Can Data Analysis Techniques be used for accurate short term weather prediction using past data

Table of Contents

[Abstract 3](#_Toc515267028)

[Introduction 4](#_Toc515267029)

[Literature Review 5](#_Toc515267030)

[Existing Data Mining Studies 5](#_Toc515267031)

[Current Numerical Weather Prediction Systems 9](#_Toc515267032)

[Numerical Weather Prediction Computing Power 11](#_Toc515267033)

# Abstract

# Introduction

Accurate weather prediction has the potential to greatly influence a number of industries with agriculture being one of them. Up until recently, numerical weather prediction would have been considered a task for supercomputers, and to a certain extent, still is. With recent advances in computing power and algorithm development, this task can now be attempted without the excessive hardware requirements.

Agriculture is just one industries that can benefit from these advances. (Sivakumar, 2006) outlines that accurate long term weather prediction can be used to mitigate risk in agriculture by helping predict the success or failure of an agricultural season. This would also lead to economic benefits, where inputs could be restricted in the event that the long term weather forecast is not agreeable. Similarly, accurate weather prediction could also be used to improve crop establishment and overall yields by utilising forecasts to determine optimal planting and harvesting times.

Precise weather prediction using predictive analytics would also assist local authorities in informing the population about incoming extreme weather events, potentially saving lives and livelihoods. (Huang and Ran, 2003) outline a traffic speed prediction model based on a neural network that determines the optimal speed under certain adverse circumstances such as severe weather events.

In addition, the ability to accurately predict certain weather phenomenon would greatly impact particular industries. Given a dataset with the relevant information, weather prediction could be used to estimate the total hours of sunshine per day. This information could then be put to use in the solar energy industry for more efficient and cost-effective energy generation. These examples epitomise the range of applications that accurate weather prediction can influence.

Therefore the aim of this research is to determine if data analysis techniques can be used to accurately predict short term weather forecasts using past data. To implement this research, past data is programmatically retrieved from Met Éireann’s website to be analysed and used as a basis for the predictions.

# Literature Review

## Existing Data Mining Studies

There are multiple existing papers that engage in the task of weather prediction, each with their own unique outlook on the problem. (Talib et al., 2017) use the J48 and decision tree algorithm to perform an analysis on weather data from 2007 to 2016. Unlike many other studies that use machine learning algorithms to predict weather values or the occurrence of particular weather events, the authors instead determine association rules for the weather i.e. under what circumstances particular weather events will occur.

(Sharma et al., 2014) use a combination of the Density Based Spatial Clustering of Applications with Noise (DBSCAN) and K-Nearest Neighbour (KNN) algorithms to cluster similar data points then assign each cluster to a specific weather class. The authors state that their system will predict the occurrence of fog, rain and snow to within 90%, 67% and above 93% accuracy respectively. Although this is a relatively successful system, it is worth nothing that this system does not quantify the weather event, it simply classifies its occurrence.

(Kalyankar and Alaspurkar, 2013) use the K-means clustering algorithm to create clusters of weather data which they then perform an analysis on for knowledge discovery. For the purpose of knowledge discovery any clustering algorithm should be appropriate but using the DBSCAN clustering algorithm has two distinct benefits; the number of clusters does not need to be supplied to the algorithm and it also determines and marks outliers in the data meaning it can be used for outlier analysis.

(Olaiya and Adeyemo, 2012) apply two forms of neural networks and a decision tree on data spanning ten years to predict a combination of weather phenomenon such as maximum temperature, rainfall, evaporation and wind speed. In summary the authors used the decision tree to determine association rules resembling the work of (Talib et al., 2017). Following this they implement a time lagged feed forward neural network and recurrent neural network to make predictions. Overall the time lagged feed forward neural network performed better with an error of ~24%.

