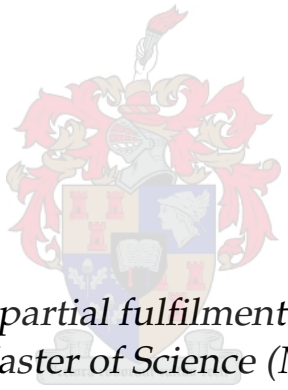
Comparison of approaches for spatial interpolation of weather data on a specific date

by

Gloria Burengengwa



Thesis presented in partial fulfilment of the requirements for the degree of Master of Science (Mathematics) in the Faculty of Science at Stellenbosch University

Supervisor: Dr David M. Drew
Co-supervisor: Prof. Cang Hui

March 2020

Declaration

By submitting this thesis electronically, I declare that the entirety of the work contained therein is my own, original work, that I am the sole author thereof (save to the extent explicitly otherwise stated), that reproduction and publication thereof by Stellenbosch University will not infringe any third party rights and that I have not previously in its entirety or in part submitted it for obtaining any qualification.

Date: March 2020

Abstract

Comparison of approaches for spatial interpolation of weather data on a specific date

G. Burengengwa

Department of Mathematical Sciences,
University of Stellenbosch,
Private Bag X1, Matieland 7602, South Africa.

Thesis: MSc

March 2020

This study compares four approaches to spatial interpolation of minimum and maximum temperature, and rainfall weather variables using data from 92 weather stations in the region of KwaZulu-Natal in South Africa. The approaches are Kriging with external drift (KED), Gaussian filter (GF), random forest (RF) and multilayer perceptron (MLP). The comparison was done against the background that the need for permanent gridded weather data for the region is important for agricultural and forest management. Also, there is little information regarding the suitability of methods for prediction in terms of performance variables for gridded data generation. The present research addresses these challenges by demonstrating the application of KED, GF, RF and MLP at a $1km^2$ spatial resolution accross three weather variables: minimum and maximum temperature, and rainfall to assess their performance. Four specific dates were selected to represent both dry and wet seasons for the years 2016 and 2017. The dates are 15^{th} of January 2016 and of 2017 for the summer season, and 15^{th} of July 2016 and of 2017 for the winter season respectively. Both years were considered because from available data, they are on the records as the driest (2016) and wettest (2017) in the region for the period 2008 to 2018. A cross-validation scheme was employed to assess the model performances and error evaluations were compared using RMSE, MAE and R^2 measures. The results were found to be almost similar across the four methods except for the RF model that outperformed in the periods considered for both years. Particularly, RF performed with the lowest RMSE and MAE errors for minimum and maximum temperature for both 15^{th} of July 2016 and of 2017 as against the other

models. The performance of RF is explained by the method's properties of being an ensemble technique. RF prediction follows from the principle of random selection of variables with high importance which allows for the decrease of uncertainty. The result of this research has importance for guiding decisions regarding forest management and climate driven businesses.

Acknowledgements

I would like to express my sincere gratitude to my two supervisors Dr David Drew and Professor Cang Hui for their guidance and continuous support during my studies. I am grateful to my family who have supported me along this journey. I am very thankful to Dr Ilaria Germishuizen for her contribution in understanding Meteoland and for her unconditional help throughout the work. I would like also to express my thanks to Anton Kuneke, the technician in the Forestry department for his technical assistance and his helpful discussions. I am grateful for the financial support from MONDI, SAPPI, Forestry South Africa and DTI THRIP funding. Lastly, I am grateful for all the people who contributed in one way or another towards the accomplishment of this thesis. They are a blessing in my life.

Dedications

I dedicate this work to my dear parents Mr Aloys Ndemeye and Mrs Gloriose Hatungimana

Contents

List of Figures

List of Tables

Acronyms

KZN Kwa-Zulu Natal
UK Universal kriging
GF Gaussian filter
RF Random forest
MLP Multilayer perceptron
RMSE Root mean squared error
MAE Mean absolute error
R^2 Coefficient of determination
CE Coefficient of efficiency
MODIS Moderate resolution imaging spectroradiometer
KED Kriging with external drift
DEM Digital elevation model
OK Ordinary kriging
SPEI Standardized Precipitation Evapotranspiration Index
EEA European Environmental Agency
RPE Relative predictor error
GWR Geographically weighted regression
MAP Mean Annual Precipitation
RBFN Radial basis function network
SASRI South African Sugarcane Research Institute
SAWS South African Weather Service
WMO World Meteorological Organisation
MAR Missing at random
MCAR Missing completely at random
QGIS Quantum geographic information system
UTM Universal transverse mercator

Chapter 1

Introduction and literature review

1.1 Research background

Access to accurate and reliable data relating to weather in a changing global climate is increasingly an area of interest to both policymakers, researchers and people whose livelihoods depend on favorable weather conditions (Gleason *et al.*, 2008). Spatial interpolation is an appealing array of methods by which such accurate and reliable data are derived. Interpolation is a process of estimating unknown values in between existing values. Predicting values of a primary variable at points within the same region of sampled locations is defined as spatial interpolation (Li and Heap, 2014). The product of the process is the generation of spatially continuous (interpolated) data.

Spatially continuous meteorological variables are important and serve as key inputs for many applications related to climate-driven issues. For instance, in the domain of forestry, the study done by DeCaceres *et al.* (2018) shows that process-based models usually requires daily meteorology for evaluating impacts of varying weather conditions on forest ecosystems. Another example is in the case of meteorological sciences where spatially continuous data are used to develop efficient spatial climate models that address climate-driven studies (McKenney *et al.*, 2011). Also, in finance, spatially continuous data are used to detect various economic indicators during crisis periods (Vermeulen *et al.*, 2015). Spatially continuous data are also used in the agriculture sector. The information they provide help farmers to determine the level of phosphorus content of the soil and thus, guide them to utilize the right amount of fertilizer in order to avoid soil pollution (Webster and Oliver, 2007). Agricultural engineers equally need full rainfall records for estimation of crop yields and growth models. Spatially continuous data are indeed vital for better planning and design.

However, such data are not easily available, difficult to obtain and often

exist at a relatively coarse resolution that leads to a loss of spatial heterogeneities in the regional physiography, particularly in areas with steep climate gradients, as shown by a number of authors (Foster *et al.*, 1996; Rigol *et al.*, 2001). Given that spatially continuous climatic data are frequently unavailable and essential, there is a need to generate regional information on climate at high resolution for individual sites or locations. The spatial resolution of the climate surfaces is relative and varies from one application to another depending on the data available and on the needs for that application. Only limited parts of the world have data at a fine resolution, $\leq 1\text{km}^2$ (Hijmans *et al.*, 2005; New *et al.*, 2002).

A number of methods are available for spatial interpolation. These vary in efficacy and suitability, depending on a number of factors (Li *et al.*, 2011; Li and Heap, 2008,1).

1.2 Interpolation

Interpolation is a process of estimating unknown values in between existing values. Specifically, interpolation involves adjusting parameters that fit a function to a set of k-dimensional data points, where k is a positive integer. Tobler's first law of geography which says "everything is related to everything else, but near things are more related than distant things" is the fundamental basis of all interpolation techniques (Tobler, 1970). Spatial interpolation is used to evaluate physical data in a continuous domain. It consists of calculating unknown values from a limited number of sample data points based on the assumption that spatially distributed objects are spatially correlated (Chorti and Hristopulos, 2008). Spatial interpolation is generally carried out by estimating a regionalized value at points not sampled from a weight of observed regionalized values (Mitas and Mitasova, 1999).

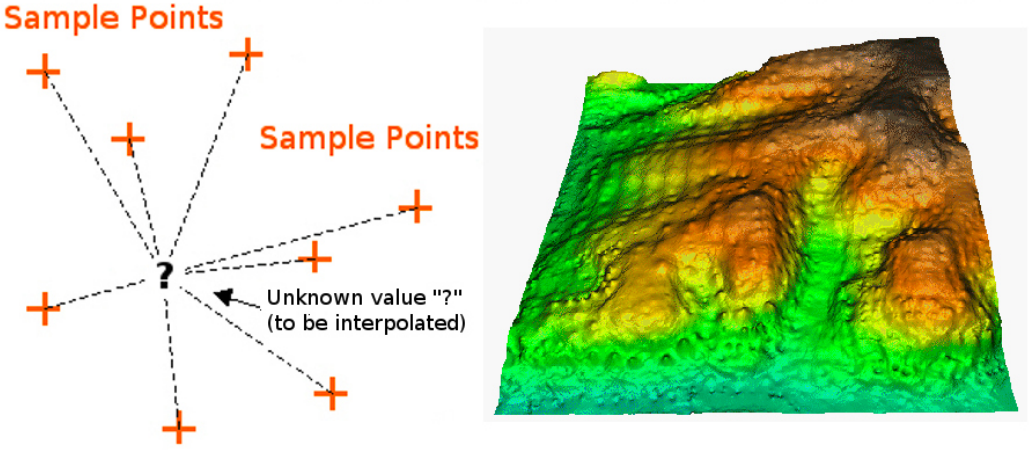


Figure 1.1 An example of a random field. In the left plot, the value for the unsampled point is calculated by weighting the values of the sample points. The right plot shows a random field with differences in slope, elevation, vegetation.

Mathematically, a general formulation of the spatial interpolation problem can be defined with the following statement:

Let us suppose a studied phenomenon z_j, $j = 1, \ldots, n$ with n values measured at different points $r_j = (x_j^{[1]}, x_j^{[2]}, \ldots, x_j^{[d]})$, $j = 1, \ldots, n$ within a certain region of d-dimensional space, (see example in Figure 1.1). Thus, Mitas and Mitasova (1999) defines a spatial interpolation as a problem that consists of finding a d-variate function $F(r)$ which passes through the given points, and fulfil the condition,

$$F(r_j) = z_j, \quad j = 1, \ldots, n \tag{1.1}$$

An infinite number of functions can be used to define the d-variate function $F(r)$ and can fulfil the requirement, however, the selection of an adequate method with suitable parameters is crucial given a particular phenomenon and the type of application desired. The many different techniques of interpolation that exist offer diverse performances, according to the characteristics of original data points (Caruso and Quarta, 1998).

Interpolation occurs with some errors given that an exact value cannot be estimated. If a considerable amount of errors occurs in predictions, setting such values as input surfaces for other studies may propagate with significant consequences within a sensitive system (Mowrer, 1997). Within the context of applied environmental modelling, effective techniques of spatial interpolation are required to efficiently predict gridded data while maximizing the information available from data that are often sparsely distributed

(Rigol *et al.*, 2001).

However, none of the interpolation methods is generalized to be optimal for all types of areas regardless of some assumptions. Thus, to be able to achieve an accurate construction of climate change scenarios at daily or monthly scales for a particular region, it is crucial to determine the best method, given values from nearby weather stations. The best way of assessing the performance of an interpolation method is to apply, on the same dataset of climate variables, different interpolation methods and compare their predictions under the same topographical conditions.

1.3 Classification of spatial interpolation methods

Spatial interpolation methods are classified into three main groups as non-geostatistical methods, geostatistical methods and combined methods (Li and Heap, 2008). In the three different categories of interpolation methods, different methods are combined in different ways using features described in table 1.1. The non-geostatistical methods include a range of interpolation methods such as nearest neighbour, inverse distance weighting, trend surface analysis, thin plate splines, etc. Geostatistical methods include also a significant range of interpolation methods Kriging and its derived interpolation methods such as ordinary Kriging, simple Kriging, Kriging with an external drift, universal Kriging, model based Kriging, etc. For the combined methods, there are several methods that have been trained together and found that they give more accurate results while completing each other. There we have regression Kriging, linear mixed model, trend surface analysis combined with Kriging, trend surface analysis combined with Kriging, etc. Also, a number of hybrid methods have been expanded in the field of machine learning such as support vector machine (SVM), neural network (NN), random forest (RF), etc (Li and Heap, 2014).

A classification of spatial interpolation methods is based on specific assumptions and features such as:

- Global or local: Interpolation methods are called `global` when they estimate the value of a given attribute according to all available data for a region of interest whereas `local` methods consider separately the variation of all of the region within a limited area.

- Exact or approximate: An interpolation method is said to be `exact` when for a sampled point, the observed value is equal to the esti-

mated value and an `approximate` method can be any other interpolation method apart from the exact method that gives a different value as the observed one.

- Stochastic and deterministic: A difference between `stochastic` methods and deterministic methods is that on top of the estimation value being interpolated, the stochastic method provides the error associated with the estimated value unlike `deterministic` method, which does not have assessment of error with the predicted value.

- Abrupt or gradual: `Abrupt` methods are interpolation methods that give a discrete surface while `gradual` methods are considered to be the interpolation methods that give smooth surfaces.

- Univariate or multivariate: The difference between `univariate` and multivariate methods is that univariate methods derive the estimate value from one primary variable whilst `multivariate` add to a primary variable one or more other variables, called second variables.

- Convex or non-convex: An interpolation method is `convex` if the estimated value is a value between the maximum and the minimum of the observed ones and is `non-convex` otherwise.

- Linear and non-linear: A `non-linear` interpolation method is a method that estimates values from a normal distribution of the observed values and a `linear` interpolation, does not take into account the normal distribution.

Table 1.1 Comparison of some interpolation techniques according to certain features, as described in the text.

Method	Global/local	Exact/ approximate	Stochastic/ deterministic	Abrupt/ gradual	Univariable/ multivariable
Inverse distance weighting (IDW)	Local	Inexact with regular smoothing window but can be forced to be exact	Deterministic	Gradual	Univariate
Thiessen polygons or nearest neighbours (NN)	Local	Exact	Deterministic	Abrupt	Univariate
Regression models (LM)	Global with local refinements	Approximate	Stochastic	Abrupt/ gradual if inputs have gradual variation	Univariate/ multivariate
Classification (CI)	Global	Approximate	Deterministic	Abrupt	Univariate
Trend surface analysis(TSA)	Global	Approximate	Stochastic	Gradual	Multivariate
Thin plate splines (TPS)	Local/ Global	Exact within smoothing limits	Deterministic with local stochastic component	Abrupt/ gradual	Univariate/ multivariate
Kriging	Local with global trends or with global variograms when stratified or not	Exact	Stochastic	Gradual	Univariate/ multivariate

In the case of the methods listed in table 1.1, knowing the properties of these methods helps in choosing an appropriate method depending on the requirements of each of them.

1.4 Examples of applications of spatial interpolation by previous scholars

1.4.1 A brief review of Kriging and random forest performance

Various studies acknowledged the use and performance of Kriging method in comparison with a number of other methods, including, nearest neighbour, inverse distance weighting and random forest (RF) (Berndt *et al.*, 2018; Hengl *et al.*, 2018; Pebesma *et al.*, 2016). Of interest is the fact that all three scholars in the following paragraph affirm the good performance of Kriging. But it is Hengl *et al.* (2018) that acknowledged the complementarity value of RF to Kriging. Their study demonstrates that RF can equally perform in pair with Kriging. In the current research, it will be of interest to observe how these different methods perform relative to one another.

In lower Saxony, in northern Germany, Berndt *et al.* (2018) compared the performance of Nearest neighbour, inverse distance weighting and Kriging with external drift (KED) under the influence of temporal resolution and network density in predicting the following weather variables, precipitation, sunshine duration, relative humidity, cloud coverage and wind speed for the period 2008 to 2013. They found that ordinary Kriging performed better than others for wind, cloudiness and sunshine duration, regardless of network station density and temporal resolution. From a study done by Pebesma *et al.* (2016), various spatio-temporal covariance models and spatio-temporal interpolations were compared to a purely spatial Kriging approach. It was found that spatio-temporal covariance structures carry useful information, but that a spatio-temporal Kriging model does not guarantee to outperform over a pure spatial prediction unless there exists a strong correlation between the locations and the temporal dimension. In Hengl *et al.* (2018), Random forest (RF) was compared to ordinary Kriging and regression Kriging for spatial prediction in order to incorporate the effect of geographical proximity. They found RF to produce equally accurate predictions as the Kriging methods. In their study, RF was found to easily adapt locally in time and space in the case of spatio-temporal predictions in comparison to Kriging.

1.4.2 A brief review of Meteoland

From the following studies, Meteoland De Cáceres (2019), a new R package within the R programming language (R Core Team, 2013), which is based on the Gaussian filter method and is used to predict daily data at unknown locations, is shown to have been successfully used for spatial predictions since its launch in 2016 (Germishuizen, 2018; Karavani *et al.*, 2018; Sánchez-Pinillos *et al.*, 2018). Therefore, given that it is a new technique in the field of spatial interpolation and has been employed for different purposes, this raised the interest to compare it with other well-known methods and assess its performance.

Karavani *et al.* (2018) used Meteoland to predict weather variables in order to understand the impact of climate change and soil moisture on mushroom productivity in Mediterranean forests. Strong relationships between weather conditions and mushroom productivity were found to exist. Karavani *et al.* (2018) observed that high temperatures tend to negatively impact the period of yield at the beginning of the fruit season while they enhance it by its end. In another study, Meteoland was effectively applied in calculating the monthly mean temperature and precipitation (Sánchez-Pinillos *et al.*, 2018). Sánchez-Pinillos *et al.* (2018) were interested in the assessement of post-fire recruitement of non-serotinous pines. They used Meteoland to derive the spatially continuous data needed for the analysis of the climatic variables relating to the mechanisms of facilitation or competition between resprouters and pines. Meteoland was also used to derive mean monthly temperature and precipitation and the result was then utilized to calculate the Standardized Precipitation-Evapotranspiration Index (SPEI). The SPEI calculated assessed the relationship between the effects of the variables employed and changes operated in water availability between years. This way, they demonstrate the applicability of the method to problems regarding mean value calculation.

1.4.3 An example of an interpolation method performance at a 1km^2

"WordClim2" was developed by Fick and Hijmans (2017) to create a dataset of spatially interpolated monthly temperature, precipitation, solar radiation, wind speed and vapour pressure at high resolution (approximately 1km^2) after "WorldClim 1" Hijmans *et al.* (2005), which created only average monthly temperature and precipitation. In Fick and Hijmans (2017), remotely sensed land surface temperature data obtained from the MODIS satellite were used to improve the estimation of areas where weather station density is low. They implemented thin-plate smoothing spline algorithm in

the program SPLINA from ANUSPLIN to interpolate the climatic variables. The results indicated high correlation values for temperature and solar radiation while the values were low for precipitation and wind speed. They found the RMSE for the average temperature to be lower between 1.1° and 1.4°C. After adding satellite-derived data, Fick and Hijmans (2017) found no improvement or even negative effects for all of the other climatic variables apart from temperature in particular, which showed the greatest positive effect in regions with high elevation. Contrary to other findings Hutchinson (1995); Kilibarda *et al.* (2014); Neteler (2010) which illustrate that adding more predictor variables in interpolating climate surfaces can increase prediction precision as demonstrated using splines and other methods, Fick and Hijmans (2017) work seems to suggest otherwise. Their optimal spline model formulations used variables that vary across regions and accounted for local context effects on climatic processes (e.g. the effect of elevation on precipitation), thereby emphasizing the effect of the latter rather than the co-variables. Their adaptive approach allows for better model fits in remote regions.

The impact and findings of "WordClim2" which are a positive effect in predicting temperature where elevation is high and also the constat of an increase in predictions while increasing co-variables, inspired us to run interpolation at the same resolution of 1km^2 and to evaluate the effect of more than one co-variable on temperature and rainfall predictions for KZN which is a region with a presence of mountains.

1.4.4 A brief review of seasonal and geographical effects on interpolation

In Yuan *et al.* (2015), thin-plate spline is applied to generate gridded climate datasets for China from 3 co-variables, longitude, latitude and elevation on a daily basis. They found a relatively poor performance for precipitation. The thin-plate spline did not show distinct differences in RMSE for the three temperature variables. But it was found that in spring and autumn, the averaged R^2 were larger for the three variables. The model showed different performances for precipitation in the four seasons with the lowest performance observed in summer. Their findings showed different performances of the same interpolation method in different seasons, therefore observing the behaviour of an interpolation method across different seasons is found worthy to be evaluated in this research.

