Comparison between Statistical Method and Machine Learning Method for Downscaling

Candidate Number: YQNN3

Course Code: STAT0035

Word Count: 14238

April 26, 2023

I, Candidate Number: YQNN3, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the work.

Abstract

This work compares the statistical and machine learning method for downscaling, for use in studying climate impact and adaptation to the increased flood risk. The statistical method used is generalised linear model, and the machine learning methods used include Bayesian network and random forest. A case study is conducted at 87 stations over Rhineland-Palatinate in Germany to simulate multivariate and multisite daily precipitation and temperature series using both methods. The following aspects of both methods are compared: the simulation results; computational efficiency; requirement for a complete dataset; model-building process; and the relationship between weather variables.

Keywords: climate impact, flood, downscaling, comparison

Contents

1 Introduction **6**

2 Study Area and Data **9**

2.1 Precipitation and Temperature Data 9

2.2 Topographic Data 11

2.3 Geographical Data for the Stations 11

2.4 Atmospheric Predictor Data 12

2.5 Summary 14

3 Methods for Downscaling **16**

3.1 Statistical Method 16

3.1.1 Theory of Generalised Linear Model 16

3.1.2 Univariate Weather Generation 18

3.1.3 Bivariate Weather Generation 20

3.1.4 Simulation 21

3.2 Machine Learning Method 22

3.2.1 Bayesian Network 22

3.2.2 Random Forest 27

3.3 Summary 30

4 Model Building **32**

4.1 Statistical Model 32

4.1.1 First Stage: Univariate Models 33

4.1.2 Second Stage: Bivariate Models 37

4.1.3 Third Stage: Model including Atmospheric Covariates . . . 38
4.2 Machine Learning Model . . . 39
4.2.1 Bayesian Network . . . 40
4.2.2 Random Forest . . . 41
4.3 Summary . . . 42

5 Simulation Results **43**
5.1 Simulations from Statistical Model . . . 43
5.1.1 Monthly Summary Statistics . . . 44
5.1.2 Seasonal Summary Statistics . . . 49
5.2 Simulations from Machine Learning Model . . . 54
5.2.1 Monthly Summary Statistics . . . 55
5.2.2 Seasonal Summary Statistics . . . 60
5.3 Comparison between Two Methods . . . 64
5.4 Summary . . . 66

6 Conclusion **67**

Appendices **71**

A Model Structure **71**
A.1 Precipitation Occurrence Model . . . 72
A.2 Precipitation Amount Model . . . 74
A.3 Temperature Model . . . 76

B Figures **79**

Bibliography **92**

Chapter 1

Introduction

Climate change has become one of the most significant problems facing humanity. One of its potential impacts is the increase in extreme events such as flooding. For example, in 2021, Germany was severely affected by floods, which caused deaths and significant socioeconomic impacts (Fekete and Sandholz, 2021). Among all the areas in Germany, one of the worst damages occurred in Rhineland-Palatinate, where more than 100 people died (Zimmermann, 2022). In order to get more accurate forecasts and effectively adapt to the increased flood risk, it is important to understand the potential impacts of various climate factors on flooding.

In the past, many studies have concentrated on investigating the effect of climate change using global climate models (GCMs). However, GCM outputs have a very coarse spatial resolution, which could be unreliable for representing the local-scale climate factors on scales below 200 km (Meehl et al., 2007). For example, urban flooding may be sensitive to the timing and amount of precipitation on a few hundred-metre spatial scales (Chandler, 2020). Therefore, it is helpful to downscale GCM outputs to a finer spatial resolution, and thus obtain better local-scale climate information for the floodings (e.g. Rucker et al. 2021; Tofiq and Güven 2015; Nam et al. 2014; Fakhri et al. 2012; Do Hoai et al. 2011)

Downscaling is based on the assumption that large-scale weather will have a strong influence on local-scale weather, but the reverse effects from local-scale weather to large-scale weather can be ignored. There are two approaches to downscaling. One approach is dynamical downscaling, which obtains higher resolution

outputs by nesting a regional climate model (formulated based on physical principles) into the GCM. Another approach is statistical downscaling, which establishes statistical relationships between large(r)-scale weather (as predictors) and observed local-scale weather (as predictands) (Maraun et al., 2010). Key benefits of statistical downscaling include its lower computational requirements and ease of implementation (Jang and Kavvas, 2015). Therefore, this study will focus on statistical downscaling.

According to Maraun et al. (2010), the statistical downscaling methods can be classified into three categories: perfect prognosis, model output statistics and weather generators. Perfect prognosis is based on the assumption that the simulated large-scale predictors can be perfectly modelled by the GCM. Model output statistics are designed to handle situations where the GCM does not reproduce the large-scale predictors perfectly, because it does not rely on observed values of the large-scale structure. These two approaches are often used simply to derive overall distributions of a particular variable for some future period. In contrast, weather generators can capture the detailed structure of weather events in space and time, which is likely to be important when assessing the risk of flooding and other climate impacts.

