one criteria we used is that these has to be atleast five instances of snowfall in that particular city. Table 3 shows the chance of white Christmas across some of the global cities -

City	maxProbStation	avgProbOfStations	0.5StationsCount	totalStations	Chance
Moscow	1.0	1.0	1.0	1.0	100.0
Berlin	0.2173913	0.15584197	0.0	4.0	15.40079418
Seoul	0.05555555	0.05555555	0.0	1.0	4.722222222
Kiev	0.5833334	0.58333334	1.0	1.0	64.58333333
Ottawa	0.86956521	0.63089093	8.0	10.0	61.97932928
Tashkent	0.14285714	0.14285714	0.0	1.0	12.14285714
Stockholm	0.42857142	0.123552012	0.0	27.0	21.1776006
Tokyo	0.0	0.0	0.0	1.0	0.0
Oslo	0.8695652	0.4879917	5.0	10.0	54.83436853
Astana	0.40909	0.40909	0.0	1.0	34.77272727
Tallinn	0.590909	0.590909	1.0	1.0	65.22727273
Riga	0.5625	0.5625	1.0	1.0	62.8125
Vilnius	0.58333334	0.58333334	1.0	1.0	64.58333333
Amsterdam	0.25	0.10648084	0.0	27.0	14.07404212
Brussels	0.052631	0.052631	0.0	1.0	4.473684211
Bishkek	0.3125	0.3125	0.0	1.0	26.5625
Dushanbe	0.047619	0.047619	0.0	1.0	4.047619048
Vienna	0.15	0.15	0.0	1.0	12.75
Vaduz	0.83333334	0.833333334	1.0	1.0	85.83333333
Ashgabat	0.0	0.0	0.0	1.0	0.0
Luxembourg City	0.090909	0.090909	0.0	1.0	7.727272727
Sarajevo	1.0	0.766666	2.0	2.0	88.33333333
Chisinau	0.47826	0.47826	0.0	1.0	40.65217391
Tbilisi	0.0	0.0	0.0	1.0	0.0
Zurich	0.13043478	0.13043478	0.0	1.0	11.08695652
Washington, D.C.	0.1875	0.0635507	0.0	35.0	9.740035284

Table 3: Chance of White Christmas as predicted by advanced model.

Table 4 shows the chance of a White Christmas for capital city of every state in the United States as predicted by our advanced model -