Given some of the geographical similarities of Australia to South Africa, it is important to mention the study done by Beesley *et al.* (2009) which com-

pared two archived Australia-wide 5km gridded gauge-based daily rainfall. They used common cross validation statistics, that are of interest in this research. There were used in order to revise previous error comparisons done by Jeffrey *et al.* (2001) and Jeffery (2006) who considered different observations for their error analyses comparison. The two archived sets are BAWAP (Bureau of the Australian Water Availability Project) and "SILO" (Queensland, 1889) and they contain publicly available patched and gridded datasets. The study found that the SILO direct daily rainfall and the monthly disaggregation method compute almost identical errors for three different error types, the mean error (ME), the mean absolute error (MAE) and the root mean squared error (RMSE). Both BAWAP and SILO have similar errors, however SILO performs slightly better than BAWAP. They found that the ME decreases considerably for cross validation of wet days, days with a rainfall record higher than 0.0mm, which implies an underestimation of the wet days.

1.5 Work done in South Africa

A number of studies done in South Africa have employed spatial interpolation methods in different areas of research.

A geographically weighted regression (GWR) approach was used by Schulze *et al.* (2007) to interpolate daily, monthly and annual rainfall values throughout South Africa. Maps of the mean annual precipitation (MAP) of South Africa were also generated using GWR. Specifically, they used the GWR to derive a 50-year time series of continuous daily rainfall at stations that have representative values of the 1946 quaternary catchments covering the area of study (Schulze *et al.*, 2007). The inverse distance weighted method was used to evaluate the differences between observed and estimated MAPs values. Therefore, a correction of the interpolated MAPs values was made using the MAPs values generated by the inverse distance weighted to fit where there are observed values and at the ungauged locations.

In Makhuvha *et al.* (1997), six different regression methods of estimating missing values at a target site were tested in order to patch rainfall data. The methods were subdivided into two main approaches, in accordance with the types of missing values. For this study, Makhuvha *et al.* (1997) mostly focused on the background theories of the methods rather than making their comparison in order to find the most efficient method.

Germishuizen (2018), employed the Meteoland package in evaluating

the ecological niche distribution of the Eucalyptus gall wasp *Leptocybe invasia* in the plantation forestry areas of South Africa and predict the risk of outbreaks at different spatial and temporal scales. With Meteoland, Germishuizen (2018), developed monthly grids of different weather variables for the years 2015, 2016 and 2017 and these datasets constitute actual environmental parameters to define the climatic niche of *Leptocybe invasia* at monthly intervals and identify seasonal and annual changes in risk of outbreaks.

To estimate rainfall values and patch missing data for different catchments of the Southern Africa, Hughes and Smakhtin (1996) developed some spatial interpolation approaches. One of the approaches is the VTI model which they defined as "a semi-distributed model that operates with a daily time step". They found that in the Southern Cape, the VTI model was more successful than the patching model, which considers at most five available sites within the site to be estimated. They found that the patching algorithm overestimated the high flows.

Niekerk *et al.* (2011) investigated the suitability of four co-variables namely slope, aspect, hillshade and distance to the coast for interpolating climate surfaces in the Western Cape Province. They found no singificant effect of slope, aspect and hillshade, only the hillshade with a 180° azimuth was found to positively affect rainfall predictions but not temperature predictions. Distance to the coast which is related to large water bodies was found to decrease the mean error of monthly mean maximum daily temperature by 27%. The latter co-variable was also found to improve the accuracy of monthly mean minimum daily temperature interpolations for seven months, from October up to April.

1.6 Problem statement

Various approaches of spatial interpolation have proven to perform well in estimating temperature and rainfall. However, most of the scholars have compared either Kriging with inverse distance weighting and nearest neighbour (Berndt *et al.*, 2018; Coulibaly and Becker, 2007) or with random forest (Hengl *et al.*, 2018). There is insufficient information on how the performance of the Gaussian filter through Meteoland compares to several other methods across variables in the literature. Also, to the best of my knowledge, no work has looked at the KZN region comparing interpolation methods performance as set out in this study. In the current research, the focus is on the minimum and maximum temperature, and rainfall predictions. The reason for choosing specifically these weather variables is because temperature and rainfall estimations are essential in monitoring the environment, predicting crop yield, determining the spatial distribution of plant develop-

ment, conditioning agricultural soil suitability and in many more important analyses (Flores, Lillo *et al.*, 2010). The oversight observed in the literature is significant because the region is of great agricultural importance (Wojtasik, 2014). Therefore, in the context of an increasingly variable climate, good quality estimates of weather data are needed for modelling to guide decisions and policies to wildlife management and agricultural practices.

1.7 Research objective

The main objective of this research is to test and compare the performances of four methods of interpolation namely Kriging with external drift (KED), Gaussian filter (GF), random forest (RF) and Multilayer Perceptron (MLP) at a relatively fine scale of 1km^2 resolution applied to minimum and maximum temperature, and rainfall in the region of KwaZulu Natal (KZN) province of South Africa. The theory and principles of these four methods are described in more detail in chapter 2.

1.8 Research questions

Towards addressing this objective, three questions were asked:

- How well do the four methods, KED, GF, RF and MLP, compare to one another in terms of their predictions of minimum and maximum temperature, and rainfall?

- Does a method show better predictions between the three weather variables namely minimum and maximum temperature, and rainfall?

- Does the performance of the methods change between dry and wet years, or between summer and winter seasons?

Chapter 2

Overview of the methods applied

The second chapter gives an overview of the methods used in this research. The chapter begins with a detailed description of each of them and how they were used. The accuracy metrics used to evaluate the models' performances are discussed. Also in this chapter, the study area in which the research is focused is explored and the source and selection of the co-variables necessary for each interpolation approach are detailed. The chapter closes by a highlight of the packages employed to run the models.

2.1 Kriging

Kriging is often used across many disciplines. The name is derived from a South African mining engineer, Daniel Krige $(1919 - 2013)$, but it was built as a method by a French mathematician called Georges Matheron who developed the general concept and theory of linear geostatistics and for Kriging interpolation (Stein, 2012). Kriging is a spatial prediction known to be a best linear unbiased prediction (BLUP) because it estimates values for unknown locations of a sample of observations and, in addition to the interpolation, it generates errors for each predicted value which are not available for other interpolations methods.

Kriging models are fitted to data obtained for large areas, which is the reason why they are defined as being global rather than local. Despite the fact that they are used for prediction, they can also be used for sensitive analysis of complex computer codes which often need much computer time and for optimal design that include automobiles, computer monitors and airplanes (Meckesheimer *et al.*, 2002; Simpson *et al.*, 2001). Some sensitive analysis and optimization require to interpolate the observed input and output data and this is done using a metamodel of the underlying simulation model such as Kriging (Van Beers and Kleijnen, 2004).

Other studies using Kriging might consider higher dimensional inputs, greater than three. For instance, Sacks *et al.* (1989) used the Kriging models with k dimensional input where k is a given positive integer however geostatistics considers two-dimensional or sometimes three-dimensional inputs (Kleijnen, 2009).

Kriging requires a good understanding of the spatial behaviour of the phenomenon represented and of the principles of spatial autocorrelation. Kriging assumes spatial isotropy and stationarity of the field of study. This implies that the properties of the field, such as the mean, standard deviation and autocorrelation do not change over time and that they are spatially uniform in all orientations. One of the advantages of Kriging compared to other interpolation methods is that in the presence of irregularity in the variation of data, Kriging gives unbiased predictions, while other simple methods of interpolation may give unreliable predictions (Berndt *et al.*, 2018).

Spatial estimation using Kriging involves the computation of a covariance matrix and the estimation of a semivariogram, which is a variance function that relates spatial dispersion in a set of data (Sakata *et al.*, 2003). A semivariogram $\gamma(h)$, is defined as half the mean squared difference between two observations of a variable separated by a distance vector h (Uyan and Cay, 2013). Therefore, considering that statistics and geostatistics are sciences of the unknown, the true semivariogram is never known (Olea, 2006). However, the following unbiased estimator is used in practice:

$$\gamma(h) = \frac{1}{2N(h)} \sum_{i=1}^{N(h)} (z_i - z_{i+h})^2$$

where $\gamma(h)$ is the experimental semivariogram of the distance class h, $N(h)$ the number of observations, z_i and z_{i+h} the observed values at locations i and $i + h$ separated approximately by h. The locations represent vectors of coordinates which are denoted as easting and northing for our purposes, since we are working in two dimensions and h is the lag and has a magnitude defined by a distance and an orientation.

2.1.1 Variogram model parameters

2.1.1.1 Nugget

The nugget (Fig. 2.1) is the estimated non-zero semivariogram as distance approaches zero (Kerry and Oliver, 2008). It is one of the most important parameters in Kriging prediction and needs to be chosen well according to

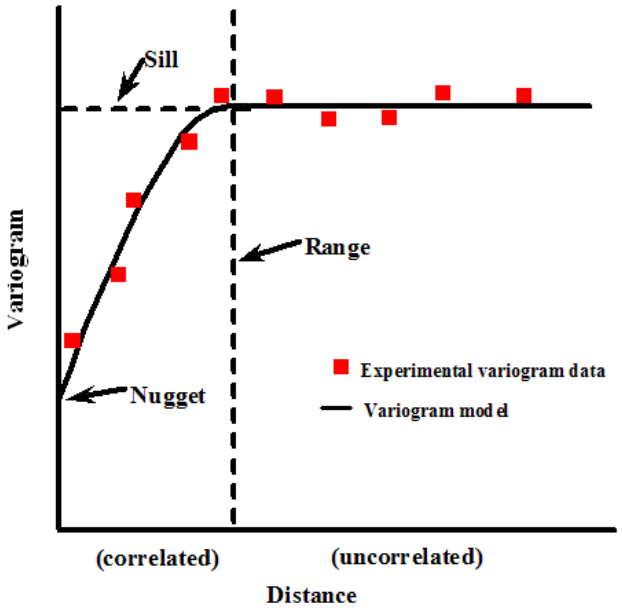


Figure 2.1 The figure shows a fitted variogram model. The x-axis represents the distance between sample points and the y-axis the calculated value of the variogram. The red squares represent the lags of the variogram.

the spatial variability of the data. Increasing the nugget size has certain effects on the model outputs, such as an increase in uncertainty, a decrease in the relative influence of nearby data and, at more isolated locations, some surfaces rebound more towards the mean (Pardo-Igúzquiza, 1999).

2.1.1.2 Range

The length of a range (Fig. 2.1) affects predictions. For short range, variance is higher and predicted values approach the mean, while for long range the variance is lower and nearby observations influence predicted values the most (Kerry and Oliver, 2008).

2.1.1.3 Sill

The sill (Fig. 2.1) is a plateau in semivariogram that occurs at a distance defined by the range (Kerry and Oliver, 2008). When there are many isolated observations, high sill values make estimates move towards the mean while low sill values pull estimates move towards values nearby observations. In

the case of sparse observations, the effects are more pronounced.

Geostatistical analysis employs variograms to present correlations of spatial variability (Oliver and Webster, 1990). If the variability of a variogram, where a variogram is a distance function of a spatial variance, does not vary across spatial directions, it is called isotropy. If the variability varies with spatial directions, it is called anisotropy. The anisotropic variogram model is a function of both distance and direction, and its equation can be written as follows: $(h,\theta) = \left(\frac{1}{2}N(h,\theta)\right)$

$$(h,\theta) = (\frac{1}{2}N(h,\theta)) \sum_{i=1}^{N(h,\theta)} [z(x_i,\theta) - z(x_i+h,\theta)]^2$$

where

θ, is the angle along point x_i and x_{i+h},

$N(h,\theta)$, the pairs of samples with interval h in the angles along x_i and x_{i+h}.

With 2 regionalized variables z and y, we obtain the joint variogram as follows:

$$\gamma_{zy}(h) = \frac{1}{2N(h)} \sum_{i=1}^{N(h)} [z(x_i) - z(x_{i+h})] \times [y(x_i) - y(x_{i+h})]$$

2.1.2 Types of variogram models

A variogram is usually estimated at various lags and a parametric model is fitted to those estimates before prediction can be performed on spatial datasets (Gorsich and Genton, 2000). There exist different types of theoretical variogram models and the variogram properties presented in the previous paragraph give insights into how a model represents the best fit. There are four most commonly used models: linear, spherical, Gaussian and exponential.

2.1.2.1 Spherical model

A curve known as a spherical model (Figure 2.2), is considered to fit the variogram well in cases where the nugget variance is important but not large and there is a clear range and sill. In Ly *et al.* (2011), a spherical model is given by the following equation:

$$\gamma(h) = \begin{cases} C_0(1 - H(0)) + C_1(\frac{3}{2}\frac{h}{a} - (\frac{1}{2}(\frac{h}{a})^3), & 0 \leq h < a, \\ C_0 + C_1, & a \leq h, \end{cases} \tag{2.1}$$

where,

C_0 is the nugget effect,

h, represents the lag or distance between the observations,

$H(0)$, is the heaviside function which is 1 at lag 0 and 0 otherwise,

a, the range,

and $C_0 + C_1$ represents the sill.

2.1.2.2 Exponential model

An exponential model observed in Figure 2.2, is considered when there is a clear nugget and sill but only a gradual approach to the range. Its equation is as follows:

$$\gamma(h) = C_0(1 - H(0)) + C_1(1 - e^{-\frac{3h}{a}}), \qquad 0 \leq h \tag{2.2}$$

The variables are the same as in equation 2.1.

2.1.2.3 Linear model

Hartkamp *et al.* (1999) define a linear model (Figure 2.2), in presence of a sill by the following equation:

$$\gamma(h) = (C_0 + C_1)\frac{h}{r} \quad , \qquad h \in (0, r] \tag{2.3}$$

The variables are the same as in equation 2.1.

2.1.2.4 Gaussian model

A Gaussian model (Figure 2.2), is used in cases where the variation is very smooth and the nugget variance is very small compared to the spatially random variation (Hartkamp *et al.*, 1999). The Gaussian model is as follows:

$$\gamma(h) = C_0(1 - H(0)) + C_1(1 - e^{-3(\frac{h}{a})^2}), \qquad 0 \leq h \tag{2.4}$$

The variables are the same as in equation 2.1.

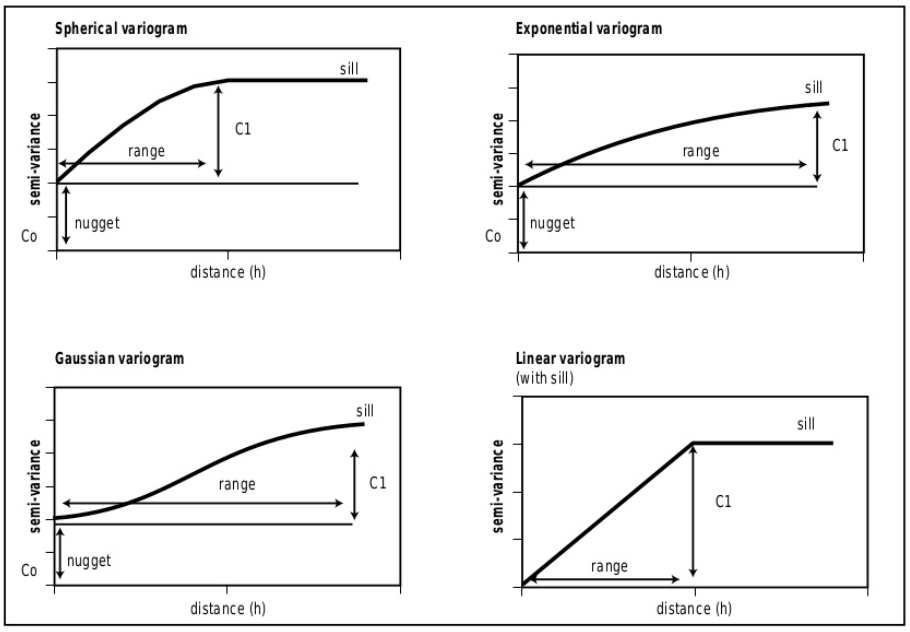


Figure 2.2 Examples of the four most commonly used variogram models.

Despite numerous publications on Kriging variogram models, the most common methods of fitting semivariogram models to experimental semivariograms are performed using manual fitting procedures (Ly *et al.*, 2011). This is not an appropriate approach. Thus, finding a good fit requires mostly a very good knowledge and experience in the field. Olea (2006) emphasizes that modelling a semivariogram remains to the uninitiated the most difficult and intriguing aspect in the application of geostatistics.

2.1.3 Kriging with external drift (KED)

The Kriging with external drift method allows the prediction of a particular variable, Z, known only at some locations, taking into account another variable, u, known everywhere. A random function, $Z(x)$, is chosen to model the Z data. The second variable is represented as a regionalized variable, $u(x)$ (Bourennane *et al.*, 1996). The two quantities are assumed to be linearly related as $Z(x)$ and $u(x)$ are two ways of expressing the same phenomenon. It is assumed that $Z(x)$ is on average equal to $u(x)$ up to a constant a_0 and a coefficient b_1 by (Bourennane *et al.*, 1996, 2000):

$$E(Z(x)) = a_0 + b_1 u(x) \tag{2.5}$$

This method merges both sources of information, uses the variable $u(x)$ as an external drift function for estimating $Z(x)$. Let us consider the prob-

lem of improving the estimation of a second-order stationary random function $Z(x)$. This is solved by introducing the shape function $u(x)$ providing detail at smaller scale than the average sample spacing for $Z(x)$. The estimator is the linear combination of the sample values at location $x_\alpha (\alpha = 1, \ldots, n)$ with unit sum weight w_α,

$$\hat{Z}(x_0) = \sum_{\alpha=1}^{n} w_\alpha Z(x_\alpha) \tag{2.6}$$

with weights constrained to sum to 1 $\sum_{\alpha=1}^{n} w_\alpha = 1$

In this way, the prediction error is expected to be zero:

$$\begin{aligned} E[\hat{Z}(x_0) - Z(x_0)] &= 0 \\ E[\hat{Z}(x_0)] &= E[Z(x_0)] \end{aligned} \tag{2.7}$$

This equality can be developed into:

$$\begin{aligned} E[\hat{Z}(x_0)] &= \sum_{\alpha=1}^{n} w_\alpha E[Z(x_\alpha)] \\ &= a_0 + b_1 \sum_{\alpha=1}^{n} w_\alpha u(x_\alpha) \\ &= a_0 + b_1 u(x_0) \end{aligned} \tag{2.8}$$

This implies that the weights should on average be consistent with an exact interpolation of $u(x)$.

$$u(x_0) = \sum_{\alpha=1}^{n} w_\alpha u(x_\alpha). \tag{2.9}$$

Therefore, the objective function (O) to minimize in this case consists of the estimation variance σ_E^2 and of two constraints:

$$O = \sigma_E^2 - \mu_1 \left(\sum_{\alpha=1}^{n} w_\alpha - 1 \right) - \mu_2 \left(\sum_{\alpha=1}^{n} w_\alpha u(x_\alpha) - u(x_0) \right) \tag{2.10}$$

with μ_1 and μ_2 the Lagrange parameters, and
σ_E^2 the estimation variance.

The estimation variance σ_E^2 is equal to

$$\begin{aligned}\sigma_E^2 &= var\left[\hat{Z}(x_0) - Z(x_0)\right] \\ &= \sum_{\alpha=1}^{n}\sum_{\beta=1}^{n} w_\alpha w_\beta C(x_\alpha - x_\beta) - 2\sum_{\alpha=1}^{n} w_\alpha C(x_\alpha - x_0) + C(0)\end{aligned} \tag{2.11}$$

where C is the covariance function.