Many statistical downscaling models are available for these statistical downscaling methods, and the choice of models significantly impacts the accuracy of the results. The COST action VALUE (http://www.value-cost.eu) provided a framework to assess the performance of different downscaling methods. However, it did not include any machine learning techniques. With the development of machine learning techniques in environmental science, the recent literature compared different statistical downscaling methodologies created using machine learning techniques and traditional statistical techniques. For instance, Legasa and Gutiérrez (2020) simulated multisite daily precipitation occurrence in southeast Germany based on Bayesian network (a machine learning method) using weather generators and assessed the performance of results using the Wilks model (a statistical method) as a benchmark. Also, Legasa et al. (2022) compared the simulations of

daily precipitation amounts on wet days over Europe generated by a posteriori random forest (a machine learning method) and generalised linear model (GLM) under perfect prognosis conditions.

However, these two research assessed the simulations of precipitation occurrence and amount separately, but did not generate and assess the whole precipitation sequences that combine the simulations of precipitation occurrence and amount. Also, other weather variables like temperature were not considered.

Therefore, this study aims to compare the statistical and machine learning methods to assess their performance for downscaling. To achieve this aim, the multivariate and multisite daily precipitation and temperature series simulations are generated using both methods, and then the performance of both methods is evaluated. Based on the above two research, the model selection of the statistical method is GLM, and the machine learning method combines the results of both Bayesian network and random forest. To illustrate the idea, a case study is conducted over Rhineland-Palatinate in Germany to generate daily precipitation and temperature.

The remainder of this report is structured as follows: Chapter 2 describes the study area and data used. Chapter 3 explains the theoretical background of the statistical and machine learning methods used in this study for downscaling. Chapter 4 illustrates how the models are built and trained. Chapter 5 presents how to simulate the results and compares the performance of the two methods. Chapter 6 summarises the research and discusses the limitations and future work.

Chapter 2

Study Area and Data

As an illustrative case study, Rhineland-Palatinate is chosen as the study area because in 2021 it was one of the worst-hit areas by flooding in Germany. This area lies between 50.0N to 52.0N, 5.6E to 8.9E. Figure 2.1 is the map of this area within Northern Europe.

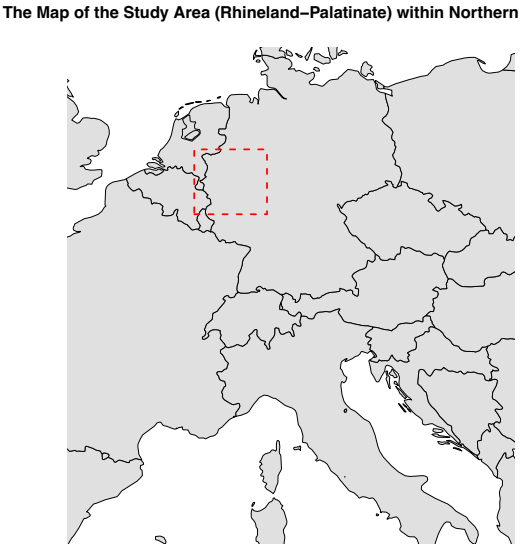

Figure 2.1: Map of the study area within Northern Europe. The red rectangle indicates the study area, which lies between 50.0N to 52.0N, 5.6E to 8.9E.

2.1 Precipitation and Temperature Data

The main cause of 2021 European flooding is heavy rain (Lehmkuhl et al., 2022). Therefore, precipitation is one important factor in this study. Also, the temperature

is relevant because it controls the rates of water evaporation and transpiration.

The daily precipitation and temperature data from 1959 to 2022 are used, and they are downloaded from the European Climate Assessment & Dataset (ECA&D) at http://www.ecad.eu (Klein Tank et al., 2002). The dataset contains both blended data that have filled in the missing values and non-blended data that contains the missing data. The non-blended data are used in this study because the filling in of the missing data may cause the problems, such as the possibility of the filled values having different statistical features from the rest of the data (Chandler, 2020). Of the original 453 stations in the study area, 87 stations are chosen because these stations are the only ones providing information on both precipitation and temperature from 1959 to 2022. There are 1,113,429 observations in total.

Before analysis, the data were checked to examine the record lengths, completeness, and other errors. First, check the annual proportions of wet days at each site for wet day thresholds of 0, 0.5, 1, 2 and 5 *mm* respectively. It can be found that some stations have unrealistically high proportions of wet days, and thus they are removed from the training data. Also, check any anomalies by looking at the plots of annual proportions of wet days at each site for different thresholds (Appendix, Figure B.1). Some sites that show anomalies were unusually wet or dry in certain periods and need further investigation. Next, draw the plot of the frequency of observations per year for these unusual sites (Appendix, Figure B.2) to speculate on the reasons for the unusual precipitation. For example, site S041 has several missing observations in 2003 and 2004, suggesting that there may be some operational difficulties during this period. Thus we removed data for these two years from the record of S041.

Another common problem with precipitation data is that the recording resolution can differ between stations and change over time due to differences in measurement equipment. For example, some sites may record to the nearest 0.5 *mm*, whereas others may record to the nearest 0.2 *mm*. This means, for example, that the definition of a ”wet” day (i.e. a day with non-zero recorded rainfall) may differ between sites, which complicates the subsequent analysis. Therefore, it is helpful

to check for major differences in recording resolution - which can be done by comparing the frequency distributions of the first decimal digit in the precipitation data both between stations and over time.

The graph of first decimals' frequencies at each site (Appendix, Figure B.3) shows that most of the recording resolution is fairly homogeneous except for sites S037 and S054. They have wildly different recording resolutions compared to other sites, so they are removed from the training data because they may not be comparable with the other sites. Another thing to check is whether the recording resolution has changed over time. From the graph (Appendix, Figure B.4), there are some clear but unsystematic differences between years, in particular with an increasing frequency of records whose first decimal digit was 1 in the last decade. To remove the effects of such inhomogeneities, all precipitation values were rounded to the nearest 0.5 *mm*.