(Jan et al., 2008) perform seasonal climate prediction using the KNN algorithm based on ten years of past data. In this scenario, the data consisted of seventeen features based on ten locations. The authors found that when predicting a Boolean attribute, such as the presence of fog or snow, they could achieve accuracies of 96.6% and greater.

(Petre, n.d.) uses a decision tree for temperature prediction modelled as a classification problem where the output temperatures are transformed into certain ranges determined by the author. The author uses data collected from 2002 to 2005 for Hong Kong. The model is evaluated under numerous classification metrics for each of the temperature range classes. The system resulted in training accuracies of 83.33%. Unfortunately the author did not evaluate the trained model on an independent test set meaning the true significance of the model can not be determined.

(Al­Roby and Alaa M, 2011) performed numerous data mining techniques to determine wind speed, which was again treated as a classification problem. The authors used ten years of historical daily data for use in their case study. The authors perform some interesting transformations of the data so each observation contains the windspeed values for the previous two days. Following this the target column of windspeed is discretized. The authors approach the problem using multiple techniques such as association rule mining, classification and clustering. In terms of classifying future wind speeds, the authors use two algorithms KNN and a feed forward neural network. The authors note that KNN and the feed forward neural network reach 62.70% and 67.37% accuracy respectively.

Similarly (Kohail and El-Halees, 2011) perform numerous data mining techniques on weather data from 1977 to 1985. The techniques performed include outlier analysis, clustering, numerical prediction, classification and association rule mining. The authors perform an interesting operation called linear interpolation which is used to fill in missing values between a known amount of data points by fitting a polynomial curve to the data. This operation is often used to fill in missing data in time series problems. Like (Al­Roby and Alaa M, 2011), the authors create three new variables in the dataset that represent the previous three days temperatures, for each observation. After performing an outlier analysis, the results indicate that the outliers contain both real and input error observations. Instead of removing the incorrect observations only, the authors decided to remove all outliers.

In terms of prediction (Kohail and El-Halees, 2011) use an Artificial Neural Network (ANN) and least median squares linear regression to predict daily average temperature. This results in the ANN having a lower correlation coefficient between the actual and predicated temperatures. The authors also perform daily temperature prediction from a classification perspective using four models. Like previous studies, the data is discretized so temperatures are classed as cold, warm or hot. The algorithms used are Naïve Bayes, KNN, decision trees and an ANN. The individual accuracies range from 81.40% to 85.77% where the best accuracy was produced by the ANN.

(PAL et al., n.d.) outline how they used a back-propagated ANN to predict minimum and maximum ground level temperatures. The authors perform some basic feature engineering by shifting the data so it includes measurements from the previous two days like many of the previous authors have already done. By doing this, the ANN should be able to look at the effects of previous weather events on each observation. The authors found that the optimal ANN had an error rate of 2 °C 80% of the time.

(Nagalakshmi et al., 2013) provide a description of existing papers and provide some recommendations for future works. The authors note that a radial based function network was the best form of ANN but they also say to get the best results overall an ensemble approach should be used. Interestingly meteorologists use an ensemble approach when determining the optimal forecast using their existing statistical numerical weather prediction methods (Flynn, n.d.).

(De and Debnath, 2009) use three back-propagated ANN’s to predict minimum and maximum temperature in the months of June, July and August. The ANN’s were trained on data from the months of December to May. The authors note that the ANN for predicting Augusts temperatures was very accurate with a prediction error of approximately 5%.

(Maqsood et al., 2004) use an ensemble approach to predict weather forecasts 24 hours ahead. The predicted weather forecast consists of temperature, wind speed and humidity. The authors ensemble consists of a multi-layered perceptron (MLP), Elman recurrent neural network (ERNN), radial basis function network (RBFN) and Hopfield Model (HFM). The authors train the models on hourly data based on all four seasons but use an interesting validation strategy where they remove one extreme weather observation from each season and place them in their test set. The authors compare the models on an individual basis and as part of two ensemble methods. The fist ensemble method uses a weighted average to determine the result whereas the second ensemble model uses a winner takes all approach. RBFN was deemed to be the best individual model in terms of accuracy and training time. Regarding the ensemble approach, the authors found the winner takes all ensemble to have the smaller prediction error of the two.