The partial derivatives of the objective function $O(w_\alpha, \mu_1, \mu_2)$ are set to zero to find the minimum of the quadratic function σ_E^2:

$$\begin{cases} \dfrac{\partial O}{\partial w_\alpha} = 0 \\ \dfrac{\partial O}{\partial \mu_1} = 0 \\ \dfrac{\partial O}{\partial \mu_2} = 0 \end{cases} \tag{2.12}$$

$$\begin{cases} \displaystyle\sum_{\beta=1}^{n} w_\beta C(x_\alpha - x_\beta) - \mu_1 - \mu_2 u(x_\alpha) = C(x_\alpha - x_0) \\ \displaystyle\sum_{\alpha=1}^{n} w_\alpha = 1 \\ \displaystyle\sum_{\alpha=1}^{n} w_\alpha u(x_\alpha) = u(x_0) \end{cases} \tag{2.13}$$

The result of minimization is a system of linear equations called universal Kriging equations:

$$\begin{cases} \displaystyle\sum_{\beta=1}^{n} w_\beta C(x_\alpha - x_\beta) - \mu_1 - \mu_2 u(x_\alpha) = C(x_\alpha - x_0) \\ \displaystyle\sum_{\beta=1}^{n} w_\beta = 1 \\ \displaystyle\sum_{\beta=1}^{n} w_\beta u(x_\beta) = u(x_0) \end{cases} \tag{2.14}$$

with the minimal estimation variance

$$\sigma_E^2 = C(0) - \sum_{\alpha=1}^{n} w_\alpha C(x_\alpha - x_0) + \mu_1 + \mu_2 u(x_0) \tag{2.15}$$

Thus, KED consists of incorporating into the Kriging system supplementary universality conditions about one or several external drift variables,

$u_i(x), i = 1, \ldots M$, measured exhaustively in the spatial domain (Bourennane *et al.*, 2000). The functions $u_i(x)$ need to be known at all locations x_α of the samples $Z(x_\alpha)$, as well as at nodes of the estimation grid.

The following conditions are added to the Kriging system independently of the class of covariances, hence the qualificative external. A Kriging system while in presence of translation invariant (internal drift) and multiple external drift, can be written as:

$$\sum_{\beta=1}^{n} C(x_\alpha - x_\beta) - \sum_{l=0}^{L} \mu_l f_l(x_\alpha) - \sum_{i=1}^{M} \mu_i u_i(x_\alpha) = C(x_\alpha - x_0) \tag{2.16}$$

for $\alpha = 1, \ldots, n$

$$\sum_{\beta=1}^{n} w_\beta f_l(x_\beta) = f_l(x_0) \tag{2.17}$$

for $l = 1, \ldots, L$

$$\sum_{\beta=1}^{n} w_\beta u_i(x_\beta) = u_i(x_0) \tag{2.18}$$

for $i = 1, \ldots, M$

2.1.3.1 Parameterization

The parameters of Kriging with external drift used in this study are the following:

- A mask: A rectangular grid is first defined, represented by a mask that contains all the spatial coordinates of the KZN region at a 1km resolution.

- Among all the drift terms possibilities within universal Kriging. The "specified drift" was selected.

- The variogram used is a linear variogram.

- The number of specified drifts is also specified and represented by elevation and coastal distance.

- The grid to consider for prediction is also defined by the x-points and y-points.

2.2 Gaussian interpolation

The truncated Gaussian is a stochastic interpolation method which is based on facies distribution. The spatial distribution and relationships can be easily tuned to produce numerous different textures such as high frequency layering or oriented facies (Beucher and Renard, 2016). For the purpose of this project, an R package, Meteoland (De Cáceres, 2019), which relies on the truncated Gaussian filter, is used.

To elaborate and develop Meteoland, De Cáceres (2019) referred to the concept of interpolation which provides daily values illustrated by Thornton *et al.* (1997) whose idea comes from the existence of a model called MTCLIM. The model provides daily values from meteorological variables. It was developed by extrapolating daily observations from a maximum of two stations to a remote and uninstrumented site. The procedure of extrapolating temperatures with elevation which apply throughout the year is accomplished by user-specified lapse rates that are derived from regional observations and holds constant in space and time. The extrapolation of daily precipitation is achieved using a ratio of mean annual total precipitation between the sites of observation and prediction, with the predicted occurrence of precipitation events duplicated from the observed time series of daily precipitation. However, the method does not give allowance for temporal variation (Thornton *et al.*, 1997).

The MTCLIM model assumes that a reasonable horizontal meteorological variability can be represented over a region of approximately 2000 km^2. The assumption is not valid for studies over larger regions where the area tends to exceed 2000 km^2. Based on the MTCLIM logic, Thornton *et al.* (1997) included then interpolations between an unspecified number of heterogeneously spaced observations in complex terrain that are not restricted to the area of the terrain.

In developing the method, the idea of the assertion that the area of relative influence for a given observation should be inversely related to the local observation density, was derived from the nearest neighbour method. But, due to the lack of possibility of generating continuous surfaces using the nearest neighbour method and given a desire to generate a continuous interpolation surface which does not have to be perfectly smooth. A surface not perfectly smooth releases the condition of continuity for the derivatives and allows the first and the higher-order derivatives to be discontinuous. Thornton *et al.* (1997) postulated that the relative influence should decrease with increasing distance from an observation. The statement was borrowed from the inverse-distance method. However, the desire to generate a smoother method where the resultant surface is not required to pass through the ob-

servations was not fulfilled using the inverse-distance method. The reason is that the asymptotic condition of the inverse-distance method forces the surface through all observations and therefore generates anomalies in the spatial distributions.

For these reasons, the approach of a truncated Gaussian filter with a surface containing the horizontal projections of the observation locations is adopted. The truncated filter serves to reduce the number of observations included in predictions at a given point. A Gaussian function satisfies the needs for implementing Meteoland given the desired features of being both an inverse-distance algorithm and a smoothing filter.

Meteoland is then developed from the above mentioned. It provides estimation of daily weather variables such as precipitation, maximum and minimum temperature, wind, relative humidity, solar radiation at any location of a landscape. The interpolation of the first four variables is done using truncated Gaussian filters, and consists of spatially defining the weight at radial distance r to the point of interest p:

$$W(r) = \begin{cases} e^{-\alpha(r/R_p)^2} - e^{-\alpha} & \text{if} \quad r \leq R_p \\ 0 & \text{otherwise} \end{cases} \tag{2.19}$$

where α is a unitless shape parameter, R_p the truncation distance from p, and $W(r)$ the filter weight associated with the radial distance r from p.

Thus for each target point, we have a vector of weights associated with observations as a result of spatial convolution of the filter with a set of weather station locations. A constant value for R_p results in a large disparity in the number of observations where the weights are non zero between the points in the least and the most densely populated regions of the prediction grid (Thornton *et al.*, 1997). R_p is automatically adjusted in such a way that it is smaller in data-rich regions and increases in regions with less data. A fixed number of observations to be used at every prediction point could be specified, but this is unfortunately not the best way to proceed because it violates the requirement for a continuous surface. Instead, we specify at each point p, an average number of observations N to be included. The truncation distance R_p is then varied as a smooth function of the local station density in such a way that the average is achieved over the spatial domain.

Therefore, we have a continuous interpolation surface ensured by the smooth variation of R_p which is accomplished through the iterative estimation of local station density at each prediction point. The estimation of R_p is done by the following:

1. R_p is initialized at a value specified by the user.

2. Given R_p, the local density D_p which denotes the number of stations per area, is calculated after the calculation of interpolation weights W_i for all $i = (1, ..., n)$ stations using equation 2.19:

$$D_p = \frac{\sum_{i=1}^{n} W_i / \bar{W}}{\pi R_p^2} \tag{2.20}$$

where $\bar{W}$ is the average weights over the untruncated regions of the filter given by the equation:

$$\begin{aligned} \bar{W} &= \frac{\int_0^{R_p} W(r)dr}{\pi R_p^2} \\ &= \left(\frac{1 - e^{-\alpha}}{\alpha}\right) - e^{-\alpha} \end{aligned} \tag{2.21}$$

3. A new R_p is then calculated as a function of N, the average number of observations and D_p, by the following:

$$R_p = \sqrt{\frac{N^*}{D_p \cdot \pi}} \tag{2.22}$$

where,

$$N^* = \begin{cases} 2N & \text{for the first } I - 1 \text{ iterations} \\ N & \text{for the final iteration} \end{cases}$$

4. This new R_p is substituted in equation 2.20 where D_p is in turn used for the step 3, the process is repeated for the specified number of times I. The final R_p is then used to generate weights W_i.

Therefore, an estimation of R_p for each target point is done for each day and weather variable. To compute the algorithm, 4 parameters such as N, R, I and α are required. In Meteoland, De Cáceres (2019) define the two parameters R and I to be set by default, therefore $R = 140000$ and $I = 3$ while the number of observations N and the shape of parameter α depend on the variable to be interpolated. The values for the interpolation parameters are specified once and held constant over all days and all prediction points. Given an arbitrary variable x_i, measured at each of the observations points $i = 1, \ldots, n$, the interpolated value x_p of a single prediction point on a single day is determined in general as:

$$x_p = \frac{\sum_{i=1}^{n} W_i x_i}{\sum_{i=1}^{n} W_i} \tag{2.23}$$

As mentioned previously, the process of interpolation is done differently from one variable to another, making the general method specific to predictions of daily minimum and maximum temperature and daily total precipitation, incorporating the influence of elevation differences.

2.2.1 Temperature

Suppose daily temperature is denoted T. Meteoland predicts daily maximum temperature and daily minimum temperature in the same way. Let us denote T_p the variable temperature to be predicted at a single target point p and for a single day based on the observations T_i and interpolation weights W_i, with $i = 1, \ldots, n$ representing the weather stations. A correction for the effects of elevation differences between the observation and the prediction points is included in equation 2.23 to predict T_p. The correction is based on an empirical analysis of the relationship of T to elevation, which is performed once for each day of prediction (Thornton *et al.*, 1997). De Cáceres (2019) used a weighted least squares regression to assess the relationship between temperature and elevation. This is done by assessing the difference in elevations associated with a pair of observations instead of regressing z_i on T_i, where z_i are the elevations recorded for stations i with $i = 1, \ldots, n$. The dependent variable is the corresponding difference in temperatures associated with a pair of stations and the independent variable is the difference in elevations associated with the pair. We then have a regression of the form

$$(T_1 - T_2) = \beta_0 + \beta_1(z_1 - z_2). \tag{2.24}$$

Here β_0 and β_1 are the regression coefficients. The daily regression is performed over all unique pairs of stations. The weights within the pairs of stations are used to find the regression weights associated with each point. Thornton *et al.* (1997) found this approach to be more robust than the simpler method of regressing t_i against z_i, using the W_i as regression weights.

The daily maximum or minimum temperature T_p is the predicted as follows:

$$T_p = \frac{\sum_{i=1}^{n} W_i[T_i + \beta_0 + \beta_1(z_p - z_i)]}{\sum_{i=1}^{n} W_i} \tag{2.25}$$

where z_p is the elevation assigned to the target point.

2.2.2 Precipitation

The process of predicting precipitation is different to the prediction of temperature and complicated because it requires predicting both daily precipi-

tation occurrence and, conditional on that result, daily precipitation amount. The patterns occurrence of precipitation show some spatial coherence on wet versus dry when measured at the time scale of a day. Thornton *et al.* (1997) defined under that assumption, a binomial predictor of spatial precipitation occurrence as a function of the weighted occurrence at surrounding stations.

Therefore, a precipitation occurrence probability, denoted POP_p, is estimated for a case of a single prediction point on a given day and given observations of daily total precipitation P_i and interpolation weights W_i:

$$POP_p = \frac{\sum_{i=1}^{n} W_i PO_i}{\sum_{i=1}^{n} W_i} \tag{2.26}$$

where PO_i are the binomial variables related to observed precipitation occurrence, $i = (1, \ldots, n)$ the observation locations, n the total number of observations,

$$PO_i = \begin{cases} 0; & P_i = 0 \\ 1; & P_i > 0 \end{cases} \tag{2.27}$$

Once POP_p is calculated, we then have PO_p, the daily binomial predictions of precipitation occurrence at a given point, that are based on the comparison of POP_p with a specified critical value, POP_{crit}, and this critical value is held constant for the whole spatial and temporal domain of the simulation,

$$PO_p = \begin{cases} 0; & POP_p < POP_{crit} \\ 1; & POP_p \geq POP_{crit} \end{cases} \tag{2.28}$$

The prediction of daily total precipitation, P_p, is calculated conditional on precipitation occurrence of $PO_p = 1$. Then precipitation values are transformed using a temporal window of 5 days. Once again, a weighted least squares is used to account for elevation effects on precipitation. Then, the normalized difference of the precipitation observations, P_i, for any given pair of stations define the dependent variable, giving the regression of the form

$$\left(\frac{P_1 - P_2}{P_1 + P_2}\right) = \beta_0 + \beta_1(z_1 - z_2) \tag{2.29}$$

where β_0 and β_1 are the regression coefficients, z_1 and z_2 are the elevations recorded for two stations denoted as 1 and 2 of a unique pair.

The predicted daily total, P_p, is obtained as follows:

$$P_p = \frac{\sum_{i=1}^{n} W_i P_i PO_i \left(\frac{1+f}{1-f}\right)}{\sum_{i=1}^{n} W_i PO_i} \tag{2.30}$$

where $f = \beta_0 + B_1(z_p - z_i)$.

The form of prediction requires that $|f| < 1$. In Meteoland, a parameter $f_{max} = 0.95$ by default is introduced to force $|f| = f_{max}$ whenever $|f| > f_{max}$.

2.2.3 Parameterization

All the parameters for Meteoland reported in this thesis are given in the following table:

Parameter	Units	Description	Climatic variable	Value
I	none	Number of station density iterations	All	3
R	m	Truncation radius	All	140,000
α	none	Gaussian shape parameter	Tmax Tmin Prec	3 3 5
N	none	Average number of stations with non-zero weights	Tmax Tmin Prec	30 30 20
S_T	days	Temporal smoothing width for elevation regressions	Tmax Tmin Prec	15 15 5
POP_{crit}	none	Critical precipitation occurrence	Prec	0.52
f_{max}	none	Maximum value for prec regression extrapolation	Prec	0.95

2.3 Random Forest

Random forest is a tree-based supervised learning algorithm that can be applied to both classification and regression problems. It is an ensemble of K tree predictors $\{T_1(X), \ldots, T_K(X)\}$, where $X = \{x_1, \ldots, x_p\}$ is a p-dimensional vector of predictors (Svetnik *et al.*, 2003).

The ensemble produces K outputs, $\{\hat{Y}_1 = T_1(X), \ldots, \hat{Y}_K = T_K(X)\}$, where $\hat{Y}_k$ with $k = 1, \ldots, K$, is the prediction given by the kth tree. These outputs are aggregated for the final predicition according to the type of problem.

In regression, the final prediction is formed by taking the average predictions in the following:

$$\hat{Y}(X) = \frac{1}{K} \sum_{k=1}^{K} T_k(X) \tag{2.31}$$

where k is the individual bootstrap sample; K, the total number of trees and $T_k(X)$, the individual learner.

The mathematical formulation of RF puts emphasis on training the algorithm iteratively until a strong learner is produced (Breiman, 2001). The ensemble methods train multiple learners, weak and strong to solve the same problem. There exist two well-known methods of classification trees, boosting and bagging, that generate many classifiers and aggregate their results. Random forest adds an additional layer of randomness to bagging. While each tree is independently constructed using a bootstrap sample of the dataset in bagging, random forest randomly chooses the best split amongst a subset of predictors in each node.

Given a training dataset $\{(X_i, Y_i)\}$ with $i = 1, \ldots, n$, as explained by Segal (2004) and Svetnik *et al.* (2003), the RF model procedure, illustrated schematically also in figure 2.3, is explained as follows:

- Randomly draw a bootstrap sample from the training dataset.
- Grow a tree from each bootstrap sample.
- At each node, specify the number of covariates, which has to be less than or equal to the number of predictors p and choose the best split based on these covariates.

- Each tree is grown to the maximum depth, until no further splits are possible.

- Repeat the steps until an adequate large number of trees is grown.

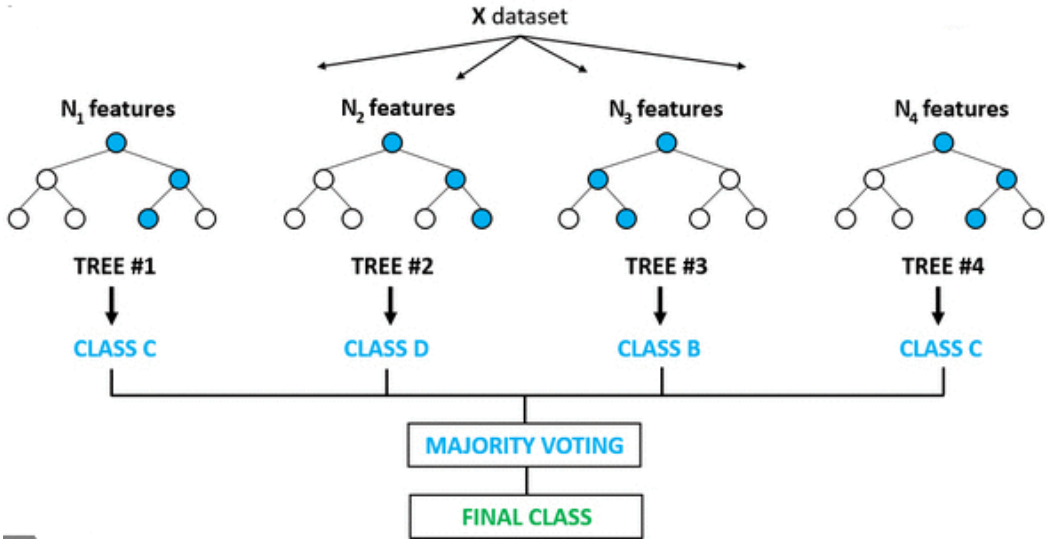


Figure 2.3 Illustration of the random forest model split. The samples are drawn from the training dataset and trees are grown from the features (co-variables) selected. Only the best splits from each tree are taken into consideration to form the output which is collected in the rectangle of final class.

Random forest depends on two parameters, the number of trees in the forest and the number of variables at each node in the random subset (Breiman, 2001). To ensure good performance of the model, the number of trees needs to grow with the number of predictor variables (Svetnik *et al.*, 2003).

Two useful pieces of information on the sample of the dataset is produced by the model. It measures the importance of the variables and gives a measure of the proximity that defines the internal structure of the data. To achieve stable estimates of these two above-mentioned measures, a sufficient number of trees is necessary. Important terms that are used through the process, are:

- Entropy,
  measures the randomness or unpredictability in the dataset,

$$\sum_{i=1}^{C} -f_i log(f_i) \tag{2.32}$$

where, f_i is the frequency of label i at a node and C is the number of unique label.

- Information gain,
  measures the decrease in entropy after splitting the data on an attribute,

$$Gain(T, X) = Entropy(T) - Entropy(T, X) \tag{2.33}$$

  where T is the target variable, X the feature to be split on and $Entropy(T, X)$, the entropy calculated after the data is split on feature X.

2.3.1 Parameterization

For this study, training the random forest model was done including four co-variables: slope, aspect, distance to coast and elevation in addition to the spatial coordinates of the weather stations. Random forest depends on two parameters, mtry which takes the number of variables that are randomly sampled at each node, and ntree that defines the number of trees used to grow the forest.

In the implementation applied here, the parameter ntree employed to grow the ensemble of trees is set to be equal to 500 in order to train the model. While tuning ntree, it was observed that from about 200 trees there was no more decrease in error, therefore 500 trees can be considered as enough to fit the model (Figure 2.4).