2.2 Topographic Data

Topographic map data include altitude data at roughly 1 km^2 resolution. The source of the data are at http://webmap.ornl.gov/wcsdown/dataset.jsp?ds_id=10003 from the GTOPO30 Digital Elevation Model. Figure 2.2 is a topographical map of the area produced using these data, with the locations of the original 87 stations on it. It also has country borders: the area includes small parts of the Netherlands and Belgium. However, there was no data from Belgium; and none of the Dutch stations provided both precipitation and temperature. Therefore, the areas of Belgium and the Dutch on the map are empty.

2.3 Geographical Data for the Stations

In addition to the latitude and longitude of each station, various other geographical variables are also available. These include the stations' codes, names, altitude and mapped altitude. The mapped altitude is the altitude of the station according to the topographic dataset, corresponding to the altitude of the 1 km grid square containing the station. It can be useful to check the correctness of altitude because they are supposed to be fairly similar. Also, the mean altitude, standard deviation of

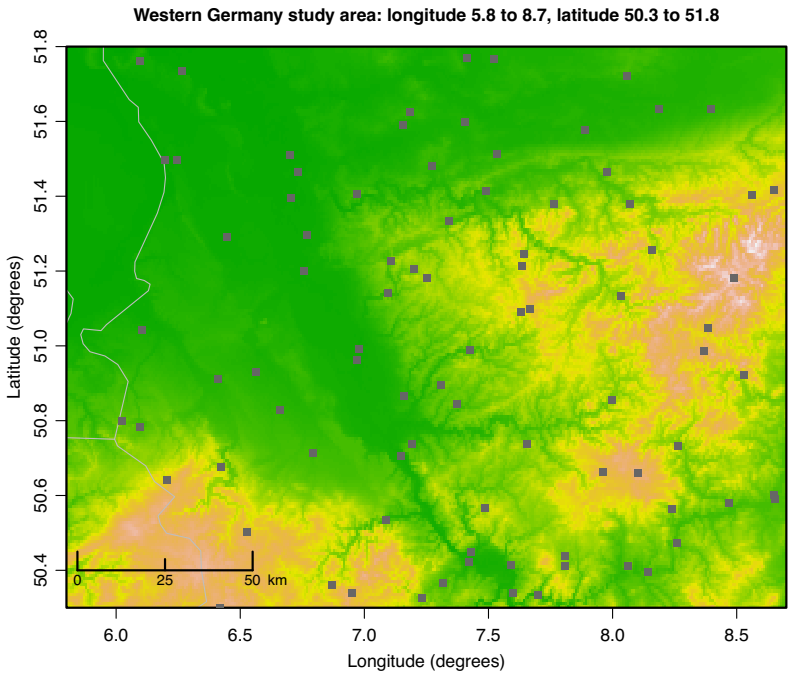

Figure 2.2: Topographical map of the study area, with the 87 station locations indicated using grey squares. Grey curves are the boundaries between Germany, Belgium and the Netherlands. The altitudes on the map range from 6 m to 817 m.

altitude, east-west and north-south average slopes are calculated over domains with a centre distance of 3×3, 10×10, and 30×30 km^2 for each site (Appendix Figure B.5 shows the graph of these variables). The topographic variability is represented by altitude standard deviations for the same domains. These topographic variations could be relevant predictors of precipitation and temperature. For instance, temperatures on south-facing slopes will usually be higher than those on north-facing slopes.

2.4 Atmospheric Predictor Data

As statistical downscaling establishes a statistical relationship between large-scale and local-scale variables, data on potential large-scale atmospheric predictors are needed to develop downscaling relationships. The atmospheric predictor data used in this study are daily time series of u- and v- wind components, temperature, geopotential (i.e. the potential energy of a unit mass relative to sea level) and specific humidity, each at 1000, 850, 700, and 500 hPa levels (these four levels corresponding

to four different heights in the atmosphere, pressure is usually around 1000 hPa at sea level and it decreases when height increases), averaged over the region 3.25E to 11.25E, 47N to 55N (Figure 2.3 shows this region) from 1st January 1959 to 31st December 2021. The selection of atmospheric predictors is guided by that in Quesada-Chacón et al. (2022) who used the same predictors averaged over an 8×8 degree domain, for downscaling over a study area in Eastern Germany that is of similar size to the one considered here.

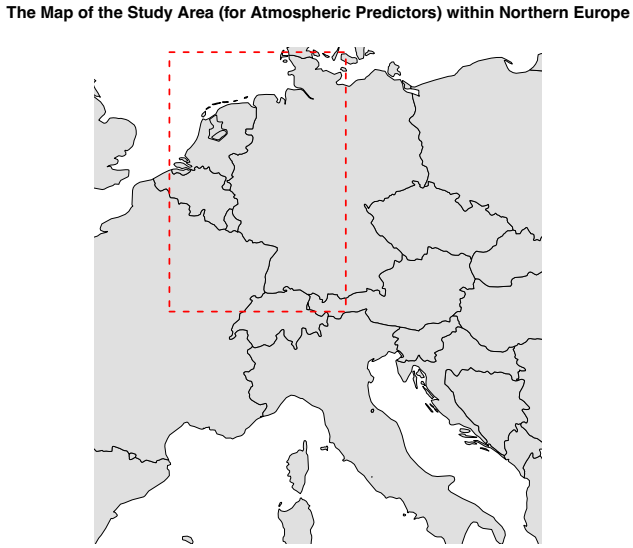

Figure 2.3: Map of the study area for atmospheric predictors within Northern Europe. The red rectangle indicates the study area, which lies between 3.25E to 11.25E, 47N to 55N.