Capital City	State	maxProbStation	avgProbOfStations	0.5StationsCount	totalStations	Temp Correlation	Chance
Montgomery	AL	0	0	0	5	0	0
Juneau	AK	0.8888889	0.5218933	6	10	0.4387894	59.26164
Phoenix	AZ	0	0	0	4	0	0
Little Rock	AR	0.04347826	0.02173913	0	4	0.02173913	2.608696
Sacramento	CA	0	0	0	9	0	0
Denver	CO	0.75	0.3042407	23	104	0.6795036	52.41051
Hartford	CT	0.173913	0.07246377	0	3	0.07246377	9.42029
Dover	DE	0.04347826	0.04347826	0	1	0.04347826	4.130435
Tallahassee	FL	0	0	0	15	0	0
Atlanta	GA	0.04347826	0.00310559	0	14	0.00310559	1.304348
Honolulu	HI	0	0	0	2	0	0
Boise	ID	0.6086957	0.2676329	1	5	0.2676329	34.95169
Springfield	IL	0.2608696	0.1520069	0	10	0.1520069	17.16222
Indianapolis	IN	0.3333333	0.1265343	0	27	0.1265343	17.19074
Des Moines	IA	0.8571429	0.5231111	4	11	0.7325762	66.14849
Topeka	KS	0.4285714	0.1726708	0	5	0.1726708	22.80124
Frankfort	KY	0.1666667	0.1309524	0	10	0.1309524	13.33333
Baton Rouge	LA	0	0	0	14	0	0
Augusta	ME	0.5652174	0.5652174	2	2	0.4936889	56.5498
Annapolis	MD	0.1428571	0.05668934	0	7	0.05668934	7.539683
Boston	MA	0.4782609	0.2771739	0	8	0.2771739	31.3587
Lansing	MI	0.6666667	0.5173913	3	5	0.5364559	56.456
Saint Paul	MN	0.8695652	0.5233521	4	8	0.9082486	72.42067
Jackson	MS	0	0	0	9	0	0
Jefferson City	MO	0.1428571	0.1199534	0	4	0.1199534	11.96817
Helena	MT	0.3043478	0.1521739	0	2	0.1521739	18.26087
Lincoln	NE	0.5714286	0.2018634	1	10	0.2018634	28.91615
Carson City	NV	0.6521739	0.2065217	1	12	0.2065217	31.17754
Concord	NH	0.3913043	0.2608696	0	2	0.2608696	28.04348
Trenton	NJ	0.1666667	0.04227053	0	12	0.04227053	7.125604
Santa Fe	NM	0.5384615	0.1076923	1	5	0.1076923	22
Albany	NY	0.5714286	0.3674948	1	4	0.4138804	42.65192
Raleigh	NC	0	0	0	26	0	0
Bismarck	ND	0.8571429	0.3990683	2	4	1	69.8913
Columbus	OH	0.2608696	0.173913	0	3	0.173913	18.69565
Oklahoma City	OK	0.1666667	0.08266997	0	29	0.08266997	9.953565
Salem	OR	0	0	0	20	0	0
Harrisburg	PA	0.3333333	0.2080659	0	6	0.2080659	22.89795
Providence	RI	0.4347826	0.1043478	0	5	0.1043478	18.17391
Columbia	SC	0	0	0	10	0	0
Pierre	SD	0.2173913	0.2173913	0	1	0.2173913	20.65217
Nashville	TN	0.1666667	0.05130467	0	50	0.05130467	7.757993
Austin	TX	0	0	0	34	0	0
Salt Lake City	UT	0.9565217	0.4464336	8	17	0.9065311	71.31926
Montpelier	VT	0.8571429	0.6162112	3	5	0.5664868	65.823
Richmond	VA	0.1666667	0.09903382	0	3	0.09903382	11.09903
Olympia	WA	0.1666667	0.1	0	5	0.1	11.16667
Charleston	WV	0.1428571	0.1281343	0	3	0.1281343	12.54083
Madison	WI	0.8571429	0.4423199	4	11	0.7556972	63.61046
Cheyenne	WY	0.5714286	0.2027031	3	26	0.2027031	29.05185

Table 4: Chance of White Christmas as predicted by the advanced model.

To present a proper visualization of our findings, we plotted the above predictions in a data-map. [Figure 17] The location of each capital city is associated with a star sign and the size of the star denotes the chance of White Christmas i.e., the bigger the star, the higher the chance of White Christmas. As expected, most of the capital cities in the northern part of United States have high probability of a White Christmas when compared to the southern part of the country.

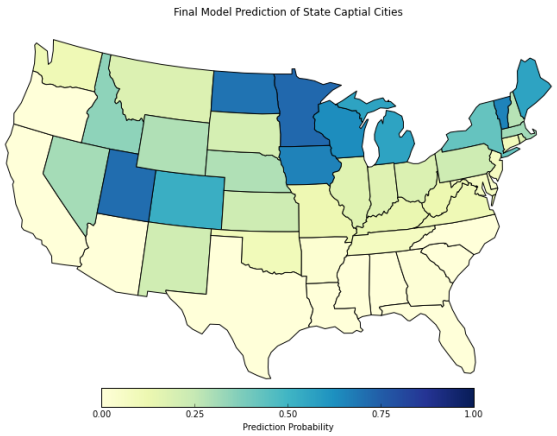


Fig. 17: Data-map showing White Christmas probability - Advanced Model

8.3 Comparing Advanced Model with the Baseline Models

In this section we compare the predictions made by our advanced model with that of the baseline model and the extended baseline model.