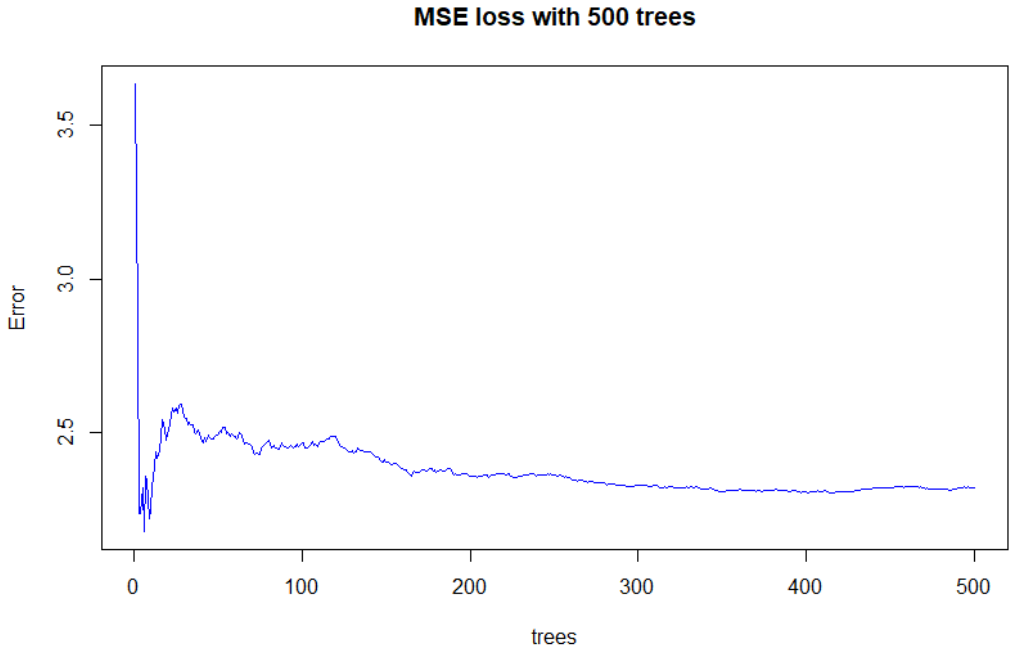


Figure 2.4 The figure shows the decrease in mean squared error (MSE) on the y-axis as the number of trees increase on the x-axis.

However to make sure that the ntree selected does not represent a small number of trees, ntree up to 3000 was checked. There was not much difference between choosing 500 trees or 3000 trees (Figure 2.5). If a considerable number of trees is selected, every input row finds an opportunity to get predicted at least a few times.

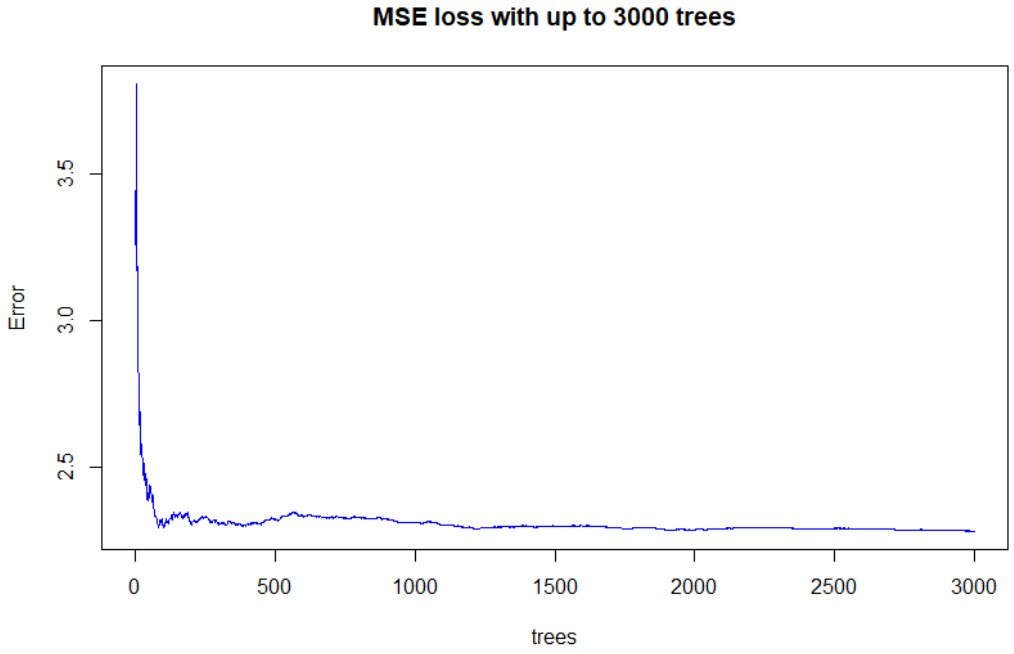


Figure 2.5 The figure shows the decrease in mean squared error (MSE) on the y-axis with increase in the number of trees on the x-axis. Not much change in error is observed over 500 trees.

Using the random forest package, the variables that are important to the prediction were checked. The spatial coordinates were considered for all the models, the co-variables that have greater influence on the predictions, including elevation, distance to the coast, slope and aspect were analysed. The importance is displayed at a scale ranging from 0 to 1. Elevation is the co-variable with the highest importance to the prediction followed by the distance to the coast in Figure 2.6.

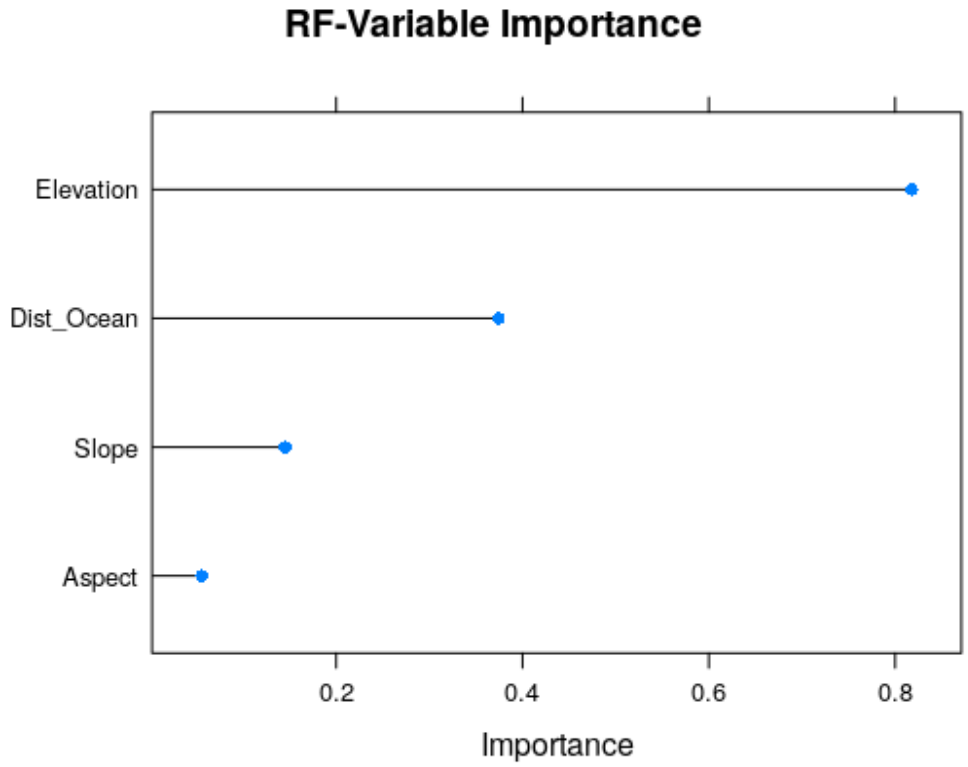


Figure 2.6 The importance, ranging from 0 to 1, of the four non spatial co-variables used in the predictions. The blue dot shows the importance value for each covariable.

2.4 Artificial Neural Network

Artificial neural networks offer methodological advantages over traditional spatial analysis methods. They are well suited to processing noisy data and handling non-linear modelling tasks. There are no critical assumptions with neural networks about the nature of spatial data. A neural network model is a computer model whose architecture essentially imitates the learning capabilities of a human's brain (Yeh *et al.*, 2013). Artificial neural networks are used to make accurate predictions for highly non-linear systems because of their capacity to approximate non-linear relations and their derivatives without knowing the true non-linear function (Joshi, 2016) .

The function of a neural network is determined by the model of the neurons, the network structure and the learning rate (Zhou, 2012). A number of possible network structures exist, however the multi-layer-feed-forward network is the most commonly used. Amongst these network structures, Multi-Layer Perceptron (MLP) and radial basis function network (RBFN) are the two most well-known neural networks for spatial interpolation (Salcedo-

Sanz *et al.*, 2016; Seo *et al.*, 2015; Yeh *et al.*, 2013).

For this study, MLP, a feed-forward neural network which is structured in layers of neuronal units interconnected by weighted links is selected. Each layer of an MLP is fully connected to the following layer and passes signals from all its neurons to each neuron in the following layer.

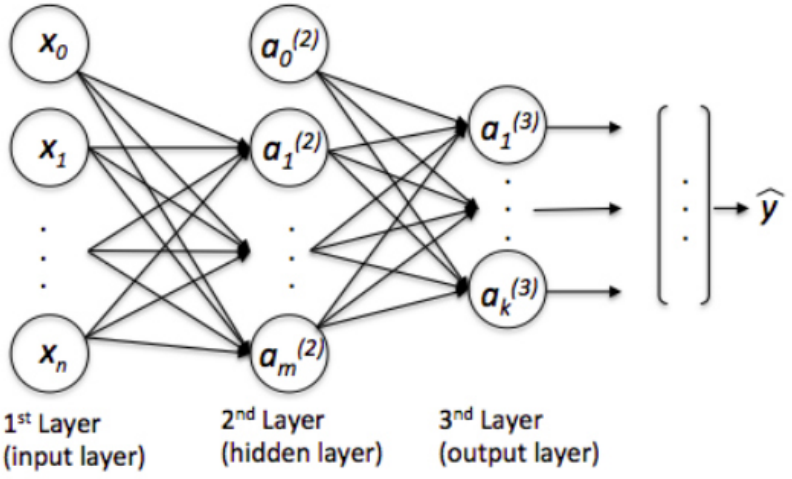


Figure 2.7 An artificial neural network architecture with 3 layers, an input layer with n neurons, a hidden layer with m neurons and an output layer with k neurons.

A neural network (Figure 2.7) has 3 types of layers: an input layer, which accepts input signals from outside. The independent variables used for prediction are set in this layer and represent its neurons. Hidden layers that can be one or more layers, are represented by neurons that detect the features found hidden in the inputs. Any continuous function can be represented by only one hidden layer and more than one hidden layer can represent continuous and even discontinuous functions. An output layer, in turn, accepts output signals from the hidden layer, where the number of outputs equals the number of dependent variables to be predicted.

The mathematical formulation of an MLP can be written in the following form:

$$\hat{Y} = f\left(\sum_{j=1}^{m} w_{kj} f\left(\sum_{i=1}^{n} w_{ji} x_i + b_j\right) + b_k\right) \tag{2.34}$$

where $\hat{Y}$ is the output matrix that contains the predicted values; f is an activation function, which is a non-linear complex function that converts the input signal to an output signal; w_{kj} defines the weights associated to the connection between hidden and output layers while w_{ji} represents the weights connecting the input and hidden layers; m and n are respectively the number of neurons in hidden and input layers; x_i represents the input variables given to the input layer; b_j and b_k are the bias of the neurons in hidden and output layers respectively.

The weights in equation 2.34 are adaptively changed to minimize the difference between the desired output and the actual output (Abutaleb, 1991). This is a time consuming process computed by trial and error due to the non-existence of a method to determine the optimal number of neurons in the hidden layers. Such computation is essential to determine the values of connection weights and the biases of the neurons. Once decided, the goal of training the algorithm is achieved. The learning process is done using a gradient descent optimization method that updates the weights and bias in the network as long as the activation function is differentiable. A gradient descent is an iterative algorithm, that starts from a random point on a function and descends its slope in steps until it reaches the lowest point of the function. The most commonly used learning algorithm for neural networks is called back-propagation (Abutaleb, 1991; Beltratti *et al.*, 1996; Pascanu *et al.*, 2013).

The back-propagation algorithm consists of taking the error calculated from the comparison of the desired output and the ouput obtained with the feed-forward network. The error obtained is then back propagated to the hidden layer, while updating the weights and biases, in order to minimize the error (Beltratti *et al.*, 1996; More and Deo, 2003; Zhou, 2012). This process is repeated until the training error is minimized.

2.4.1 Activation function

While training the model, there exists different activation functions denoted as f in equation 2.34 that can be tested to ensure that the representation in the input space is mapped to a different output space (Pascanu *et al.*, 2013). The following four most commonly used activation functions were implemented in the study reported in this thesis.

- Sigmoid or logistic function (logistic; Fig. 2.8), which ranges between 0 and 1 and therefore is not zero centred and can make optimization harder. It is in the form of

$$f(x) = \frac{1}{1 + e^{-x}} \tag{2.35}$$

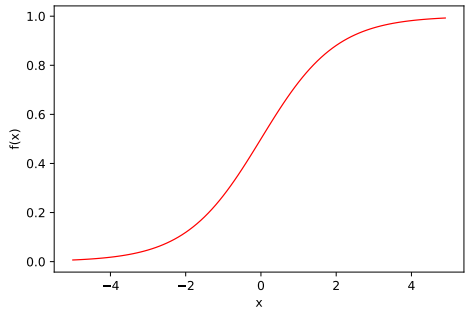


Figure 2.8 Sigmoid function

- Hyperbolic tangent function (Tanh; Fig. 2.9), a zero-centred function that ranges between -1 and 1 and is in the form of

$$f(x) = \frac{e^{2x} - 1}{e^{2x} + 1} \tag{2.36}$$

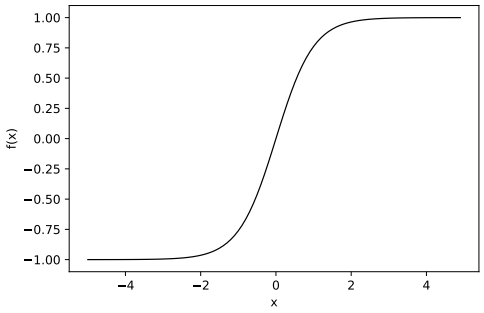


Figure 2.9 Tanh function

- Linear function (linear; Fig. 2.10), that ranges from $-\infty$ to $+\infty$ is in the form

$$f(x) = ax \tag{2.37}$$

where a is a positive integer

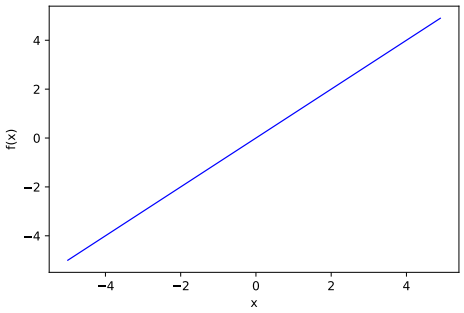


Figure 2.10 Linear function

- Rectified Linear units (ReLu; Fig. 2.11), which is the most used activation function given its fast convergence, is written as

$$f(x) = max(0, x) \tag{2.38}$$

$$if \begin{cases} x < 0 & f(x) = 0 \\ \text{otherwise} & f(x) = x \end{cases}$$

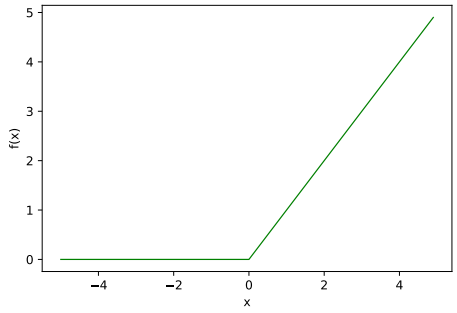


Figure 2.11 ReLu function

2.4.2 Structuring the network

The parameters of a neural network such as the number of hidden layers and of neurons in the hidden layers are not quantified for any particular application (FrontlineSolvers, 2019). Although, a choice of many hidden layers in a network can be made, any continuous function can be approximated using a feed-forward network with only one hidden layer (Zhou, 2012). Many hidden layers can introduce divergence in the network, preventing the network from converging on a stable state (Zhou, 2012). This

suggests the use of a more complicated algorithms to prevent such a problem. Other important parameters that need to be determined while training a neural network model exist, such as:

- A batch size, which is a hyper-parameter indicating the number of samples the model needs to work through before updating the internal model parameters. This parameter is given by an integer that can be set at 1 or more than 1 but cannot exceed the number of samples of the training dataset. If

  - batch size $= 1$, the batch is a stochastic gradient descent (sgd) which randomly picks one data point from the entire dataset at each iteration to decrease the error. A stochastic gradient descent is a method to find the optimal parameter configuration of the network.

  - $1 <$ batch size $<$ training dataset size, is a mini-batch gradient descent. A mini-batch tries to find a balance between the goodness of gradient descent and speed of sgd.

  - batch size $=$ training dataset size, is a batch gradient descent or full-batch gradient descent. It is less efficient than mini-batch gradient descent.

- An epoch, is a hyper-parameter that gives the number of times the learning algorithm needs to go through the entire dataset. This hyper-parameter is an integer that varies from 1 to infinity, and a high epoch number allows the model to minimize the error.

- A loss function, which learns to decrease the prediction error by adjusting the weights and biases to match the target output.

2.4.3 Normalization of the dataset

In this study, prior to training, the training dataset is normalized which improves the learning in a neural network (Sola *et al.*, 1997). The inputs are scaled in the range of 0 to 1. Normalization of the dataset is one of the most important preprocessing steps while working with neural networks (Raad *et al.*, 2012). The normalization formula is in the form

$$x_{ijs} = \frac{x_{ij} - x_{im}}{\sigma_{ix}} \tag{2.39}$$

where x_{ijs} is the standardized value of variable x_i for weather station j; x_{ij} is the value of variable x_i for weather station j; x_{im} the mean value of the

variable x_i and σ_{ix} the standard deviation of the variable x_i.

Two architectures are built. One for maximum temperature and minimum temperature together, which gives two neurons in the output layer. Another architecture for rainfall prediction with one neuron in the output layer. The neurons in the input and hidden layers are tested and the best architecture is considered.

2.4.4 Parameterization

For multilayer perceptron, prediction of maximum and minimum temperature is done by changing the hyper-parameters by trial and error. The best selections for each dataset are shown in the table below:

Dataset	Hidden layer	Neurons	Input	Variables	Activation function	Optimizer
Jan 2016 MaxT, MinT	1	18	6	All	ReLu	sgd
Jan 2016 Rain	1	7	6	All	linear	sgd
July 2016 MaxT, MinT	1	9	6	All	ReLu	sgd
July 2016 Rain	1	8	6	All	linear	sgd
Jan 2017 MaxT, MinT	1	140	6	All	ReLu	sgd
Jan 2017 Rain	1	56	3	x,y,z	ReLu	adam
July 2017 MaxT, MinT	1	14	6	All	ReLu	sgd
July 2017 Rain	1	5	3	x,y,z	ReLu	sgd

In the Variables column, 'All' represents x coordinates, y coordinates, elevation, slope, aspect and coastal distance. z, represents elevation.

In the optimizer column, sgd represents a stochastic gradient descent algorithm and adam an adaptive moment algorithm. Adam uses the squared gradients and the estimations of first and second moments of gradient to scale the learning rate for each weight of the neural network. An n^{th} moment of a random variable is the expected value of the random variable to the power of n.

2.5 Accuracy assessment

Evaluating the model accuracy is an essential part in describing how well the model performs in its prediction. In this study, to assess the performance of the four models, a random split of the dataset was performed by randomly assigning 70% of the data to a training sample and the remaining 30% to a testing sample. The procedure is called cross-validation and it is

performed to evaluate the ability of the model to predict unseen data (testing sample). We used a 10-fold cross-validation which ensures that observations from the original dataset are given a chance of appearing in training and testing set (Kohavi *et al.*, 1995). The entire dataset was randomly split into 10 folds, where a single fold was retained as the validation data for testing the model and the remaining 9 folds were used as training data. The process was repeated 10 times, with each of the 10 folds used exactly once as the validation data. Then, the results were averaged to produce the estimation.