The original downloads were hourly data on a 0.5×0.5 degree grid over the region 3.25E to 11.25E, 47N to 55N. They were downloaded from ERA5 Reanalysis (i.e. combine the past observations with modern data and model to generate complete datasets of the past weather) dataset (https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysi era5-pressure-levels). The raw ERA5 data are available on a 0.25×0.25 degree grid. However, since we only need very coarse scale average values, 0.5×0.5 degree grid data were used to reduce the file size for downloads. To convert the hourly data to daily values, they were averaged over the period 00:00 to 23:00 Coordinated Universal Time (UTC) each day. The reason for choosing this time

interval is that the precipitation is usually for a 24-hour period ending around either 8 am or 9 am local time (and possibly adjusting for daylight saving in summer). In contrast, daily mean temperatures are calculated as the average of the daily maximum and minimum over a similar period. However, the precise periods used can differ between stations: this makes it difficult to align the reanalysis data perfectly with the observations, and the period 00:00 to 23:00 UTC has therefore been used as a simple approximation that overlaps to a large degree with the 24-hour periods represented by the daily station data.

The raw data were processed to produce the following units shown in Table 2.1. This ensures that the magnitudes of most variables are mostly between -20 and +20 so that there will not be any danger of computational instability due to floating-point arithmetic overflows or underflows. The original units of geopotential from ERA5 data are $m^2 s^{-2}$, but the values are very large and hence liable to lead to numerical instabilities in statistical modelling algorithms. Therefore, the value is divided by 10^4, which converts to "thousands of geopotential units" and hence corresponds approximately to the height (in *km*) at which the relevant pressure level is found (Mcilveen, 1992).

Geopotential	Temperature	Specific humidity	U-wind	V-wind
GPU*1000	deg C	g/kg	m/s	m/s

Table 2.1: Table of atmospheric predictors and their corresponding units.

2.5 Summary

In conclusion, this chapter describes the study area and data needed in the study. The study area is Rhineland-Palatinate. This area is located between 50.1N and 52.0N, and 5.6E and 8.9E. The following data are needed to investigate the climate change on the flooding in this region: Precipitation and temperature data from 1959 to 2022 were used because they are two important factors regarding flooding. The quality of the precipitation data needs to be checked to avoid biased results. Some unusual precipitation data were removed, and the precipitation values were rounded to the nearest 0.5 *mm* to reduce the effect of apparent differences in recording res-

olution among different sites. Topographic data are also provided and they can be visualised by the topographical map. Among all the stations in this area, 87 stations were selected because they are the only ones that provide data on both temperature and precipitation. The related attributes of stations are used. For example, code, name, longitude, latitude, altitude and mapped altitude of each station, as well as mean altitude, altitude standard deviation, east-west slope and north-south slope which are determined over domains with a centre distance of 3×3, 10×10 and 30×30 km^2 for each station. The atmospheric predictor data include daily time series of u- and v- wind components, temperature, geopotential, and specific humidity, each at 1000, 850, 700, and 500 hPa levels, averaged over the region 3.25E to 11.25E, 47N to 55N from 1st January 1959 to 31st December 2021. Average original hourly data over hours 00:00 to 23:00 UTC to get daily data and modify some measurement units to avoid computational instability. These data will be useful for subsequent model building.

Chapter 3

Methods for Downscaling

3.1 Statistical Method

This section describes the theoretical framework of GLM for building weather generators. The software package used here is Rglimclim (Chandler, 2018), which is designed to build the multisite, multivariate daily weather series based on GLM, particularly for climatological downscaling applications.

3.1.1 Theory of Generalised Linear Model

Linear regression model is one of the most popular techniques for statistical downscaling (Maraun et al., 2010). It can be written as the following equation:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i \tag{3.1}$$

where Y_i: i = 1,...,n represents the predictand for the ith case in the dataset, and x_{ij}: j = 1,...,p is the corresponding predictor. $\boldsymbol{\beta} = (\beta_0,\ \beta_1,\ \ldots,\beta_p)^T$ is a vector of unknown parameters. Errors ε_i are independent and identically follow the distribution $\varepsilon_i \sim N(0,\sigma^2)$.