Figure 18 shows the predictions made by the baseline model and the advanced model for the United States state capitals. Ignoring the cities which have 0 percent chance as predicted by the advanced model, the advanced model predicts higher than the baseline models for 24 cities and lower than the baseline model for 15 cities. Interestingly, for all the north east capital cities except Boston, the baseline model predicts a higher chance of white Christmas than the advanced model.

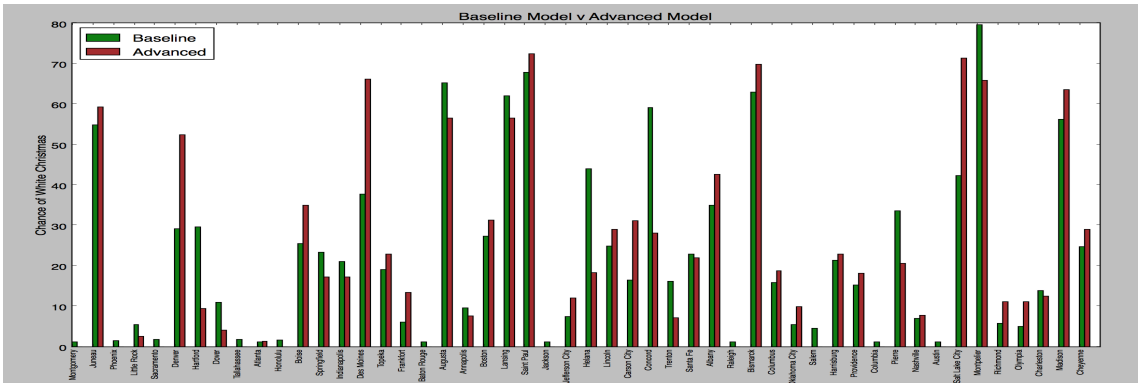


Fig. 18: Comparison between predictions of Baseline model and Advanced Model

Figure 19 shows the predictions made by the extended baseline model and the advanced model for the United States state capitals. Ignoring the cities which have 0 percent chance as predicted by the advanced model, the advanced model predicts lower than the extended baseline models for most of the cities.

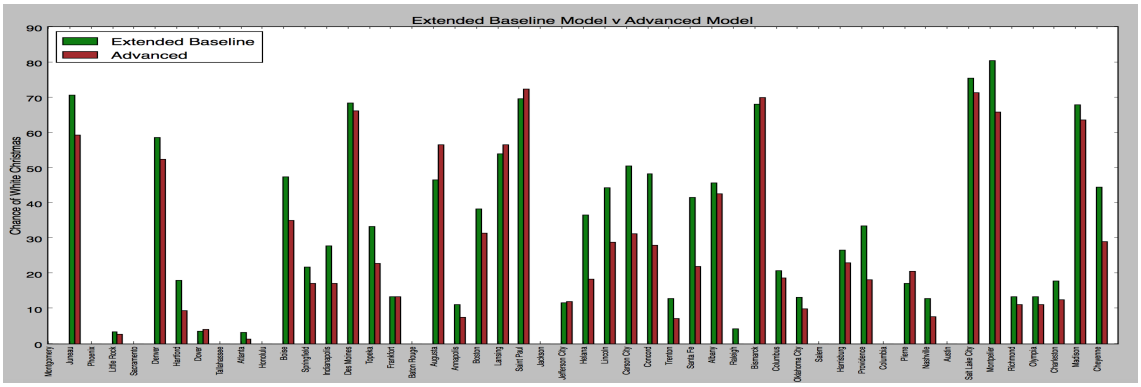


Fig. 19: Comparison between Extended Baseline model and Advanced Model

8.4 Comparing Advanced Model Predictions with NOAA Predictions

Figure 20 shows the historic probability of there being at least 1-inch of snow on the ground on December 25 based on the latest (1981-2010) U.S. Climate Normals from NOAA's National Climatic Data Center. Dark gray shows places where the probability is less than 10 percent, while white shows probabilities greater than 90 percent. [15]