The procedure is different for the MLP models, for which each dataset was randomly split into training, validation and testing sets allocating 80% to training and validation, while 20% to testing. Prior to running the model, the 80% was then split into 75% for training and 25% for validation. The process was repeated 10 times and the results were averaged in a single estimation.

Three types of goodness-of-fit measure were considered in evaluating the accuracy of the models, including Root Mean Squared Error (RMSE), which places a lot of weight on large errors; Mean Absolute Error (MAE), which is less sensitive to extreme values compared to RMSE and gives an indication of the error extent; and R^2 was assessed to evaluate how well the predicted values fit compared to the original values.

- Root Mean Squared Error (RMSE) is an error rate of a model obtained by the square root of the average of squared differences between predicted and observed values,

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(p_i - a_i)^2}{n}} \tag{2.40}$$

  where a is the actual target, p the predicted target and n the total number of observations.

- Mean absolute error (MAE) represents the difference between predicted and observed values extracted by averaged the absolute difference over the dataset,

$$MAE = \sum_{i=1}^{n} \frac{|p_i - a_i|}{n} \tag{2.41}$$

  the variables are the same as in equation 2.40

- R^2 or coefficient of determination summarizes the explanatory power of the model,

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} \tag{2.42}$$

where,

$$\begin{aligned} SST &= \sum_{i=1}^{n}(a_i - \bar{a})^2, \quad \text{sum of square total} \\ SSR &= \sum_{i=1}^{n}(p_i - \bar{p}), \quad \text{sum of square prediction} \\ SSE &= \sum_{i=1}^{n}(a_i - p_i)^2, \quad \text{sum of square error} \end{aligned}$$

with $\bar{a}$ is the mean value of a and $\bar{p}$ the mean value of p. The other variables a, p and n are the same as in equation 2.40

2.6 Study Area

This study was carried out in the KwaZulu-Natal province of South Africa located in the south-eastern part of the country and which extends over a total geographical area of approximately $94,360\text{km}^2$ (Figure 2.12). KwaZulu-Natal is bordered by the Indian Ocean to the east and the Drakensberg Mountain escarpment to the west producing a warm, subtropical climate with inland regions becoming progressively colder. Situated in the Southern Hemisphere, the summer season is from about November to February, and is hot, averaging 28°C. It gets most of the rain during the summer season (Schulze *et al.*, 2007). Autumn is from March to May, where temperatures begin to cool before the warm, dry and clear Winter season that is from June to August with average temperatures of 23°C.

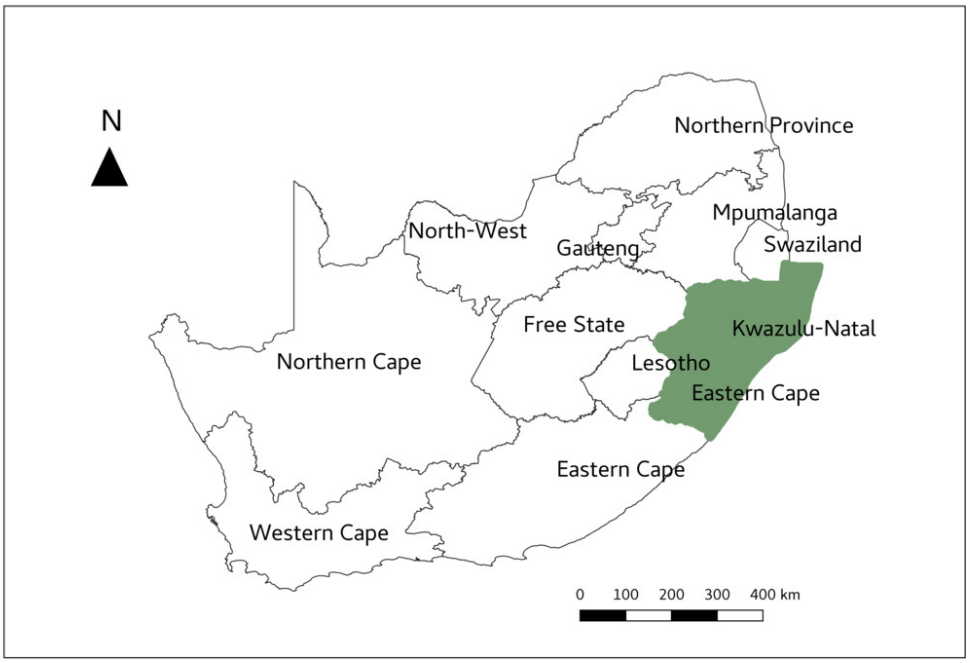


Figure 2.12 South Africa map with all its provinces. KwaZulu-Natal province, the region of interest to this study, is highlighted on the map with a dark green colour.

2.7 Data source and collection

This section covers the source for the co-variables, the relation between primary and secondary variables employed to run the models and the co-variables used for each model.

2.7.1 Co-variables

For the modelling in this study, a Digital Elevation Model (DEM) at 30m resoultion was used (Figure 2.13). The DEM was sourced from the contour 20m provided by the Department Land Affairs. A DEM is a mathematically-derived representation of the relief of the Earth's surface. This was re-sampled to 1km grid cells for the purpose of this study. Elevation, slope, aspect and distance to the coast were then calculated from the DEM in the quantum geographic information system (QGIS).

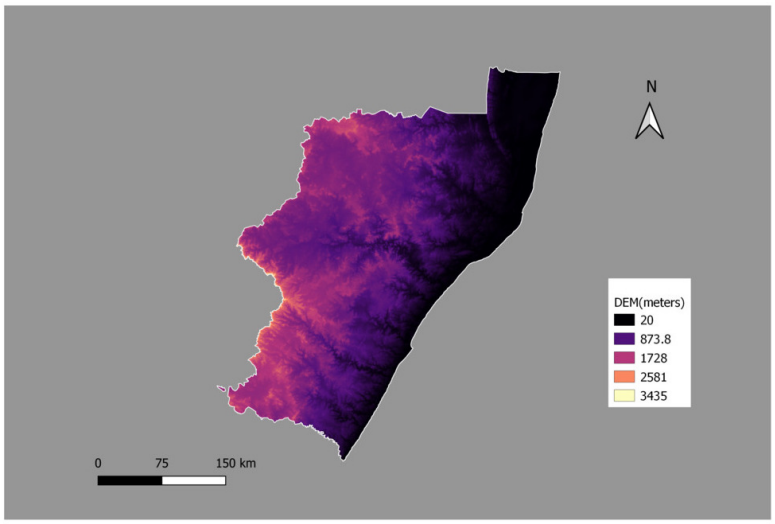


Figure 2.13 Digital Elevation Model (DEM) of KwaZulu-Natal at a 30m resolution

The spatial coordinates, longitude and latitude were converted into UTM coordinates and this was applied for all the models. Figure 2.14 shows the correlation between the features. A strong positive correlation of above 0.6 was found between minimum and maximum temperatures. There was a strong negative correlation observed between longitude and elevation. A considerable correlation ranging from -0.3 to -0.6 was observed between maximum temperature and distance to coast as well as between maximum temperature and elevation. Similar observations are found for the minimum temperature.

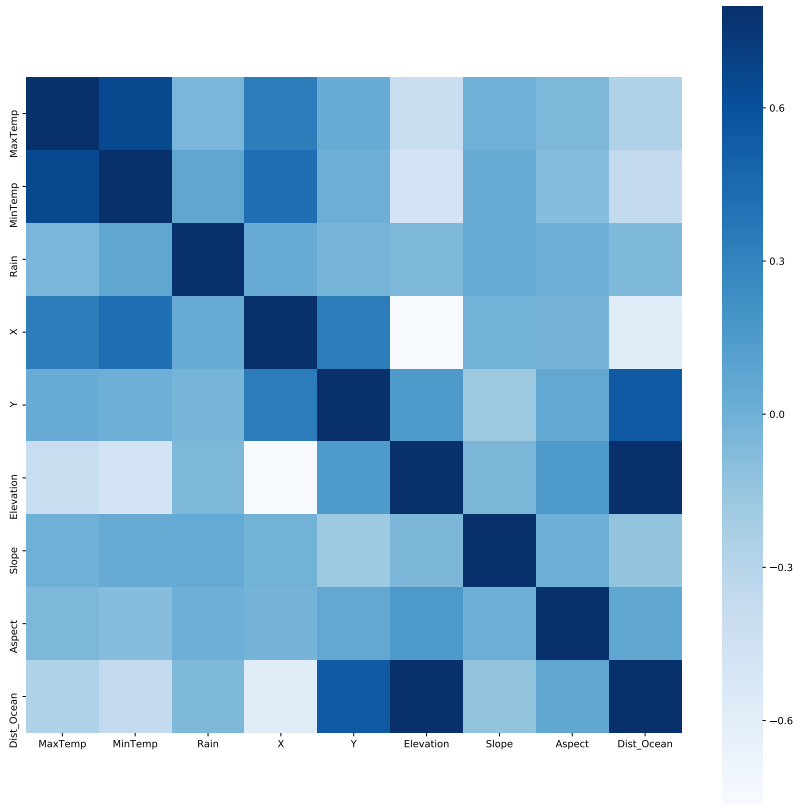


Figure 2.14 Correlation between the features used for interpolation. A strong positive correlation is given by a darker blue colour and a strong negative correlation is given by a lighter blue colour.

In this study, the choice of co-variables varies with the method of interpolation given the role of each co-variable in the increase of the accuracy and effect on temperature and precipitation predictions. Elevation was considered for all four methods given its strong influence on climate. Niekerk *et al.* (2011) has found distance to coast to simulate the effect Ocean has on climate, by improving the accuracy for spatial interpolation. Thus, this co-variable was used for the three other methods except for GF which, through Meteoland, does only incorporate elevation, slope and aspect (De Cáceres, 2019). Slope and aspect are mostly used to allow topographical effects on weather prediction however the use of a linear variogram for KED led to

not consider slope and aspect.

The following table gives the co-variables considered for each method:

Methods	Co-variables
KED	Elevation, coastal distance
GF	Elevation, slope, aspect
RF	Elevation, slope, aspect, coastal distance
MLP	Elevation, slope, aspect, coastal distance

2.7.2 Climatic data

For the analyses in this thesis, weather station observations from three sources were used. Minimum and maximum air temperatures as well as precipitation data of the 15th day of two months, January and July, representing respectively summer and winter of two years, 2016, the driest, and 2017, the wettest are selected in order to achieve the objective of comparing spatial interpolation methods on a specific day.

- Data from 69 weather stations were provided by the South African Sugarcane Research Institute (SASRI). This network of weather stations is relatively new and has been established to assist with modelling and managing sugarcane plantations. Data is accessible from (sasri/institute).

- The VitalWeather System, that gives live weather information to industry sectors, provided with 133 weather stations data for this research. The VitalWeather System is a South African locally developed system and uses the Davis Vantage Pro2 weather station as a source of their weather data. This system was primarily designed for fire danger assessment and is not always reliable for detailed modelling work (VitalWeather).

- Data from 34 weather stations were given by the South African Weather Service (SAWS). SAWS is a member of the World Meteorological Organization and is the meteorological service under the Department of Environmental Affairs and Tourism of South Africa as explained in the website (SawsService).

Stations located at a long distance away from the KZN region were removed to avoid an introduction of bias in the result. Only stations within and nearby KZN with the presence of data on the 15th of January and July in the years of interest were considered. Thus 92 weather stations were used for this study (Figure 2.15). The choice of these 92 stations was motivated by the spatial variability of their distribution allowing a presence of spatially complex daily weather data.

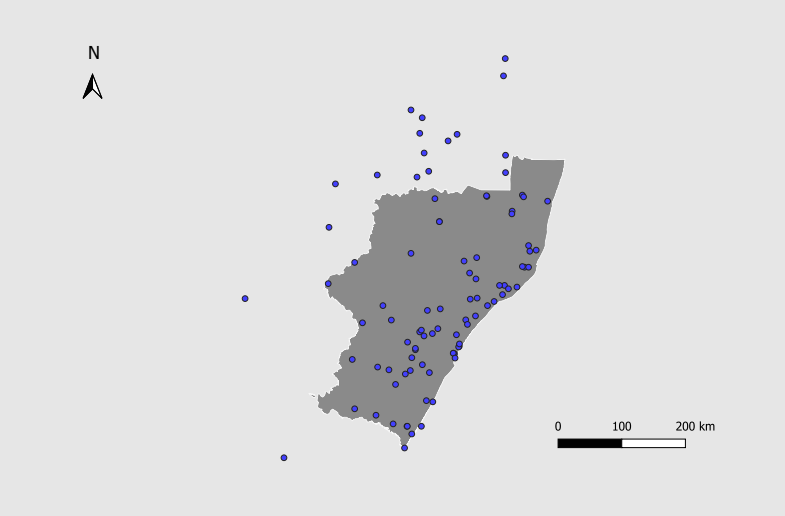


Figure 2.15 Spatial distribution of 92 weather stations used for each model in this study

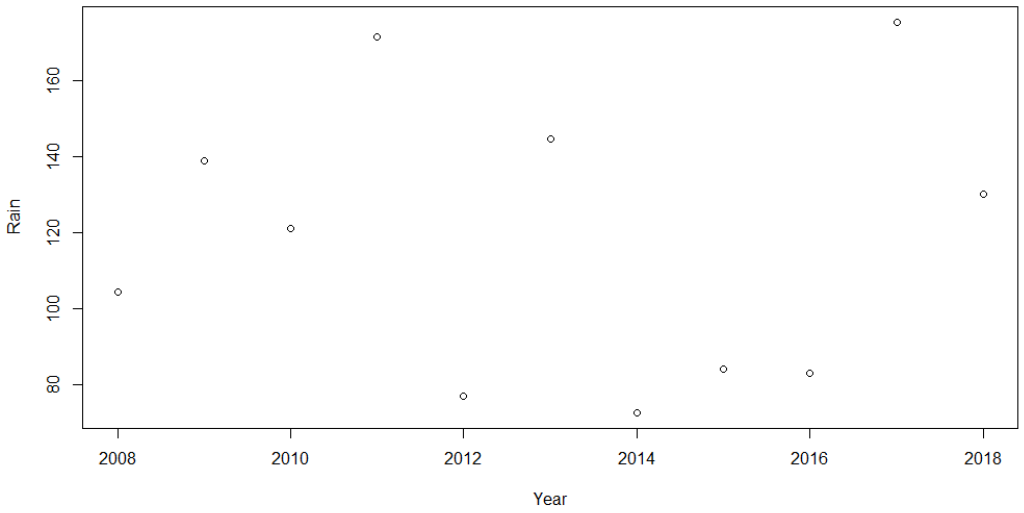


Figure 2.16 Total annual rainfall (in mm) for the period 2008 to 2018

The driest years were actually 2014, and then 2012 but they also happen to have a large number of missing weather stations with respectively 48% and 72% while 2016 was the third driest year during the 10 year period and had substantially more weather stations. In 2016, 97% of the weather stations were present.

2.8 Python and R packages

All the modelling was done using two open source programming languages, R and Python. Kriging with external drift and Multilayer perceptron were computed in Python using different python libraries while Gaussian filter and Random forest were computed using several R packages.

In R, the packages *raster* and *sp* were used for the analysis of shapefiles. The packages *randomForest*, *ranger*, *caret*, *custom* were used to run random forest models, and *sp* and *meteoland* were used for Meteoland model analyses. In Python, *pykrige* package was used for Kriging with external drift analyses. *Sci-kitlearn* and *keras* were used to analyse Multilayer perceptron.

QGIS, an open-source cross-platform desktop geographic information system was employed for generating the maps to spatially view the predictions made with the models.

Chapter 3

Results and discussion

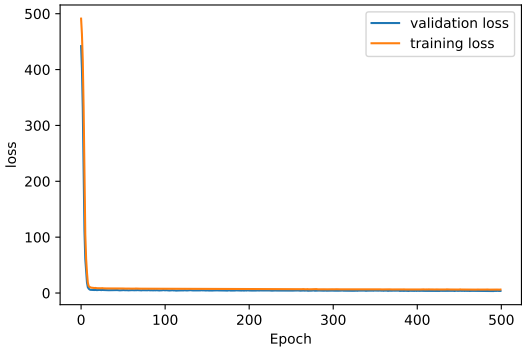
3.1 Analysis from the models

As mentioned in section 2.4 and explained by Kumar *et al.* (2005), there is no rule to find the best MLP model, this is found with trial and error. The plots obtained while training maximum and minimum temperature confirm that the training data is not overfitted because the validation loss align with the training loss. Hence, they produce a reduced mean squared error (MSE). However, for rainfall which represents different patterns, we do not get a good fit of the data. The small size of the training datasets is the reason of such results as found by Anctil *et al.* (2004). Moreover, if the values to train are around 0 which is the case of the actual rainfall values of July, where there was nearly no rainfall, we found similar observations of slow convergence and inefficient learning process as found by (Sajikumar *et al.*, 1999). The learning process would have been different with variation in the dataset in terms of values and time which implies a wider range of values to train the model and the change in time if time was considered as a co-variable.

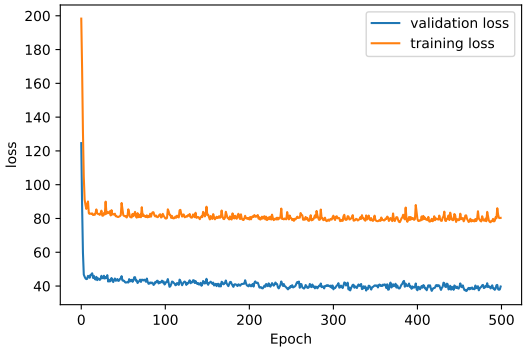
Figures 3.1 and 3.2 illustrate the best models selected for MLP. Training and validation losses of maximum and minimum temperatures are observed in the plots (a) and (c) that give two outputs in the network architecture since the two weather variables were combined in fitting the model. The plots (a) and (c) produce a reduced mean squared error (MSE) and give nicely curved plots. On the other hand, the plots (b) and (d) show the training and validation losses for rainfall. All these results were found with 500 number of epochs and variate activation functions and optimizers. Each caption indicates the parameters selected.

In this study, in figures 3.3 to 3.6 that illustrate the RF model selection employed for all the datasets, it is found that for the wet year, training rainfall does not require a large number of trees and many co-variables. Only

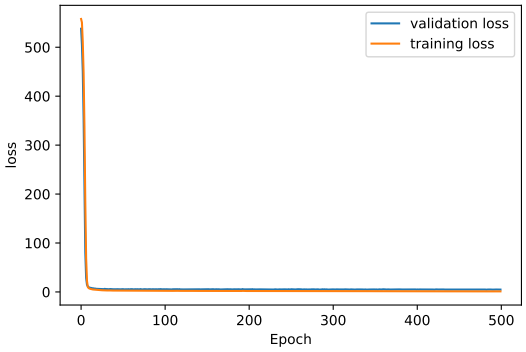
125 trees and one or two co-variables were mostly enough to achieve the best performance.