However, linear regression model assumes the predictands follow the normal distributions with constant variance. From Figure 3.1, it can be seen that the variances of both precipitation and temperature for the dataset considered in this study are not constant, which means linear regression model is not suitable for these data. To tackle this problem, GLM is used instead. GLM is an extension of linear re-

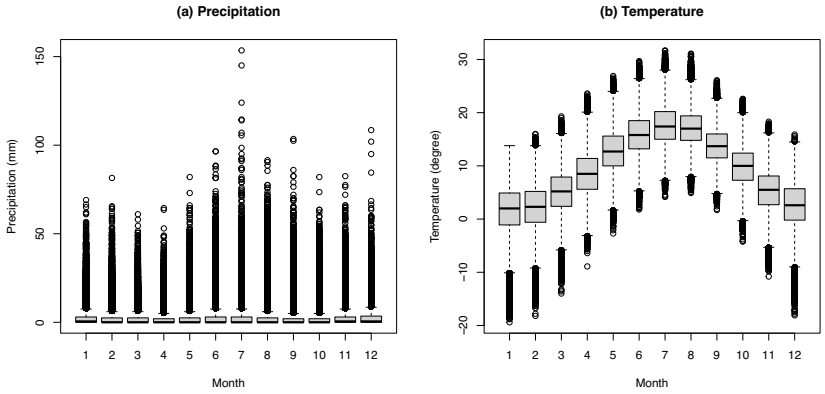

Figure 3.1: Monthly distribution for precipitation and temperature data in this study: Plot (a) indicates the precipitation distribution for each month; Plot (b) indicates the temperature distribution for each month. It can be seen that the variances of both of them vary with the month.

gression model. It relaxes two assumptions of linear regression model. First, the response variables are no longer restricted to the normal distribution and follow a wider range of distributions in exponential family. Exponential family is defined as the family of all distributions with densities of the form:

$$f(y;\psi,\phi) = \exp\left[\frac{y\psi - b(\psi)}{a(\phi)} + c(y,\phi)\right] \tag{3.2}$$

for some functions $a(\cdot)$, $b(\cdot)$, $c(\cdot)$, where ϕ is known as a dispersion parameter, ϕ and ψ are unknown parameters (Chandler, 2018). A feature of exponential family distributions is that the variance is a function of the mean: $Var(Y_i) = \phi V(\mu_i)$, where $V(\cdot)$ is the variance function for the distribution, and the dispersion parameter ϕ is assumed to be constant.

Second, the relationship between predictors and predictands can be non-linear. For the ith case in the dataset, the expectation of the predictand $E(Y_i) = \mu_i$ can be calculated using the following equation:

$$g(\mu_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} = \eta_i \tag{3.3}$$

where η_i is called the linear predictor; $g(\cdot)$ is a monotonic link function.

3.1.2 Univariate Weather Generation

It is necessary to choose an appropriate model for each weather variable. The currently available models in Rglimclim package include the logistic regression, gamma, normal and normal-heteroscedastic (Chandler, 2018). To explain the theory as it applies to the dataset in this study, the notation such that Y_t^s now denotes a predictand (or response variable) at site s on day t.

For precipitation, most of the weather generators used now treat the precipitation occurrence and amount separately (Chandler, 2020). Therefore, we applied the same method in this study. The precipitation occurrence (i.e. value = 0 denotes a dry day, and value = 1 denotes a wet day) is fitted using the logistic regression model:

$$\log\left(\frac{p_t^s}{1-p_t^s}\right) = (\boldsymbol{x_t^s})^{\boldsymbol{T}}\boldsymbol{\beta} \tag{3.4}$$

where p_t^s is the probability of rainfall for site s on day t; $(\boldsymbol{x_t^s})^{\boldsymbol{T}}$ is the row vector of covariates. $\boldsymbol{\beta}$ is the unknown parameter vector.

The precipitation amount considers wet days only. It is fitted by the gamma distribution because it only contains positive values. The mean of response variable μ_t^s can be written as follow:

$$\mu_t^s = \exp((\boldsymbol{\zeta_t^s})^{\boldsymbol{T}}\boldsymbol{\beta}) \tag{3.5}$$

where $(\boldsymbol{\zeta_t^s})^{\boldsymbol{T}}$ is the row vector of covariates. $\boldsymbol{\beta}$ is the unknown parameter vector.

Temperature is fitted by the normal-heteroscedastic model. This model is chosen because the variance of precipitation changes with the month (see Figure 3.1). Therefore, instead of using the normal distribution, the normal-heteroscedastic model is used:

$$Y_t^s \sim N(\mu_t^s, (\sigma_t^s)^2)$$

$$\text{with} \qquad \mu_t^s = (\boldsymbol{\tau_t^s})^{\boldsymbol{T}}\boldsymbol{\beta} \tag{3.6}$$

$$\text{and} \quad \log((\sigma_t^s)^2) = (\boldsymbol{\xi_t^s})^{\boldsymbol{T}}\boldsymbol{\gamma}, \tag{3.7}$$

where row vectors of covariates are $(\boldsymbol{\tau_t^s})^{\boldsymbol{T}}$ and $(\boldsymbol{\xi_t^s})^{\boldsymbol{T}}$ for mean and variance respec-

tively. $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are two unknown parameter vectors.

3.1.2.1 Model fitting

Next, the appropriate covariates and interactions should be chosen to build the model. If the response variables are regarded as independent, then the unknown coefficient vector can be estimated by maximum likelihood estimation. However, the independence assumption is usually unrealistic for a daily multisite data set. The weather variables of successive days at a site and the weather variables of neighbouring sites in a day are expected to be similar. In other words, temporal and spatial dependencies exist.

To deal with temporal dependency, the functions of the previous days' response variables are added to the model as covariates to represent the temporal autocorrelation. On the other hand, spatial dependence is accounted for by estimating the parameters using a likelihood constructed as though sites are independent, and then adjusting standard errors and likelihood ratios to account for the dependence.

3.1.2.2 Model comparison and checking

To compare different models, the methods used include likelihood ratio tests, analysis of variance (ANOVA) tests, and root mean squared error (RMSE) tests. For instance, if a much more complex model leads to only a slight decrease in RMSE, it suggests that it is not worth using the more complex model.