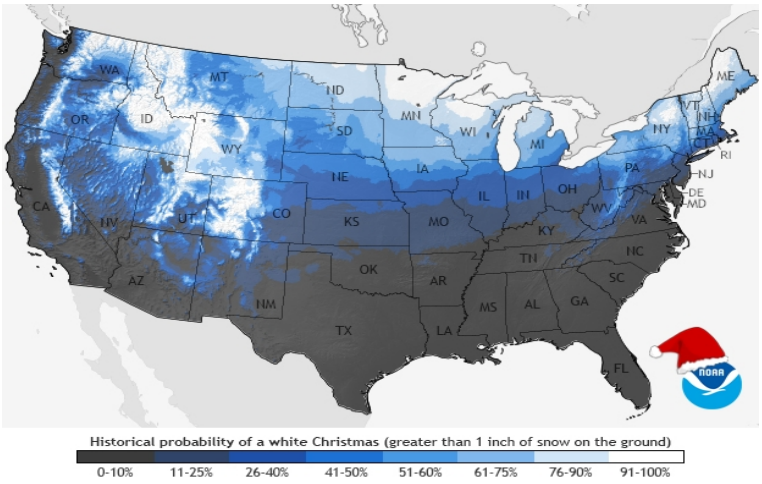


Fig. 20: Data-map showing White Christmas probability - NOAA Analysis

Using the above information and the predictions made by our advanced model, we calculated the absolute difference between the predictions made by both the models. We also went ahead and plotted this information on a datamap. [Figure 21]

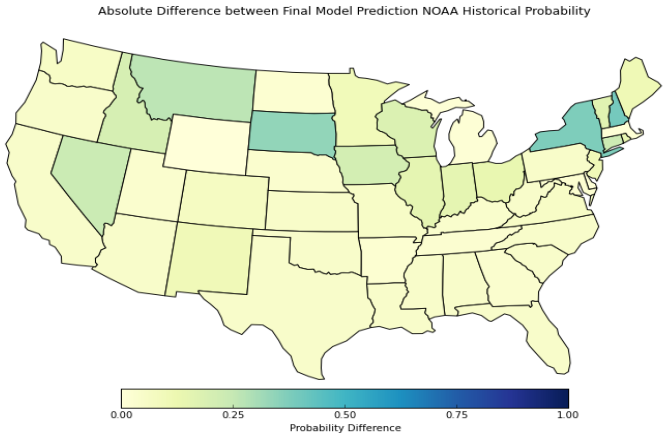


Fig. 21: Difference between Advanced Model and NOAA Predictions

As it can be seen from the figure 21, most of the states are in the 0 - 25 percent error range showing that the predictions we made using our advanced

model are relatively close to the predictions NOAA made. The three states where both the models differ a lot are South Dakota, New York and New Hampshire.

9 Final Prediction and Conclusion

Now that we have predicted the chance of a white Christmas in each of the state capital in United States, we do a probabilistic analysis to come up with a statement like "It is going to be a white Christmas in X state capitals in United States."

This can be done in two ways -

- **Independent Probabilistic Analysis** wherein we consider the cities to be independent of one another when it comes to snowfall.
- **Correlated Probabilistic Analysis** wherein we consider a correlation between the cities when it comes to snowfall.

In the next sections, we take a closer look at each of these probabilistic analysis.

9.1 Independent Probabilistic Analysis

In the independent probabilistic analysis, we treat the probability of a white Christmas in a particular city as an event independent of the probability of a white Christmas in other cities. For this, we adopt a recursive probabilistic analysis defined below.

Let $f(m, k)$ be the probability of it snowing in exactly k cities out of m cities and p_k be the probability of it snowing in the k^{th} city.