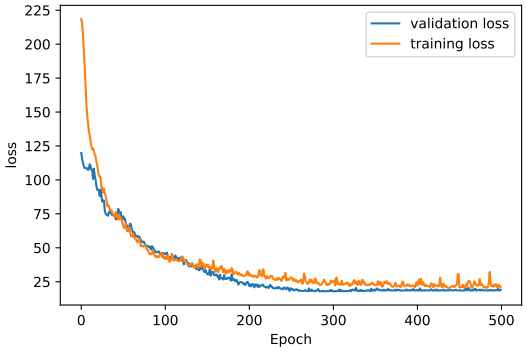
(a) Training and validation losses (MSE) obtained with MLP model using ReLu activation function and sgd optimizer with 18 neurons for maxT and minT



(b) Training and validation losses (MSE) obtained with MLP model using linear activation function and sgd optimizer with 7 neurons for rainfall

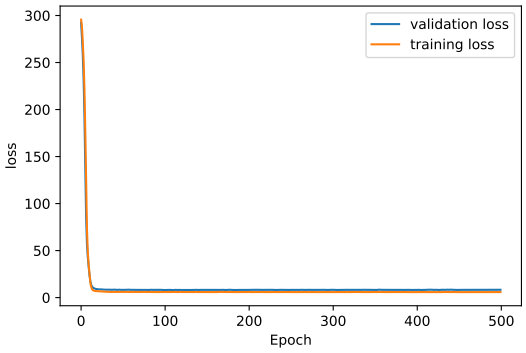


(c) Training and validation losses (MSE) obtained with MLP model using ReLu activation function and sgd optimizer with 140 neurons for maxT and minT

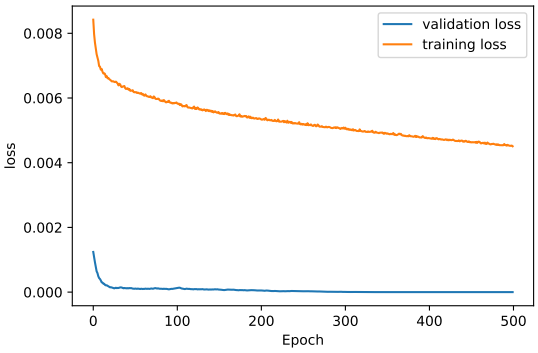


(d) Training and validation losses (MSE) obtained with MLP model using ReLu activation function and adam optimizer with 56 neurons for rainfall

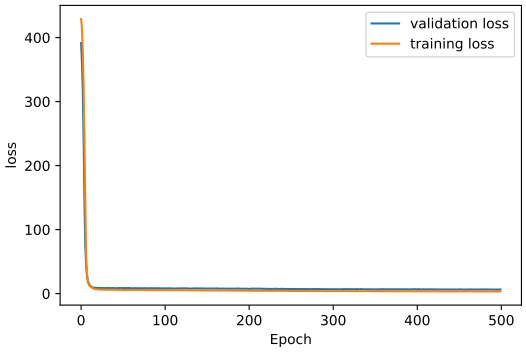
Figure 3.1 (a) and (c) show the training and validation losses of maximum temperature and minimum temperature for the 15th of January 2016 and 15th of January 2017 respectively while (b) and (d) show the training and validation losses of rainfall for the same dataset. One hidden layer, 500 epochs are used for all the models and 6 inputs for all except for (d) where 3 inputs are used.



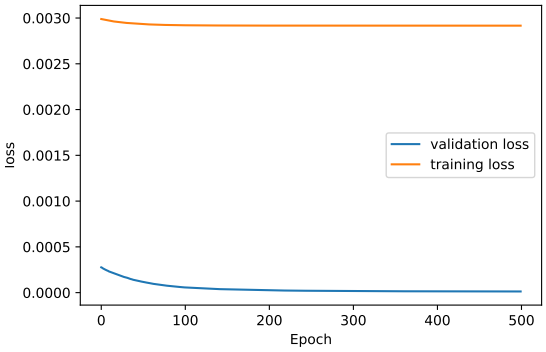
(a) Training and validation losses (MSE) obtained with MLP model using ReLu activation function and sgd optimizer with 9 neurons for maxT and minT



(b) Training and validation losses (MSE) obtained with MLP model using linear activation function and sgd optimizer with 8 neurons for rainfall

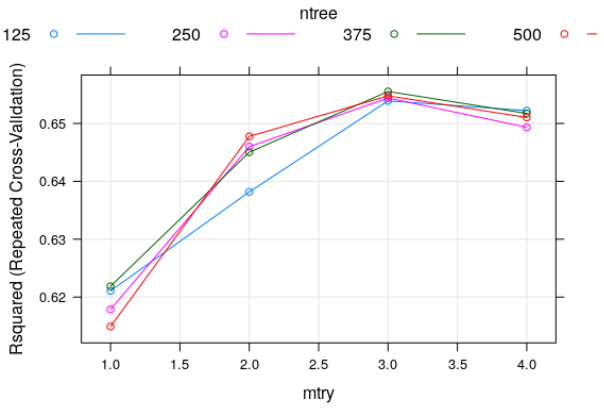


(c) Training and validation losses (MSE) obtained with MLP model using ReLu activation function and sgd optimizer with 14 neurons for maxT and minT

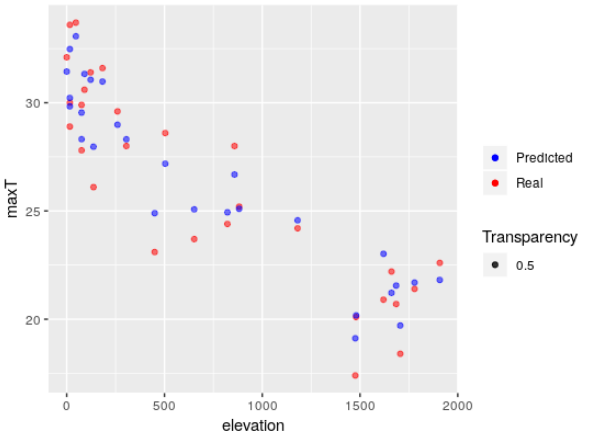


(d) Training and validation losses (MSE) obtained with MLP model using ReLu activation function and sgd optimizer with 5 neurons for rainfall

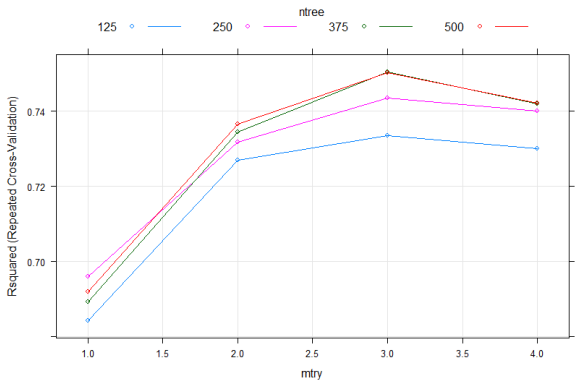
Figure 3.2 (a) and (c) show the training and validation losses of maximum temperature and minimum temperature for the 15th of July 2016 and 15th of July 2017 respectively while (b) and (d) show the training and validation losses of rainfall for the same dataset. One hidden layer, 500 epochs and 6 inputs are used for all the architectures except for rainfall that used 3 inputs.



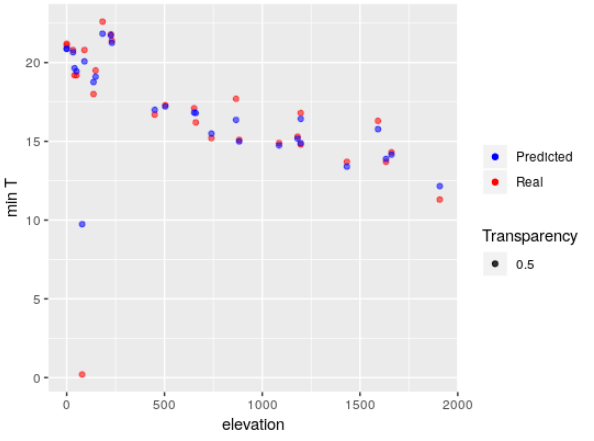
(a) Tuned parameters on R^2 (ntrees and mtry) with a custom RF model. The best set of parameters is given by ntree=375 and mtry=3



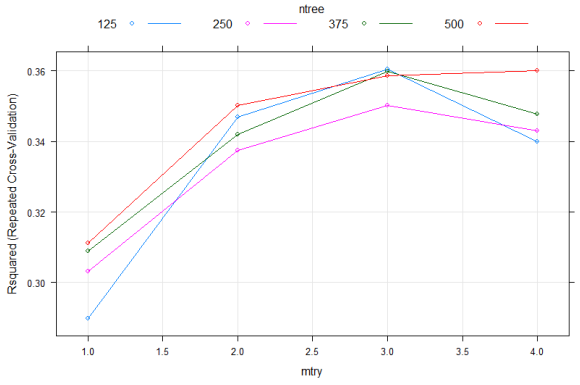
(b) Elevation versus maximum temperature (°C) for predicted maximum temperature values in blue observed values in red



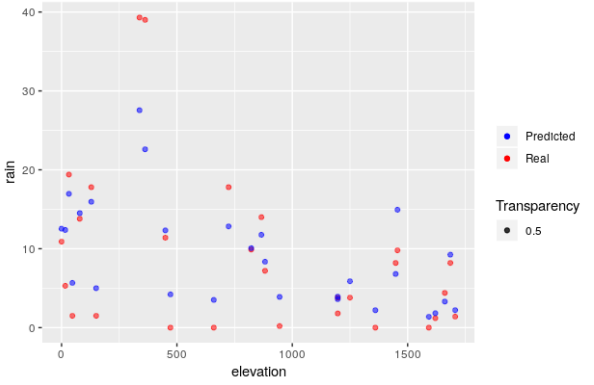
(c) Tuned parameters on R^2 (ntrees and mtry) with a custom RF model. The best set of parameters is given by ntree=375, 500 and mtry=3



(d) Elevation versus minimum temperature (°C) for predicted minimum temperature values in blue observed values in red

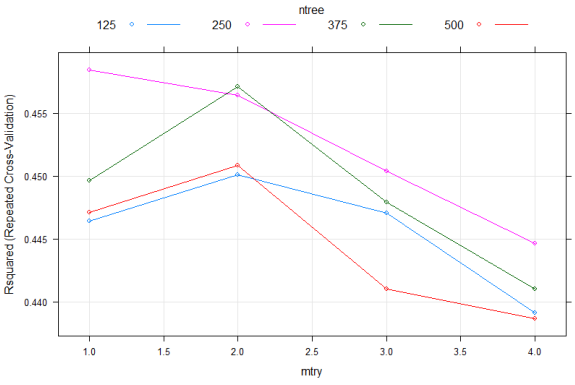


(e) Tuned parameters on R^2 (ntrees and mtry) with a custom RF model. The best set of parameters is given by ntree=125 with mtry=3 and ntree=500 with mtry=4

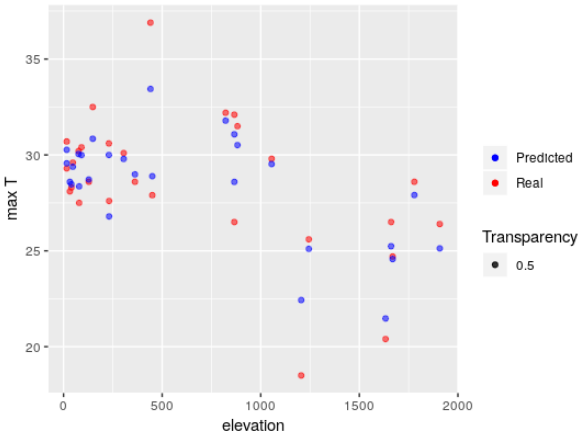


(f) Elevation versus rainfall (mm) for predicted rainfall values in blue observed values in red

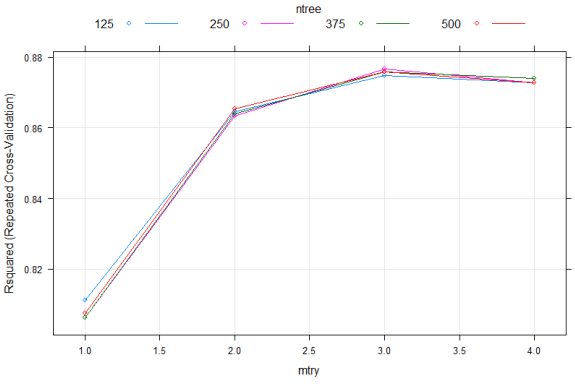
Figure 3.3 (a), (c) and (e) show the performance of RF model on the 15th January 2016 for maxT, minT and rainfall with different variations and combinations of the parameters. The tuned parameters are used in the final predictive models. (b), (d) and (f) show the actual and predicted values on 15th January 2016



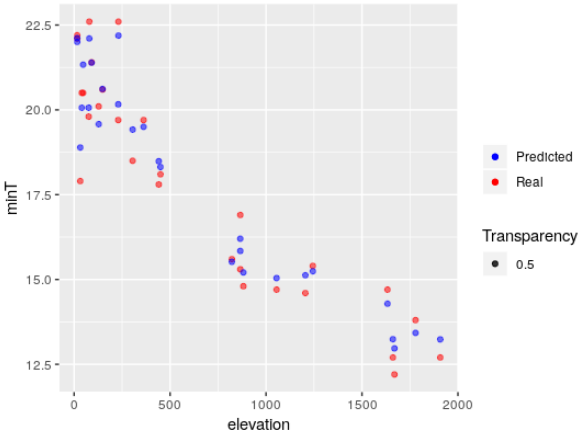
(a) Tuned parameters on R^2 (ntrees and mtry) with a custom RF model. The best set of parameters is given by ntree=250 and mtry=1



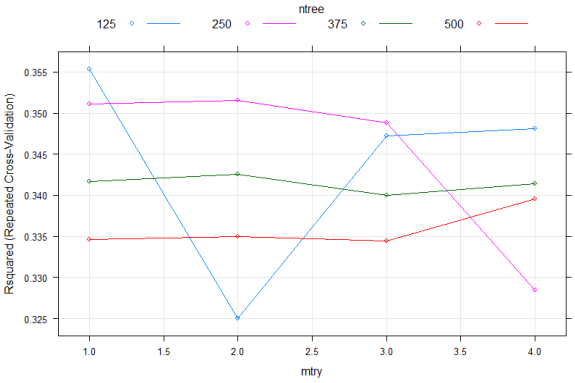
(b) Elevation versus maximum temperature (°C) for predicted maximum temperature values in blue observed values in red



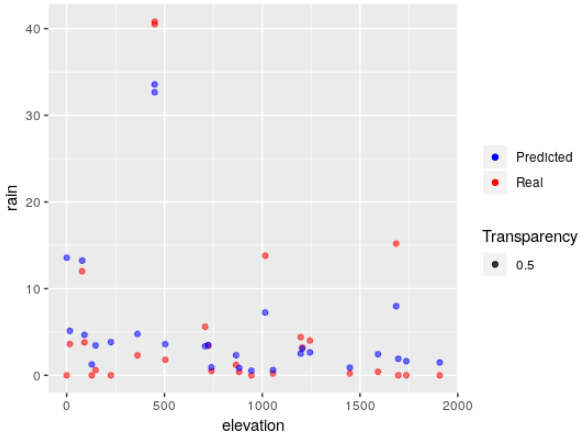
(c) Tuned parameters on R^2 (ntrees and mtry) with a custom RF model. The best set of parameters is given by ntree=250 and mtry=3



(d) Elevation versus minimum temperature (°C) for predicted minimum temperature values in blue observed values in red

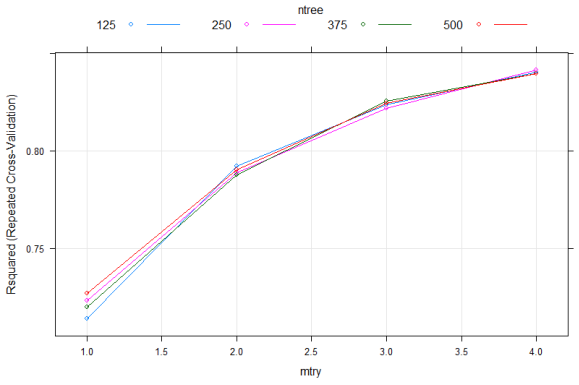


(e) Tuned parameters on R^2 (ntrees and mtry) with a custom RF model. The best set of parameters is given by ntree=125 and mtry=1

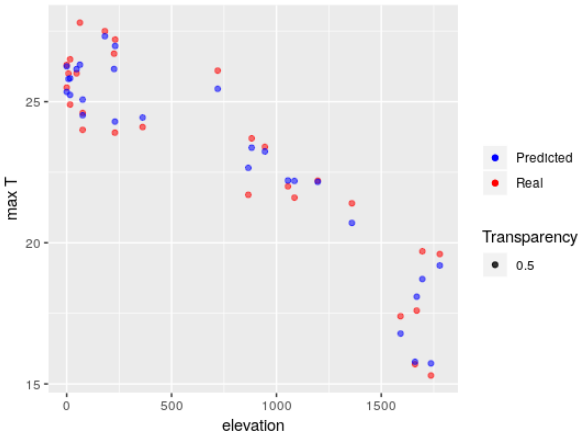


(f) Elevation versus rainfall (mm) for predicted rainfall values in blue observed values in red

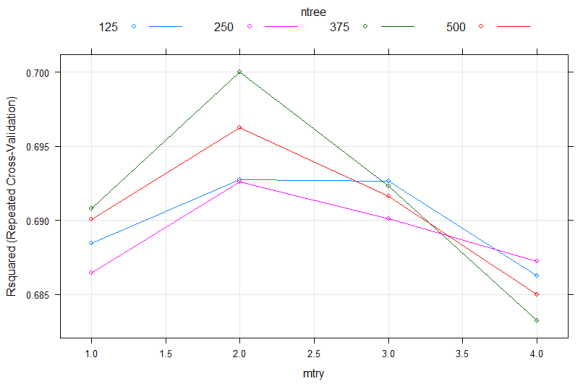
Figure 3.4 (a), (c) and (e) show the performance of RF model on the 15th January 2017 for maxT, minT and rainfall with different variations and combinations of the parameters. The tuned parameters are used in the final predictive models. (b), (d) and (f) show the actual and predicted values on 15th January 2017



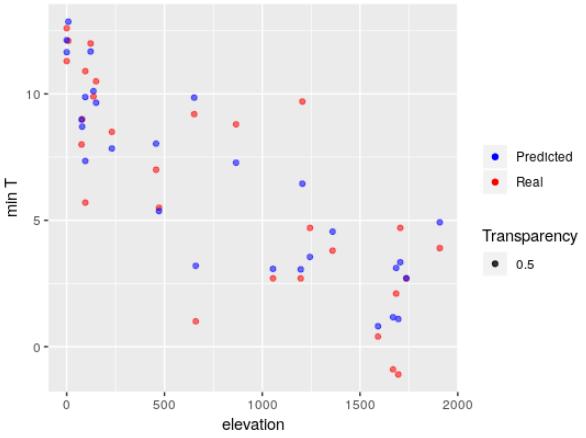
(a) Tuned parameters on R^2 (ntrees and mtry) with a custom RF model. The best set of parameters is given by any class of ntree but for mtry=4



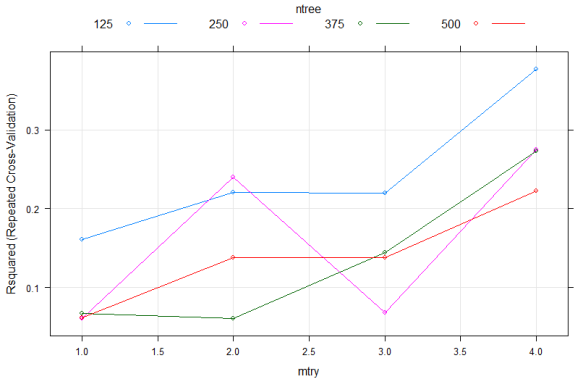
(b) Elevation versus maximum temperature (°C) for predicted maximum temperature values in blue observed values in red