Apart from the formal model comparisons, some diagnostic methods can be used to assess the model. Rglimclim package constructs the graphs mostly based on the Pearson residuals r_t^s, which is defined as $r_t^s \propto \frac{y_t^s - \mu_t^s}{\sigma_t^s}$, where y_t^s represents the response variable at site s on day t; μ_t^s and σ_t^s represent the mean and standard deviation respectively of the generating distribution. Suppose the model successfully captures the systematic structure in the data. In that case, the Pearson residuals will have zero mean and constant variance, and thus no obvious systematic structure will be shown in the residual plots (Chandler, 2020).

3.1.3 Bivariate Weather Generation

The next step is switching from univariate to bivariate weather generation. Since the precipitation and temperature behave differently and only temperature follows a normal distribution, it is impossible to build a regression model for both variables simultaneously. Alternatively, a joint probability density function is built, and the standard factorisation can be applied to it (Chandler, 2018).

Suppose $\mathbf{Y}_i = \left(Y_{i1}, Y_{i2} \right)^{\boldsymbol{T}}$ is the collection of two variables for ith case in the data. The joint density for $\mathbf{Y}_i$ can be factorised into a product of conditional densities given the corresponding vector of covariates $\mathbf{x}_i$:

$$f(\mathbf{y}_i \mid \mathbf{x}_i) = f_1(y_{i1} \mid \mathbf{x}_i) \times f_2(y_{i2} \mid y_{i1}, \mathbf{x}_i) \tag{3.8}$$

The order of two variables can affect the resulting bivariate model. For instance, if precipitation is selected as the first variable and temperature is conditional on precipitation occurrence, then the resulting marginal temperature distribution for each day will be a combination of one normal distribution for dry days and another normal distribution for wet days, which is bimodal (i.e. distribution with two modes). In contrast, if the first variable is temperature, then the marginal temperature distribution will be normal and unimodal (Equation (3.6) and Equation (3.7)) (Chandler, 2018).

To decide the order of the variables, it might be useful to consider the following (Chandler, 2018): First, check whether physical mechanisms exist between variables. If a variable appears to strongly influence or determine the values of other variables, it would be sensible to take it as the first variable and derive other variables. However, the relationship between precipitation and temperature is not obvious.

The second consideration is the number of data. Modelling data-rich variables first in a GLM is more practical because it allows for using all available data, while modelling them conditional on data-poor variables results in fewer observations being available. In our case, the number of precipitation data is 1,111,711 and the number of temperature data is 1,028,214. However, since the total number of obser-

vations is very large, the number difference between precipitation and temperature data is considered not too big.

The third thing to assess is the performance of each marginal distribution. The variable that has a better modelling performance can be selected as the first variable. This will give a better resulting model.

3.1.4 Simulation

To generate simulations, the random numbers are produced from different predicted distributions. When simulating time series at multiple spatial locations, it is essential to ensure realistic inter-site dependence. Rglimclim package allows using different dependence structures based on different distributions. For continuous distributions, transform the response variables to approximately normal distributed first, and then define the inter-site dependence based on the transformed scale. The reason for doing this is that among all multivariate distributions, the multivariate normal distribution is the only one whose dependence structure can be completely characterised by correlations. For normal-heteroscedastic model, the spatial dependency can be directly represented by correlations between standardised residuals. For gamma distribution, inter-site dependence is represented via correlations between Anscombe residuals, which is defined as $(y_t^s/\mu_t^s)^{1/3}$ (Chandler, 2018).

To represent the spatial dependency of the logistic model: If the study area is small, the model can be built based on the distribution of the total number of sites that receive precipitation on a given day. Suppose the study area is large, such as Rhineland-Palatinate in this study. In that case, the model can be built based on the correlation structure of latent Gaussian processes, so the dependence between distant pairs of sites is weaker. Suppose latent Gaussian variable Z_t^s is the standard normal random variable at site s on day t, then Y_t^s is set to 1 if $Z_t^s > -\Phi^{-1}(p_t^s)$, where $\Phi(\cdot)$ represents the standard normal distribution function and $p_t^s = P(Y_t^s = 1)$ (Chandler, 2020).

3.2 Machine Learning Method

This section explains the implementation of two machine learning methods, Bayesian network and random forest for statistical downscaling. The package used for Bayesian network is BNWeatherGen (Legasa and Gutiérrez, 2020), available at https://github.com/MNLR/BNWeatherGen. The package used for random forest is RandomForestDist (Legasa et al., 2022), available at https://github.com/MNLR/RandomForestDist. In this study, Bayesian network is used to generate precipitation occurrence, and random forest is used to generate both precipitation amount and temperature.

3.2.1 Bayesian Network

This section will first talk about the general theory of Bayesian networks (Subsection 3.2.1.1) and Bayesian Network Learning (Subsection 3.2.1.2), and then their use as weather generators (Subsection 3.2.1.3).