(Radhika and Shashi, 2009) use support vector machines (SVM) to predict maximum atmospheric temperature 24 hours ahead. The authors compare the results of an SVM to the results achieved by a back-propagated MLP. Data spanning from 2003 to 2007 was used to train the models. Data from January to July in 2008 is treated as the test data. Unlike previous studies that have used linear interpolation to fill in missing values, the authors populate empty cells with the average temperature for that specific month. The authors found that the SVM consistently performed better than the MLP at predicting the maximum temperature for the following day.

(Wang and Sheng, 2010) compare a generalised regression neural network with a back-propagated neural network for long term rainfall prediction between the years 1955 and 2009. The authors state that one of the main challenges with rainfall prediction is its characteristic of being non-linear over time. The generalised regression model is determined as the superior of the two, as it consistently has a lower mean standard error (MSE) than the back-propagated neural network.

(Gumaste and Kadam, 2016) propose a weather prediction system using a genetic algorithm and fast Fourier transform (FFT) aimed at assisting the agricultural community. In essence the proposed system implements the genetic algorithm alongside FFT to observe previous weather events for the same day in previous years, by taking averages of these past values and comparing them to the actual outcomes.

An interesting plant monitoring system is created by (Kurniawan et al., 2017) for use in the agricultural sector. Through the use of past weather data and data obtained from sensors for soil moisture levels for example, the authors created a system that determined whether a plant would need irrigation or not. The authors implement a system which uses fuzzy logic to determine the weather. The system is tested 33 times in comparison to weather recordings from external sources. On all occasions the system is reported as being 100% accurate.

(Pandey et al., 2017) implement an adaptive neuro fuzzy inference system (ANFIS) to perform weather classification. As the system is using fuzzy logic, the problem is treated as a multi-class classification task. The authors take an unusual approach to cleaning the dataset by applying the wordcount program from Hadoop on the dataset. The ANFIS model achieves a relatively low MSE of approximately 1.42 indicating it is a reasonable model.

(Saha and Chauhan, 2017) attempt to predict low temperature, high temperature, humidity and wind speed using a non-linear autoregressive neural network. The data used in this study is based on daily observations over a time span of 45 years. As the model was evaluated using multiple metrics, the authors noted that the optimal number of hidden neurons ranged between 3 and 5 for each of the targets.

(Qu Xiaoyun et al., 2016) compare an SVM to a Long Short Term Memory (LSTM) neural network for next day wind power prediction. Principal component analysis (PCA) is used as a dimensionality reduction technique in this study which finds the directions of maximum variance within the data. Generally when PCA is used the dataset contains a lot of features. In this case the original dataset contained five features, and after the PCA transformation only two components are used as input. The authors determine that these components are related to wind speed and direction. The authors compare the results of the LSTM with raw NWP input and PCA input. The results showed that the LSTM neural network with PCA had a lower normalised RMSE. When compared with the SVM’s output, the PCA LSTM neural network performed significantly better.

(Nurunnahar et al., 2017) implement wind speed prediction using Support Vector Regression (SVR) and back-propagated ANN’s. The authors performed quite an extensive study by building multiple models for each algorithm to provide forecasts for different time intervals. A sliding window validation strategy was used which is specific to time series prediction. In principle for each training iteration the training set increases in size. As the dataset used in this study was sparse, the Leaky ReLu function was used to replace the zero’s with some values that did not disturb the existing relationship in the data. In terms of feature engineering a relatively straightforward approach was taken where the average, minimum and maximum wind speeds were included, although the timescale these features related to was not supplied. The results of this study were superb. Bearing in mind that the authors created models for 1, 3, 5, 7, 10 and 15 day ahead forecasts, SVR had accuracies ranging from 90.10% to 99.60%. In this case the 99.60% represented 3 day ahead forecasts. The back-propagated ANN’s on the other hand had accuracies ranging from 64.90% to 99.80%. In this scenario the accuracy score of 99.80% represented a 5 day ahead forecast.