We can write a recursive equation for $f(m, k)$ given by -

$f(m, k) = p_m$ * $f(m-1, k-1)$ + (1 - p_m) * $f(m-1, k)$

Using the above equation we calculate the probability of it snowing in exactly zero cities, one city and so on. newline

The following is the probability distribution obtained. From figure 22, we can say that - "There is a good chance of a white Christmas in 8-15 state capitals in United States. The most likeliest occurrence is that it is going to be a white Christmas in 11 state capitals in United States."

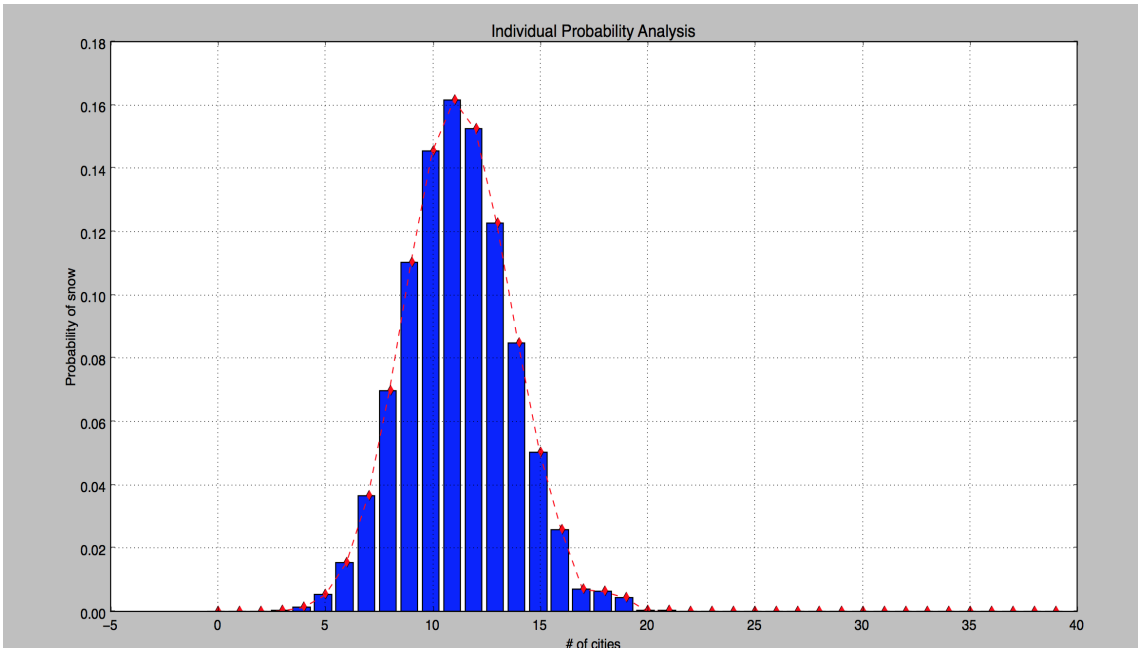


Fig. 22: Individual Probability Analysis

Next, we introduce the concept of correlation between cities into this independent probabilistic analysis.

9.2 Correlated Probabilistic Analysis

Although the independent probabilistic analysis does a decent job at predicting the number of occurrences of white Christmas, it fails to capture the snowfall and snow depth correlation between the cities based not only on the proximity of the cities but also on the weather patterns in the cities. It is for this reason we switch to a more sophisticated correlated probabilistic analysis.

So, to do a correlated probability analysis, we first calculate the pairwise correlation between snow depths in various state capitals.

Figure 23 shows the snow depth correlation between all the state capitals in the United States of America.