3.2.1.1 Theory of Bayesian Network

Bayesian network (Scutari and Denis, 2014) is a probabilistic graphical model that uses both probability theory and graphs to learn from the joint probability distribution of a group of discrete random variables. It provides a visual representation of the relationships between variables through a graph, making it easier to interpret. Suppose there is a set of discrete random variables $X_1, X_2, \ldots, X_n$ characterising the quantities of interest (in this study, precipitation occurrence in a network of n stations, where $n = 87$), the joint probability distribution $P(X_1, \ldots, X_n)$ will have 2^n possible categories, which means $2^n - 1$ parameters are needed. As a result, the number of parameters will increase exponentially with the increase in the number of variables. This can be problematic because even if it is computationally possible, it is unlikely that we will have a large enough dataset to adjust such a large number of parameters (Legasa and Gutiérrez, 2020).

To solve this problem, Bayesian network method uses a directed acyclic graph (DAG). Each node represents a variable X_i, and each directed arc represents direct probabilistic dependencies. The node at the tail of the arc is defined as the parent,

while that at the head (where the arrow is) is defined as the child (e.g. $A \rightarrow B$ means B depends on A, and B is a child of A, where A and B are two nodes and $\rightarrow$ is the arc between them in a DAG) (Scutari and Denis, 2014). The DAG is shown in Appendix, Figure B.6 indicates the spatial dependence structure among 87 stations in this study.

On the other hand, if no arc connects the two nodes, the associated variables are conditionally independent, given a subset of the other variables. As a result, the joint probability distribution can be factorised as the following equation by the chain rule of probability:

$$P(X_1,\ldots,X_n) \approx P_{dag}(X_1,\ldots,X_n) = \prod_{i=1}^{n} P(X_i \mid \pi(X_i)) \tag{3.9}$$

where $\pi(X_i)$ refers to the set of parents of node X_i. $i = 1,\ldots,n$, n = 87 in this case. This factorisation can now be used to simulate the joint distribution in such a way as to respect the dependencies between the variables. Its connection with weather generation will be discussed in the Subsection 3.2.1.3.

The right-side factorisation in Equation 3.9 requires the conditional probability tables (CPTs), which specify the conditional probability distribution of child variables given their parents' states in the DAG. Specifically, the number of parameters needed for each CPT is equal to $(s(X_i)-1)*s(\pi(X_i))$, where $s(\cdot)$ refers to the number of states of the variable or combination of variables. As a result, the overall joint probability model only requires a reasonable number of parameters (approximately 2000 parameters in this study), which leads to efficient data-driven models (Scutari and Denis, 2014).

3.2.1.2 Bayesian Network Learning

For Bayesian network, the procedure of model selection and estimation is called learning. There are two typical aspects of Bayesian network learning: structure learning (i.e. learning the structure of the DAG) and parameter learning (i.e. learning the estimation of probabilities) (Scutari and Denis, 2014).

There are two methods for structure learning: constraint-based and score-based

methods (Scutari and Denis, 2014). Constraint-based methods use conditional independence tests (e.g. χ^2) to learn the potential conditional dependencies and independencies from the data set. According to Scutari et al. (2019), the performance of score-based methods is better than that of constraint-based methods for climate problems (Scutari et al., 2019). Therefore, score-based methods are used in the study.

The two requirements to carry out score-based methods include the score (i.e. how well the graph represents the data set) and the optimisation algorithm (i.e. the approach to maximise the score) (Scutari and Denis, 2014).

Structure learning aims to find the graph structure that is most compatible with the data. Legasa and Gutiérrez (2020) propose to do this using the Bayesian Information Criterion (BIC), which is the score we want to maximise for the score-based methods. In this context, the BIC for a graph G and data D is defined as:

$$\mathrm{BIC}(G,D) = \sum_{i=1}^{n} \left(\log\left(P\left(X_i \mid \pi\left(X_i\right)\right)\right) - k\left|\Theta\left(X_i\right)\right|\right) \tag{3.10}$$

where k is defined as the regularization parameter, which equals to $log(n)/2$, where n is the total number of observations. $|\Theta(X_i)|$ means the number of parameters for each CPT corresponding to each node X_i under the network represented by the graph G. $P(X_i|\pi(X_i))$ is estimated by the parameter learning method.

The parameter learning method can be either maximum likelihood estimation (MLE) or Bayesian estimation using the posterior distribution (Scutari and Denis, 2014). In this study, Bayesian estimation is used because it might be helpful to incorporate prior information into the analysis.

Specifically, BNWeatherGen package chooses the uniform distribution as a prior and calculates the posterior estimates as a weighted mean of the prior and observed values as follows: Suppose we want to estimate $P(X_i = 1|X_j = 0)$, according to $P(X_i \mid X_j) = \frac{P(X_i,X_j)}{P(X_j)}$, then the posterior estimate $\widetilde{P}\left(X_i = 1 \mid X_j = 0\right)$ is:

$$\widetilde{P}\left(X_i = 1 \mid X_j = 0\right) = \frac{\widetilde{P}\left(X_i = 1, X_j = 0\right)}{\widetilde{P}\left(X_j = 0\right)} = \frac{\frac{n}{n+1}\widetilde{p}_{X_i,X_j} + \frac{1}{n+1}\pi_{X_1,X_j}}{\frac{n}{n+1}\widetilde{p}_{X_j} + \frac{1}{n+1}\pi_{X_j}} \tag{3.11}$$

where n is the number of observations; $\widetilde{p}_{X_i,X_j}$ is the number of observations where $X_i = 1$, $X_j = 0$ divided by the total number of observations; $\widetilde{p}_{X_j}$ is the number of observations where $X_j = 0$ divided by the total number of observations; π_{X_1,X_j} and π_{X_j} equal to 1/4 and 1/2 respectively because of the uniform prior distribution. Notably, $\widetilde{P}\left(X_i = 1 \mid X_j = 0\right) = 0.5$ if there are no observations to guide the parameters (Legasa and Gutiérrez, 2020).