(Sanusi and Corne, 2015) address the topics of feature selection and feature engineering for short term wind speed forecasting.

## Current Numerical Weather Prediction Systems

As one could imagine numerical weather prediction (NWP) is a vast and complicated field. The principle of NWP was realized by an English meteorologist, Lewis Fry Richardson (Flynn, n.d.). The system implemented by Richardson was fundamentally based on using past data, although a very small amount, to predict the air pressure in the next six hours. The fact that even the earliest attempts at NWP used past data to perform predictions highlights the need for incorporating data analysis techniques, both descriptive and predictive, into modern NWP systems. The calculations required to make a prediction took six weeks further highlighting the complexity of the task at hand.

NWP is used in order to get approximations of the atmosphere in numerical format, as the name suggests (Coiffier, 2011). The equations used to get these approximations are beyond the scope of this document but are strongly linked to fluid mechanics. A model of the atmosphere is formed using these calculations which is then discretised, so the model no longer reflects the continuous variables. This process reflects the series of actions some of the aforementioned data mining applications have taken to make weather prediction a classification problem rather than a regression problem. Although the NWP systems are still predicting a numeric value, the range of values is restricted in a sense.

In essence NWP is based on modelling the atmosphere which is treated as a fluid, hence why it is strongly linked to fluid mechanics. As NWP is considered an initial value problem, the success of a forecast as largely related to the accuracy of the initial values used to represent the atmospheres current state.

Improving weather prediction is highly correlated with improving NWP processes (Shuman, 1978). The atmosphere and its constituents are at the heart of NWP. Rainfall is possibly the most important weather factor from an individual’s perspective. Rainfall is largely related with the radiative properties of oceans and seas. As heat takes longer to dissipate from a body of water than a mass of land, evaporation can continue for longer over a body of water. This increase in evaporation over oceans leads to large quantities of clouds that become dense with water vapour. As these cloud systems are pushed in certain directions by the prevailing winds, the cloud systems make landfall. When these clouds, dense with water vapour, meet the configuration of the earth’s surface, rainfall occurs.

By utilising NWP processes to model atmospheric conditions these weather phenomena can be forecasted in advance. (Shuman, 1978) outlines that heating and friction are important for determining an atmospheric forecast which extends a week in advance. Interestingly the author gives an overview of what they perceive to be the most influential features of NWP systems; three of the five are related to advances in technology. The effectiveness of technological advances on NWP systems is further highlighted when the author states that these advances are heavily reliant on progress in computational domains.

The main analysis of this document, both descriptive and predictive, will be based on hourly data obtained from Met Éireann. Hence this section will continue by looking at the NWP processes and models employed by Met Éireann in their weather forecasting systems.

Fortunately Met Éireann provide a brief overview of their entire weather forecasting process (*How Met Éireann produces a forecast - Met Éireann - The Irish Meteorological Service*, n.d.). As is the case with all NWP systems, the first step is to collect observations of current and past weather conditions. Following this, the process enters the data assimilation stage. Numerous papers referred to in this document, such as (Shuman, 1978) , state that NWP is an initial value problem. Therefore, Met Éireann use the data assimilation stage to compare previous forecasts with the actual observations. This process of data assimilation is very similar to model training techniques in the data science world. In essence when creating a predictive model, an initial model is created with basic parameters and a subset of the data. When the model is run and evaluated under some performance metric it is often re-run with updated parameters to increase the predictive accuracy of the model which is approximately the purpose of data assimilation.

Given the complexity of NWP and its heavy reliance on computing power, its intriguing to know that forecasters in Met Éireann still adopt a manual approach when creating forecasts on certain occasions. Forecasters will sometimes draw charts by hand to determine what the weather observations are indicating (Forecasting Centre - Met Éireann - The Irish Meteorological Service, n.d.).