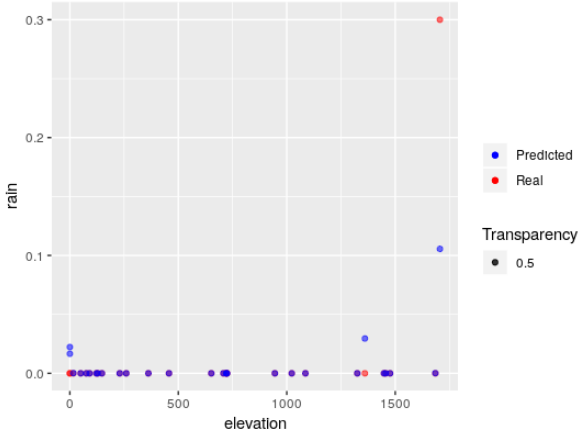
(c) Tuned parameters on R^2 (ntrees and mtry) with a custom RF model. The best set of parameters is given by ntree=375 and mtry=2



(d) Elevation versus minimum temperature (°C) for predicted minimum temperature values in blue observed values in red

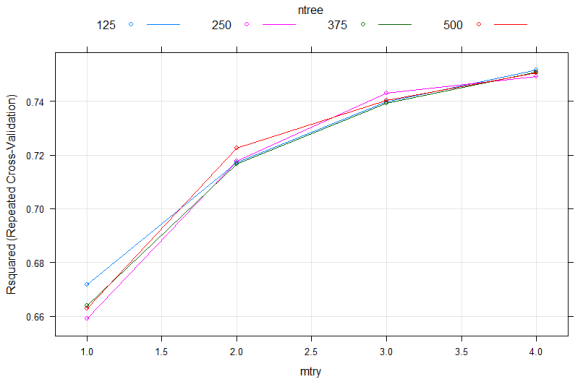


(e) Tuned parameters on R^2 (ntrees and mtry) with a custom RF model. The best set of parameters is given by ntree=125 and mtry=4

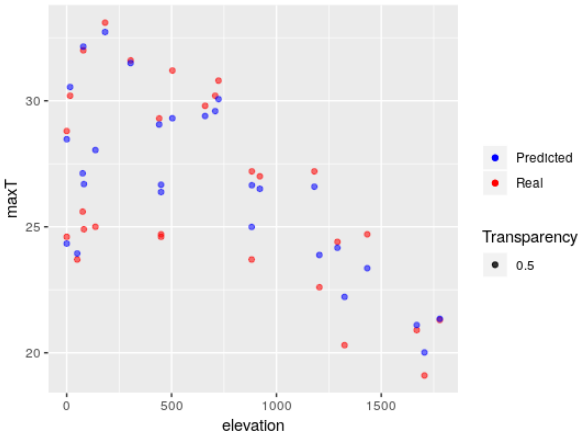


(f) Elevation versus rainfall (mm) for predicted rainfall values in blue observed values in red

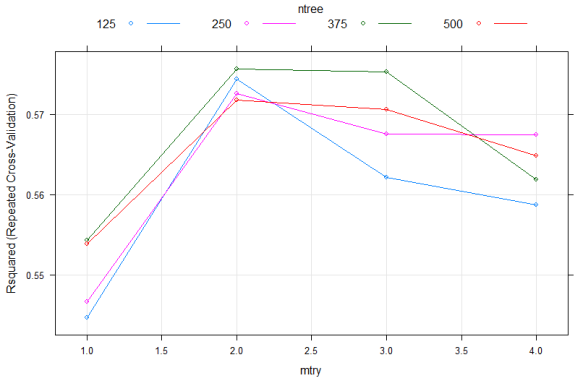
Figure 3.5 (a), (c) and (e) show the performance of RF model on the 15th July 2016 for maxT, minT and rainfall with different variations and combinations of the parameters. The tuned parameters are used in the final predictive models. (b), (d) and (f) show the actual and predicted values on 15th July 2016



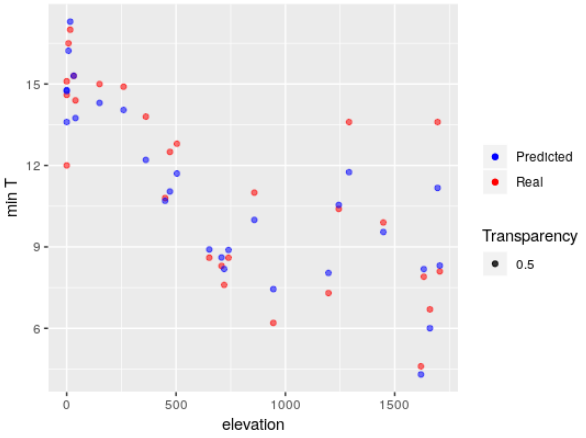
(a) Tuned parameters on R^2 (ntrees and mtry) with a custom RF model. The best set of parameters is given by mtry=4 and with all the values of ntree



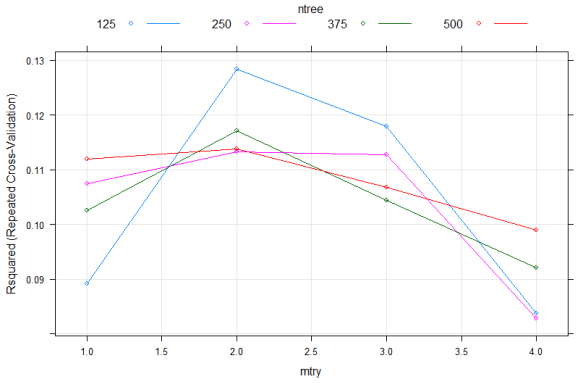
(b) Elevation versus maximum temperature (°C) for predicted maximum temperature values in blue observed values in red



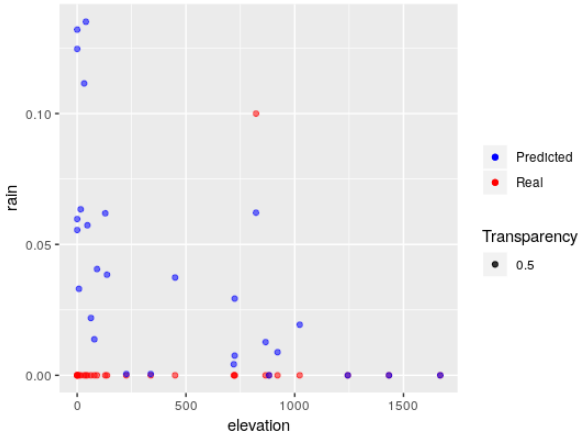
(c) Tuned parameters on R^2 (ntrees and mtry) with a custom RF model. The best set of parameters is given by ntree=375 and mtry=2



(d) Elevation versus minimum temperature (°C) for predicted minimum temperature values in blue observed values in red



(e) Tuned parameters on R^2 (ntrees and mtry) with a custom RF model. The best set of parameters is given by ntree=125 and mtry=2



(f) Elevation versus rainfall (mm) for predicted rainfall values in blue observed values in red

Figure 3.6 (a), (c) and (e) show the performance of RF model on the 15th July 2017 for maxT, minT and rainfall with different variations and combinations of the parameters. The tuned parameters are used in the final predictive models. (b), (d) and (f) show the actual and predicted values on 15th July 2017

3.2 Accuracy

This section discusses the performance of the four models namely Kriging with external drift (KED), Gaussian filter (GF), random forest (RF) and multilayer perceptron (MLP) in predicting maximum temperature, minimum temperature and rainfall.

In general, RF model was found to be the most accurate interpolation approach for all three weather variables across all years and seasons as obtained using all three accuracy measures with a few exceptions (Tables 3.1 and 3.2). The generally good performance of RF can be explained by its being an ensemble technique (see section 2.3). There were some exceptions, however, where KED outperformed RF slightly. KED performed better in terms of RMSE and MAE for maximum temperature and in terms of RMSE and R^2 for minimum temperature (Table 3.1). KED also showed higher accuracy in predicting rainfall for July 2016 (Table 3.2). GF is another model that showed some exceptions. GF obtained higher accuracy compared to the three other models for rainfall predictions using RMSE (Table 3.1). The model also showed a higher accuracy of 91% in terms of R^2 for maximum temperature prediction in July (Table 3.2).

Table 3.1 Performance measured in terms of root mean squared error (RMSE), mean absolute error (MAE) and R^2 of the four models used in prediction of the three weather variables for the 15th of January, summer period, in 2016 and 2017. The blue color highlights the best value found for each evaluation metric.

		January					
		2016			2017		
		RMSE	MAE	R^2	RMSE	MAE	R^2
MaxT (°C)	KED	1.43	1.24	0.82	1.09	0.88	0.80
	GF	1.58	1.20	0.87	1.44	1.10	0.82
	RF	1.02	0.85	0.95	1.27	0.89	0.90
	MLP	1.81	1.49	0.71	2.17	1.85	0.02
MinT (°C)	KED	0.92	0.75	0.91	1.04	0.82	0.87
	GF	1.60	1.06	0.76	1.01	0.80	0.90
	RF	1.86	0.69	0.86	0.50	0.42	0.97
	MLP	1.87	1.49	0.39	1.12	0.81	0.79
Rain (mm)	KED	5.06	4.31	0.37	4.94	3.54	0.53
	GF	7.87	6.29	0.40	1.10	4.96	0.52
	RF	4.70	3.14	0.87	4.04	2.66	0.87
	MLP	8.3	6.4	0.25	7.93	5.94	0.07

Table 3.2 Performance measured in terms of root mean squared error (RMSE), mean absolute error (MAE) and R^2 of the four models used in prediction of the three weather variables for the 15th of July, winter period, in 2016 and 2017. The blue color highlights the best value found for each evaluation metric.

		July					
		2016			2017		
		RMSE	MAE	R^2	RMSE	MAE	R^2
MaxT (°C)	KED	1.35	1.03	0.70	1.46	1.18	0.81
	GF	1.34	1.11	0.85	1.51	1.83	0.91
	RF	0.54	0.44	0.97	1.16	0.88	0.90
	MLP	1.59	1.30	0.18	2.17	1.74	0.51
MinT (°C)	KED	1.95	1.64	0.71	1.79	1.42	0.76
	GF	2.65	2.14	0.61	2.00	1.5	0.73
	RF	1.21	0.96	0.91	0.93	0.70	0.93
	MLP	1.95	1.56	0.59	1.83	1.50	0.62
Rain (mm)	KED	0.03	0.01	0.95	0.15	0.084	0.92
	GF	0.06	0.01	0.94	0.12	0.03	0.94
	RF	0.03	0.01	0.92	0.05	0.03	0.90
	MLP	0.03	0.06	0	0.06	0.032	0.11

Overall, MLP exhibited the poorest performance of all of the tested models. MLP may have been sensitive, as discussed in Chronopoulos *et al.* (2008) to the relatively small (92 station) dataset.

For the summer season in both years, the four models showed better performance for maximum and minimum temperature prediction indicating lower RMSE and MAE values which are found to be below 2.17°C than for rainfall prediction (Table 3.1). However, the observations were found to be different for winter where all the four models showed better measurement error for rainfall prediction than for maximum and minimum temperature prediction (Table 3.2).

Another observation to highlight is that there was no considerable difference between predictions of all the weather variables from the driest year to the wettest year in both seasons. All the models gave similar accuracies per weather variable in both years. Thus, in this study, goodness of fit measures were found not to have been influenced by whether or not the year was wetter or drier.

The results found in this research are comparable with the finding of Appelhans *et al.* (2015), who compared several maching learning techniques

such as RF, neural networks and others to Kriging with elevation only, Kriging with elevation and NDVI (Normalized Difference Vegetation Index) and kriging with elevation and the sky view factor for the spatial prediction of air temperature, they found that RF outperformed all other models. Moreover, analyses done by Chen *et al.* (2017) found the best results with RF in comparison to other spatial and machine learning models as well as the results from the study done by Youssef *et al.* (2016).

For KED, to perform slightly better than RF can be explained by the strong emphasis of RF on the distance to the coast and the little emphasis on the influence of elevation. Similar results were found by Appelhans *et al.* (2015), where RF interpolation patterns were brought closer to Kriging interpolations, in locations where RF put too little emphasis on the influence of elevation.

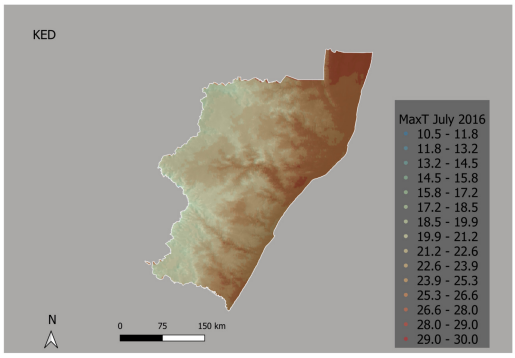
3.3 Interpolated temperature and rainfall surfaces

The summaries in Tables 3.1 and 3.2 give a good overview of model performance. However, there is a great deal of difference between the patterns of variability in the three weather variables across space when looking at the interpolated surface. This is worth exploring further and only one date is chosen in this study to illustrate the difference observed. The generated maps of the three weather variables at a $1km^2$ resolution in the KwaZulu-Natal (KZN) region on the 15th of July 2016 are displayed (Figure 3.7 to 3.9).

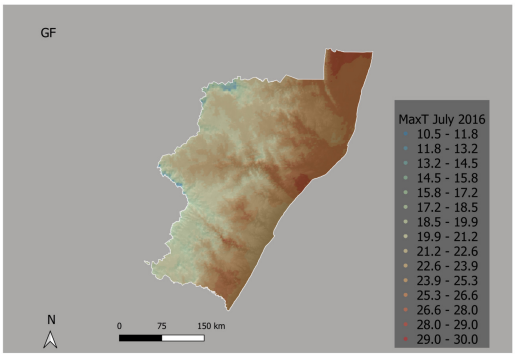
While RF performed well in terms of general goodness of fit, some issues are evident in the patterns of variation that can be seen in maps of interpolated surfaces. Referring to Figures 3.7 and 3.8, the following point is most notable. The evidence of a "banding effect" shown by the purple arrow which was really only obvious in the case of the RF estimates. This could be linked to an artefact of the model since no other model was able to capture such effect. It is not clear why it was so obvious for RF. The effect is stronger for minimum temperature.

It is also observed with all the modelling techniques that minimum and maximum temperatures values are higher along the coast and decrease as you move from the east to the west part of KZN. This can be an effect of the distance to the coast and the elevation used as co-variables which are similar to the finding in Jarvis and Stuart (2001). For their study, they found that elevation, urban index and northing, where northing has high correlation with distance from the south coast are the most influential variables for maximum and minimum temperatures. Thus, our observations validate theirs.

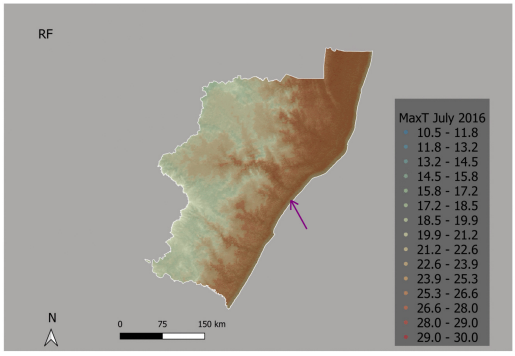
In Figure 3.9, three of the models generated similar maps for rainfall except for the MLP that showed higher rainfall amount along the coast. The predicted rainfall values with MLP range between 12 and 26.5 mm while they average 0 mm for the other models. This aligns with the results shown in Table 3.2 where MLP accuracy measures were found very bad for rainfall.



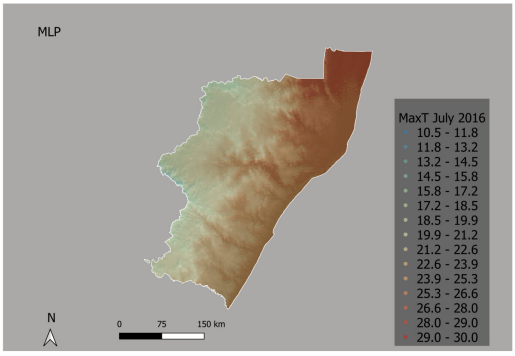
(a) Map with Kriging with external drift



(b) Map with meteoland

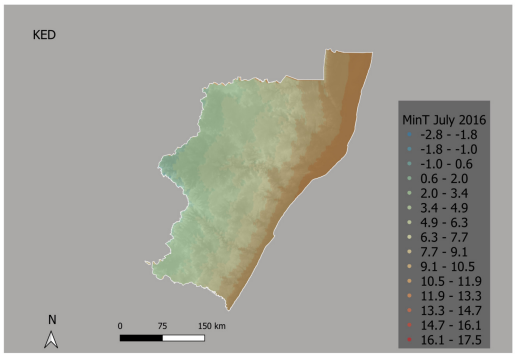


(c) Map with RF

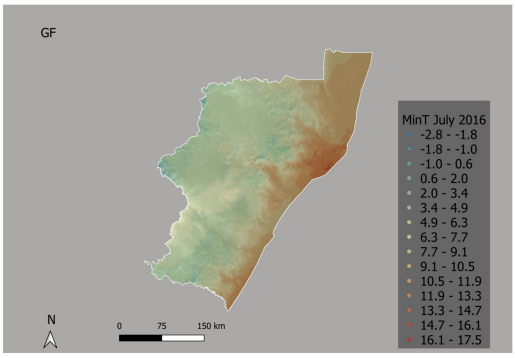


(d) Map with MLP

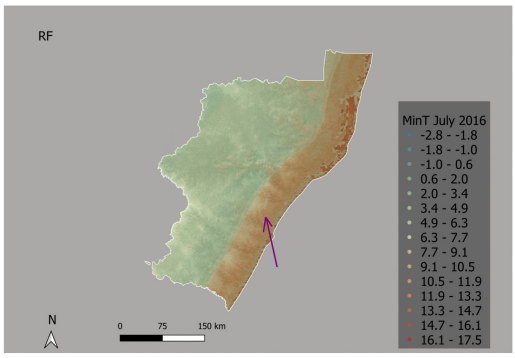
Figure 3.7 Surface interpolations of maximum temperature at a 1km^2 resolution on the 15th of July 2016 for the 4 models



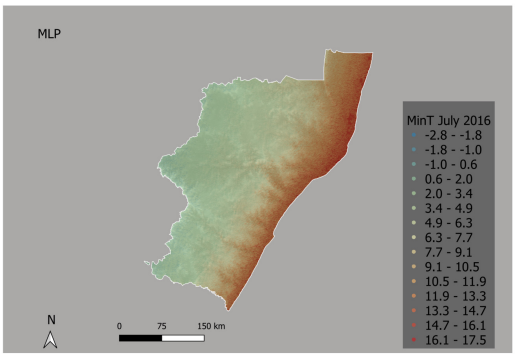
(a) Map with Kriging with external drift



(b) Map with meteoland

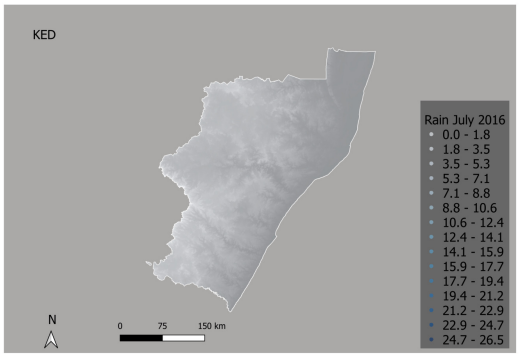


(c) Map with RF

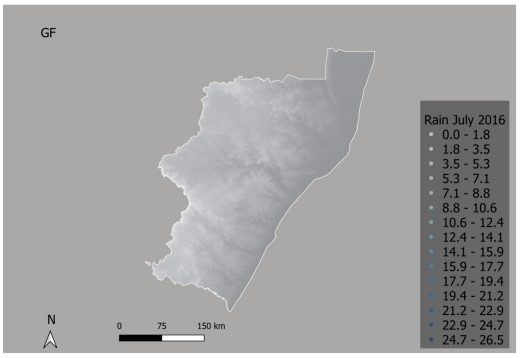


(d) Map with MLP

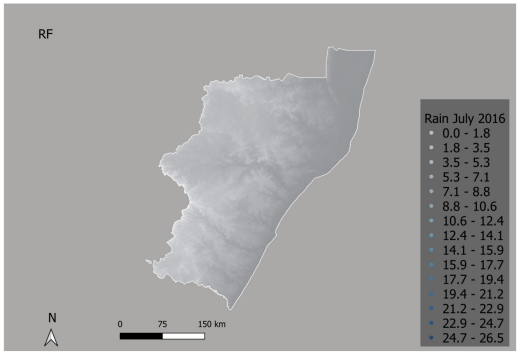
Figure 3.8 Surface interpolations of minimum temperature at a 1km^2 resolution on the 15th of July 2016 for the 4 models



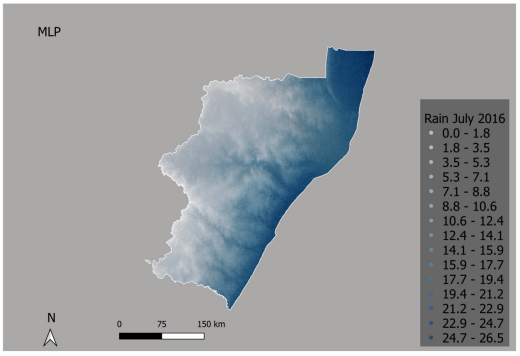
(a) Map with Kriging with external drift



(b) Map with meteoland



(c) Map with RF



(d) Map with MLP

Figure 3.9 Surface interpolations of rainfall at a 1km^2 resolution on the 15th of July 2016 for the 4 models

Chapter 4

Conclusion and future work

In this study, four approaches of interpolation were tested and compared to determine their performances in predicting minimum and maximum temperature, and rainfall weather variables at a $1km^2$ spatial resolution in the region of KwaZulu-Natal (KZN) in South Africa. Kriging with external drift (KED), Gaussian filter (GF), random forest (RF) and multilayer perceptron (MLP) are the models compared and trained on four datasets containing values of the three weather variables collected from 92 weather stations in the region of interest. It is of importance to know if a model show better predictions between the three weather variables in order to specifically guide decision makers towards choosing a robust model with regards to the weather variable of interest. Also, if the performance of the models change between dry and wet years, or between summer and winter seasons.