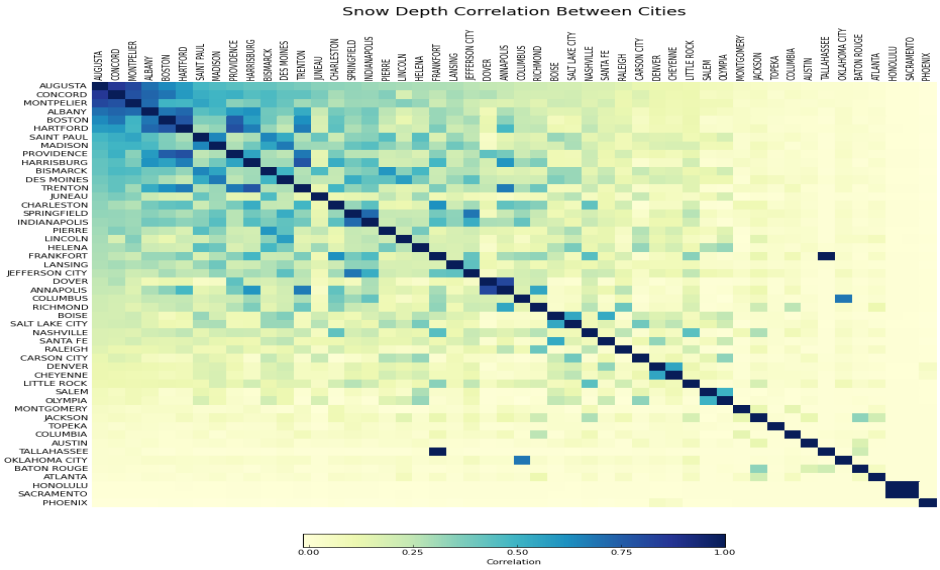


Fig. 23: Snow Depth Correlation

Now that we have all the pairwise correlations, we see which cities are strongly correlated to the 11 probable cities the independent probabilistic model predicted.

Table 5 shows the correlations we took into picture

Initial City	First Degree Correlation	Second Degree Correlation
Juneau		
Des Moines	Lincoln	
Augusta	Concord	
Lansing		
Saint Paul	Pierre, Helena	
Albany	Hartford	Providence
Bismarck	Pierre, Helena	
Salt Lake City	Boise, Carson City, Helena	
Madison		
Denver	Cheyenne	
Montpelier	Boston	Providence

Table 5: Cities correlated by snow depth.

In addition to the 11 initial cities, we now have 10 additional cities. So we concluded that - “There is a good chance of a white Christmas in 21 state capitals in United States.”

10 Bibliography

References

1. Patrick Young, http://patricklyoung.net/
2. White Christmas, http://en.wikipedia.org/wiki/White_Christmas_(weather)
3. Weather Forecasting, http://en.wikipedia.org/wiki/Weather_forecasting
4. Weather Prediction Models, http://ww2010.atmos.uiuc.edu/(Gh)/guides/mtr/fcst/mth/prst.rxml
5. Groisman, Pavel Ya, and David R. Easterling. "Variability and trends of total precipitation and snowfall over the United States and Canada." Journal of Climate 7.1 (1994): 184-205.
6. Brasnett, Bruce. "A global analysis of snow depth for numerical weather prediction." Journal of Applied Meteorology 38.6 (1999): 726-740.
7. Karl, Thomas R., et al. "Recent variations of snow cover and snowfall in North America and their relation to precipitation and temperature variations." Journal of Climate 6.7 (1993): 1327-1344.
8. Bednorz, Ewa. "Snow cover in eastern Europe in relation to temperature, precipitation and circulation." International Journal of Climatology 24.5 (2004): 591-601.
9. Murphy, Allan H. "What is a good forecast? An essay on the nature of goodness in weather forecasting." Weather and forecasting 8.2 (1993): 281-293.
10. National Oceanic and Atmospheric Administration, http://www.noaa.gov/about-noaa.html
11. Global Historical Climatology Network - Daily, http://www.ncdc.gov/oa/climate/ghcn-daily/
12. Mean Squared Error, http://en.wikipedia.org/wiki/Mean_squared_error
13. Root Mean Squared Error, http://en.wikipedia.org/wiki/Root-mean-square_deviation
14. F Score, http://en.wikipedia.org/wiki/F1_score
15. Are you dreaming of a white Christmas?, http://www.climate.gov/news-features/featured-images/are-you-dreaming-white-christmas