Met Éireann uses a combination of two forecasting models for their weather predictions, HIRLAM and HARMONIE. HIRLAM, which stands for High Resolution Local Area Model, is used to provide short term weather forecasts. This model has been developed by a number of countries as part of a consortium. A medium range forecasting model implemented by the ECMWF is used to provide weather forecasts approximately ten days in advance.

When reading about NWP, resolution is a term that surfaces more often than not. The resolution of a weather forecasting model is not necessarily an accuracy measure but an indicator of how localised a weather model can be. Met Eireann’s NWP models for example have a resolution of approximately 2.5km (Numerical Weather Prediction - Met Éireann - The Irish Meteorological Service, n.d.). This means that the model splits the country into grids of 2.5km2 and then approximates the atmospheric conditions for each of those grids. Although this resolution seems quite good from a newcomer’s point of view, it is stated that observations would need to be recorded over every couple of meters to have a completely accurate measure of the atmospheric conditions. It is also stated that as these recordings are not attainable, assumptions need to be made. This characteristic again resembles that of any machine learning problem where assumptions need to be made about the data and problem being undertaken in order to perform predictions.

Mitigating the errors introduced by making assumptions is crucial in machine learning. This is no different in terms of NWP models. To avoid the errors introduced by these assumptions an ensemble approach is used whereby numerous forecasting models are built using a different initial state of the atmosphere for each model. The idea of using an ensemble approach is again common in the machine learning and predictive analytics world whereby numerous models are built using the same (with different hyperparameter values) or different algorithms. The results of these algorithms are then averaged, using a weighted average for example, to get an overall result.

(Andersson, 2014) states that the ECMWF medium range prediction is based on an ensemble of 52 models. One of these models is at a much higher resolution than the other 51 models meaning the initialisation values used for that model are the most accurate description for the current state of the atmosphere. A second model takes the opposite approach where it uses a much lower resolution than all the rest. This results in an individual model that has a less accurate but broader outlook. The remaining models are then initialised with values within the range of the high resolution and low resolution model.

As opposed to the ensemble machine learning approach, where the results are effectively averaged, the result of ensemble weather forecasting is to provide multiple forecasts that account for the error introduced in making assumptions about the atmosphere (Jeppesen, 2017). This approach of producing multiple forecasts enables the meteorologists to refrain from making a deterministic forecast which could potentially be incorrect. The importance of ensemble prediction models is accentuated by (Barratt, 2017) whereby ensembles are considered the way forward for accurate weather prediction.

## Numerical Weather Prediction Computing Power

As outlined by (Shuman, 1978) increases in computing power is largely associated with advances in NWP. Even by today’s standards, basic CPU’s alone are not sufficient for weather forecasting. Continuous developments of GPU’s has allowed problems which are largely concerned with complex calculations to become more mainstream. The reasons that GPU’s have enabled this is their innate ability to perform linear algebraic calculations anywhere from 10 to 100 times faster than CPU’s. This, combined with their ability to parallelise operations makes them an optimal solution in any computing intensive operation.

(Michalakes and Vachharajani, 2008) demonstrate the power of GPU’s by running a computationally intensive NWP program on multiple GPU’s and provide and analysis of the results in comparison to a traditional high-powered CPU based system. The authors experiment was run on a cluster containing 64 CPU-GPU pairs. The authors found that incorporating GPU’s into a system decreased computation time by orders of magnitude. Interestingly the NWP models used by Met Éireann are run on supercomputers that contain multiple high powered GPU’s further highlighting the scale of NWP problems.

(Coiffier, 2011) outlines the scale of the calculations and computational requirements to make a 24 hour forecast using the ARPEGE model that is used by the French. The computer used in this system is powerful enough to create a 24 hour forecast in under 15 minutes.