Given the insufficiency of information in regards to accurate spatial interpolation at a fine resolution in KZN and the increasing demand of gridded data for process based models that heavily rely on climate data, it is of interest to identify a model that can produce reliable gridded data. In this regard, it is important to compare well-known models such as KED with emerging approaches, such as GF, which have recently been used to generate daily interpolated data

The four datasets in this study included values of the weather variables derived on the 15th of January that represents the summer period and the 15th of July that represents the winter period in the years 2016 and 2017 respectively. The models were evaluated based on their capacity to reduce the root mean squared error (RMSE) and the mean absolute error (MAE). Topographical information including elevation, slope, aspect and distance to the coast of the sampled locations and of the unsampled locations at $1km^2$ resolution within KZN were added as inputs to the models. There were extracted from a digital elevation model (DEM) of 30m resolution. RF and MLP included all the co-variables, elevation, slope, aspect and distance to

the coast while with KED only elevation and distance to the coast were considered given the use of the linear variogram which does not suit the values of slope and aspect. GF predictions were done with the use of elevation, slope and aspect since the model does not take any additional information. RF outperformed the other three models in prediction of the three weather variables in both seasons for both years with some exceptions. Another asset of RF is that it has the lowest computing time, followed by KED. KED outperformed for minimum and maximum temperature predictions in January 2016 and 2017 respectively. It outperformed also for rainfall predictions in July 2016. Apart from RF and KED, GF also showed almost similar performance to the KED in predicting minimum and maximum temperature with an approximate difference of 0.5°C in all experiments. Also, even though KED performs slightly better than GF and MLP, it does not have the advantage of predicting a sequence of daily data. KED can only be applied on a dataset containing a particular date or on monthly or yearly averaged data. MLP was the most complex model and required long training in fitting the data yet it was the least accurate. The low accuracy of MLP is related to the sensitivity of the model to the amount of data used for fitting it. Thus, an increase of the data while fitting the MLP model could contribute in a decrease of errors and therefore in a more reliable prediction. The present study suggests that there is utility in complementing approaches to spatial interpolation while deriving gridded data. Although, the results demonstrate the outperformance of RF compared to other models, they also point out that in certain variables, KED has the lead. Thus, having the four models allow for complementarity in which where one model fails, another provides reduced error range.

Interpolated temperature and rainfall surfaces were produced using the four models. Even though RF provided lower errors in terms of RMSE and MAE and has given higher R^2, there was a presence of a banding effect in the interpolated maximum and minimum temperature generated by RF. The effect was observed along the coast as you get further from the ocean giving surfaces that do not ressemble true temperatures.

The findings of this research allow us to conclude that RF can be used to predict minimum and maximum temperature, and rainfall at a $1km^2$ resolution in KZN with relatively low error measures. In this research, it was also shown that accurate interpolation temperature and rainfall surfaces can be created at a fine resolution in KZN.

This study is limited by scope, and the results reported should be considered in that regard. The study only focused on specific dates within two seasons in the KZN region. There is obvious room for expanding the study beyond these boundaries to include a wider range of periods to enhance its

generalizability. It is hoped that the work done in the present study will provide a pointer for future direction in research in this light.

List of References

Abutaleb, A.S. (1991). A neural network for the estimation of forces acting on radar targets. *Neural networks*, vol. 4, no. 5, pp. 667–678.

Anctil *et al.* (2004). Impact of the length of observed records on the performance of ann and of conceptual parsimonious rainfall-runoff forecasting models. *Environmental Modelling & Software*, vol. 19, no. 4, pp. 357–368.

Appelhans *et al.* (2015). Evaluating machine learning approaches for the interpolation of monthly air temperature at mt. kilimanjaro, tanzania. *Spatial Statistics*, vol. 14, pp. 91–113.

Beesley *et al.* (2009). A comparison of the bawap and silo spatially interpolated daily rainfall datasets. In: *18th world IMACS/MODSIM congress, Cairns, Australia*, pp. 13–17. Citeseer.

Beltratti *et al.* (1996). *Neural networks for economic and financial modelling*. International Thomson Computer Press London, UK.

Berndt *et al.* (2018). Spatial interpolation of climate variables in northern germany—influence of temporal resolution and network density. *Journal of Hydrology: Regional Studies*, vol. 15, no. 3, pp. 184–202.

Beucher, H. and Renard, D. (2016). Truncated gaussian and derived methods. *Comptes Rendus Geoscience*, vol. 348, no. 7, pp. 510–519.

Bourennane *et al.* (1996). Improving the kriging of a soil variable using slope gradient as external drift. *European Journal of Soil Science*, vol. 47, no. 4, pp. 473–483.

Bourennane *et al.* (2000). Comparison of kriging with external drift and simple linear regression for predicting soil horizon thickness with different sample densities. *Geoderma*, vol. 97, no. 3-4, pp. 255–271.

Breiman, L. (2001). Random forests. *Machine learning*, vol. 45, no. 1, pp. 5–32.

Caruso, C. and Quarta, F. (1998). Interpolation methods comparison. *Computers & Mathematics with Applications*, vol. 35, no. 12, pp. 109–126.

Chen *et al.* (2017). A comparative study of logistic model tree, random forest, and classification and regression tree models for spatial prediction of landslide susceptibility. *Catena*, vol. 151, pp. 147–160.

Chorti, A. and Hristopulos, D.T. (2008). Nonparametric identification of anisotropic (elliptic) correlations in spatially distributed data sets. *IEEE Transactions on signal processing*, vol. 56, no. 10, pp. 4738–4751.

Chronopoulos *et al.* (2008). An application of artificial neural network models to estimate air temperature data in areas with sparse network of meteorological stations. *Journal of Environmental Science and Health Part A*, vol. 43, no. 14, pp. 1752–1757.

Coulibaly, M. and Becker, S. (2007). Spatial interpolation of annual precipitation

in south africa-comparison and evaluation of methods. *Water International*, vol. 32, no. 3, pp. 494–502.

De Cáceres, M. (2019). Package meteoland (ver. 0.8. 1).

DeCaceres *et al.* (2018). Estimating daily meteorological data and downscaling climate models over landscapes. *Environmental modelling & software*, vol. 108, no. 6, pp. 186–196.

Fick, S.E. and Hijmans, R.J. (2017). Worldclim 2: new 1-km spatial resolution climate surfaces for global land areas. *International journal of climatology*, vol. 37, no. 12, pp. 4302–4315.

Flores, F., Lillo, M. *et al.* (2010). Simple air temperature estimation method from modis satellite images on a regional scale. *Chilean Journal of Agricultural Research*, vol. 70, no. 3, pp. 436–445.

Foster *et al.* (1996). Snow cover and snow mass intercomparisons of general circulation models and remotely sensed datasets. *Journal of Climate*, vol. 9, no. 2, pp. 409–426.

FrontlineSolvers (2019). Neural network prediction. `https://www.solver.com/neural-network-prediction`.

Germishuizen, I. (2018). Mapping risk at different spatial and temporal scales for short- and long-term risk evaluation: The case of the eucalypt gall wasp leptocybe invasa. `http://conferences.sun.ac.za/index.php/ff2018/NFFF2018/paper/view/3761`. Accessed: 2019-01-30.

Gleason *et al.* (2008). Obtaining weather data for input to crop disease-warning systems: leaf wetness duration as a case study. *Scientia Agricola*, vol. 65, no. SPE, pp. 76–87.

Gorsich, D.J. and Genton, M.G. (2000). Variogram model selection via nonparametric derivative estimation. *Mathematical geology*, vol. 32, no. 3, pp. 249–270.

Hartkamp *et al.* (1999). Interpolation techniques for climate variables.

Hengl *et al.* (2018). Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. *PeerJ*, vol. 6, p. e5518.

Hijmans *et al.* (2005). Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology: A Journal of the Royal Meteorological Society*, vol. 25, no. 15, pp. 1965–1978.

Hughes, D. and Smakhtin, V. (1996). Daily flow time series patching or extension: a spatial interpolation approach based on flow duration curves. *Hydrological Sciences Journal*, vol. 41, no. 6, pp. 851–871.

Hutchinson, M.F. (1995). Interpolating mean rainfall using thin plate smoothing splines. *International journal of geographical information systems*, vol. 9, no. 4, pp. 385–403.

Jarvis, C.H. and Stuart, N. (2001). A comparison among strategies for interpolating maximum and minimum daily air temperatures. part i: The selection of "guiding" topographic and land cover variables. *Journal of Applied Meteorology*, vol. 40, no. 6, pp. 1060–1074.

Jeffery, S. (2006). Error analysis for the interpolation of monthly rainfall used in the generation of silo rainfall datasets. In: *Technical report, the Queensland Department of Natural Resources*.

Jeffrey *et al.* (2001). Using spatial interpolation to construct a comprehensive

archive of australian climate data. *Environmental Modelling & Software*, vol. 16, no. 4, pp. 309–330.

Joshi, R. (2016). Artificial neural network (ann) based empirical interpolation of precipitation. *International Journal of Mathematical, Engineering and Management Sciences*, vol. 1, no. 3, pp. 93–106.

Karavani *et al.* (2018). Effect of climatic and soil moisture conditions on mushroom productivity and related ecosystem services in mediterranean pine stands facing climate change. *Agricultural and Forest Meteorology*, vol. 248, pp. 432–440.

Kerry, R. and Oliver, M.A. (2008). Determining nugget: sill ratios of standardized variograms from aerial photographs to krige sparse soil data. *Precision Agriculture*, vol. 9, no. 1-2, pp. 33–56.

Kilibarda *et al.* (2014). Spatio-temporal interpolation of daily temperatures for global land areas at 1 km resolution. *Journal of Geophysical Research: Atmospheres*, vol. 119, no. 5, pp. 2294–2313.

Kleijnen, J.P. (2009). Kriging metamodeling in simulation: A review. *European journal of operational research*, vol. 192, no. 3, pp. 707–716.

Kohavi, R. *et al.* (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In: *Ijcai*, vol. 14, pp. 1137–1145. Montreal, Canada.

Kumar, S. *et al.* (2005). Rainfall-runoff modelling using artificial neural networks: comparison of network types. *Hydrological Processes: An International Journal*, vol. 19, no. 6, pp. 1277–1291.

Li *et al.* (2011). A review of comparative studies of spatial interpolation methods in environmental sciences: Performance and impact factors. *Ecological Informatics*, vol. 6, no. 3-4, pp. 228–241.

Li, J. and Heap, A.D. (2008). A review of spatial interpolation methods for environmental scientists.

Li, J. and Heap, A.D. (2014). Spatial interpolation methods applied in the environmental sciences: A review. *Environmental Modelling & Software*, vol. 53, pp. 173–189.

Ly *et al.* (2011). Geostatistical interpolation of daily rainfall at catchment scale: the use of several variogram models in the ourthe and ambleve catchments, belgium. *Hydrology and Earth System Sciences*, vol. 15, no. 7, pp. 2259–2274.

Makhuvha *et al.* (1997). Patching rainfall data using regression methods.: 1. best subset selection, em and pseudo-em methods: theory. *Journal of Hydrology*, vol. 198, no. 1-4, pp. 289–307.

McKenney *et al.* (2011). Customized spatial climate models for north america. *Bulletin of the American Meteorological Society*, vol. 92, no. 12, pp. 1611–1622.

Meckesheimer *et al.* (2002). Computationally inexpensive metamodel assessment strategies. *AIAA journal*, vol. 40, no. 10, pp. 2053–2060.

Mitas, L. and Mitasova, H. (1999). Spatial interpolation. *Geographical information systems: principles, techniques, management and applications*, vol. 1, no. 2.

More, A. and Deo, M. (2003). Forecasting wind with neural networks. *Marine structures*, vol. 16, no. 1, pp. 35–49.

Mowrer, H.T. (1997). Propagating uncertainty through spatial estimation processes

for old-growth subalpine forests using sequential gaussian simulation in gis. *Ecological Modelling*, vol. 98, no. 1, pp. 73–86.

Neteler, M. (2010). Estimating daily land surface temperatures in mountainous environments by reconstructed modis lst data. *Remote sensing*, vol. 2, no. 1, pp. 333–351.

New *et al.* (2002). A high-resolution data set of surface climate over global land areas. *Climate research*, vol. 21, no. 1, pp. 1–25.

Niekerk, V. *et al.* (2011). Input variable selection for interpolating high-resolution climate surfaces for the western cape. *Water SA*, vol. 37, no. 3.

Olea, R.A. (2006). A six-step practical approach to semivariogram modeling. *Stochastic Environmental Research and Risk Assessment*, vol. 20, no. 5, pp. 307–318.

Oliver, M.A. and Webster, R. (1990). Kriging: a method of interpolation for geographical information systems. *International Journal of Geographical Information System*, vol. 4, no. 3, pp. 313–332.

Pardo-Igúzquiza, E. (1999). Varfit: a fortran-77 program for fitting variogram models by weighted least squares. *Computers & Geosciences*, vol. 25, no. 3, pp. 251–261.

Pascanu *et al.* (2013). On the difficulty of training recurrent neural networks. In: *International conference on machine learning*, pp. 1310–1318.

Pebesma *et al.* (2016). Spatio-temporal interpolation using gstat. *RFID Journal*, vol. 8, no. 1, pp. 204–218.

Queensland, G. (1889). Australian climate data from 1889 - silo. `https://www.longpaddock.qld.gov.au/silo/`.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
Available at: `http://www.R-project.org/`

Raad *et al.* (2012). Breast cancer classification using neural network approach: Mlp and rbf. *networks*, vol. 7, no. 8, p. 9.

Rigol *et al.* (2001). Artificial neural networks as a tool for spatial interpolation. *International Journal of Geographical Information Science*, vol. 15, no. 4, pp. 323–343.

Sacks *et al.* (1989). Design and analysis of computer experiments. *Statistical science*, pp. 409–423.

Sajikumar *et al.* (1999). A non-linear rainfall–runoff model using an artificial neural network. *Journal of hydrology*, vol. 216, no. 1-2, pp. 32–55.

Sakata *et al.* (2003). Structural optimization using kriging approximation. *Computer methods in applied mechanics and engineering*, vol. 192, no. 7-8, pp. 923–939.

Salcedo-Sanz *et al.* (2016). Monthly prediction of air temperature in australia and new zealand with machine learning algorithms. *Theoretical and applied climatology*, vol. 125, no. 1-2, pp. 13–25.

Sánchez-Pinillos *et al.* (2018). Relative size to resprouters determines post-fire recruitment of non-serotinous pines. *Forest ecology and management*, vol. 429, pp. 300–307.

sasri/institute (). South african sugarcane research institute. `http://safarikzn.com/kwazulu-natal-climate-weather/`.

SawsService (). The south african weather service. `http://www.weathersa.co.za/`.

Schulze *et al.* (2007). Annual precipitation. *South African Atlas of Climatology and Agrohydrology: Water Research Commission, Pretoria, RSA, WRC Report 1489/1/06, Section 6.2*.

Segal, M.R. (2004). Machine learning benchmarks and random forest regression.

Seo *et al.* (2015). Estimating spatial precipitation using regression kriging and artificial neural network residual kriging (rknnrk) hybrid approach. *Water Resources Management*, vol. 29, no. 7, pp. 2189–2204.

Simpson *et al.* (2001). Kriging models for global approximation in simulation-based multidisciplinary design optimization. *AIAA journal*, vol. 39, no. 12, pp. 2233–2241.

Sola *et al.* (1997). Importance of input data normalization for the application of neural networks to complex industrial problems. *IEEE Transactions on nuclear science*, vol. 44, no. 3, pp. 1464–1468.

Stein, M.L. (2012). *Interpolation of spatial data: some theory for kriging*. Springer Science & Business Media.

Svetnik *et al.* (2003). Random forest: a classification and regression tool for compound classification and qsar modeling. *Journal of chemical information and computer sciences*, vol. 43, no. 6, pp. 1947–1958.

Thornton *et al.* (1997). Generating surfaces of daily meteorological variables over large regions of complex terrain. *Journal of Hydrology*, vol. 190, no. 3-4, pp. 214–251.

Tobler, W.R. (1970). A computer movie simulating urban growth in the detroit region. *Economic geography*, vol. 46, no. sup1, pp. 234–240.

Uyan, M. and Cay, T. (2013). Spatial analyses of groundwater level differences using geostatistical modeling. *Environmental and ecological statistics*, vol. 20, no. 4, pp. 633–646.

Van Beers, W.C. and Kleijnen, J.P. (2004). Kriging interpolation in simulation: a survey. In: *Proceedings of the 2004 Winter Simulation Conference, 2004.*, vol. 1. IEEE.

Vermeulen *et al.* (2015). Financial stress indices and financial crises. *Open Economies Review*, vol. 26, no. 3, pp. 383–406.

VitalWeather (). The vitalweather system. `http://www.vitalweather.co.za/about`.

Webster, R. and Oliver, M.A. (2007). *Geostatistics for environmental scientists*. John Wiley & Sons.

Wojtasik, E.M. (2014). *Richness and diversity of alien ethnomedicinal plant taxa used and sold for traditional medicine in South Africa*. Ph.D. thesis.

Yeh *et al.* (2013). Spatial interpolation using mlp–rbfn hybrid networks. *International Journal of Geographical Information Science*, vol. 27, no. 10, pp. 1884–1901.

Youssef *et al.* (2016). Landslide susceptibility mapping using random forest, boosted regression tree, classification and regression tree, and general linear models and comparison of their performance at wadi tayyah basin, asir region, saudi arabia. *Landslides*, vol. 13, no. 5, pp. 839–856.

Yuan *et al.* (2015). Validation of china-wide interpolated daily climate variables from 1960 to 2011. *Theoretical and Applied Climatology*, vol. 119, no. 3-4, pp.

689–700.

Zhou, Z.-H. (2012). *Ensemble methods: foundations and algorithms*. Chapman and Hall/CRC.