Although the BIC score takes the value of $k = log(n)/2$ (Equation (3.10)), changing the value of k is a useful method to tailor the DAG complexity to meet our needs. The higher k is, the less complex DAG will be. In particular, if $k = 1$, it matches the Akaike information criterion (AIC). Legasa and Gutiérrez (2020) concluded that when $k = 1$, the Bayesian network weather generator performs the best. Therefore, $k = 1$ is also used in this study.

The second requirement for the score-based method is the optimisation algorithm. The Tabu search algorithm is a classical method to find the maximum score in the discrete space of DAGs (Legasa and Gutiérrez, 2020). It starts from a network structure (typically with no arc), and then gradually modifies the structure by adding, removing, or reversing one arc every time until no further improvements in the score can be made (Scutari and Denis, 2014). The algorithm does not guarantee achieving a global optimum, but it has been reported to find effective structures in a wide range of applications (e.g. Legasa and Gutiérrez 2020).

3.2.1.3 Bayesian Networks as Weather Generators

Suppose random variable $\boldsymbol{X}_t^s = (X_t^1, \ldots, X_t^n)$, where $\boldsymbol{X}_t^s$ is the precipitation occurrence (0 for a dry day and 1 for a wet day) at station $s = 1, \ldots, n$ (n = 87 in this study) at time t, a multivariate weather generator can be regarded as a model to obtain samples from the distribution $P(\boldsymbol{X}_t^s | \boldsymbol{X}_{t-1}^s, \boldsymbol{X}_{t-2}^s, \ldots)$. It is usually simplified as a Markov-1 process, which means the precipitation occurrence on the present day only depends on that on just the day before:

$$P\left(\mathbf{X}_t^s \mid \mathbf{X}_{t-1}^s, \mathbf{X}_{t-2}^s, \ldots\right) \approx P\left(\mathbf{X}_t^s \mid \mathbf{X}_{t-1}^s\right) = P\left((X_t^1, \ldots, X_t^n) \mid \mathbf{X}_{t-1}^s\right) \tag{3.12}$$

Combine this equation with the factorisation of JPD (Equation 3.9), the following equation can be obtained:

$$P(\mathbf{X}_t^s \mid \mathbf{X}_{t-1}^s) = P((X_t^1, \ldots, X_t^n) \mid \mathbf{X}_{t-1}^s) \approx \prod_{s=1,\ldots,n} P(X_t^s \mid \pi(X_t^s), \mathbf{X}_{t-1}^s) \tag{3.13}$$

In addition, assume X_t^s is conditionally independent of all X_{t-1}^j, where $s \neq j$ (i.e. the precipitation occurrence in station s on the present day is independent of precipitation occurrences in other stations on the day before, conditional on X_{t-1}^s and on $\boldsymbol{X}_t^s$), Equation 3.13 can be simplified as follows:

$$P(\mathbf{X}_t^s \mid \mathbf{X}_{t-1}^s) \approx \prod_{s=1,\ldots,n} P(X_t^s \mid \pi(X_t^s), X_{t-1}^s) \tag{3.14}$$

This factorisation allows each present node X_t^s to be simulated as long as we know the value of its corresponding past node X_{t-1}^s and the node's present parents $\pi(X_t^s)$. This Bayesian network is called Markov. Legasa and Gutiérrez (2020) also described other extended methods apart from Markov, but in this study, we focus on Markov for its high computation efficiency.

After learning the spatial structure (i.e. learning from stations) and the temporal structure (i.e. learning from past nodes), Bayesian network can then generate weather series by an iterative process with ancestral order (i.e. from parents to children). Begin with initial observations $\mathbf{X}_0^s(X_0^1 = x_0^1, ..., X_0^n = x_0^n)$, Bayesian network generates random synthetic initial observations using Equation (3.9). Then, calculate x_1^1 using $P(X_t^1 | \mathbf{X}_{t-1}^s = \mathbf{X}_0^s)$, then proceed with $P(X_t^2 | X_t^1 = x_1^1, \mathbf{X}_{t-1}^s = \mathbf{X}_0^s)$, and so on. After calculating $\mathbf{X}_1^s = (X_1^1 = x_1^1, ..., X_1^n = x_1^n)$, the iterative process will continue (Legasa and Gutiérrez, 2020).

All station data are regarded as predictors to the model, and precipitation occurrences are regarded as predictands. The large-scale predictors are not included in the input data This approach is consistent with that from Legasa and Gutiérrez (2020). Finally, a series of precipitation occurrences for each station will be simulated.

3.2.2 Random Forest

Random forests are used to model and simulate the precipitation amount and temperature. This section will first focus on the general theory of random forest (Subsection 3.2.2.1), and then explain how a posteriori random forest is used to generate weather series (Subsection 3.2.2.2).

3.2.2.1 Theory of Random Forest

Classification and regression trees (CART) are predictive algorithms that divide the predictor space into regions that are homogeneous with respect to the response variable (Legasa et al., 2022). CART explore the variance of a response variable based on one or more explanatory variables. The response variable is typically either categorical (as in classification trees) or numeric (as in regression trees). The explanatory variables can be categorical and/or numeric. To construct a tree, the data are repeatedly split according to the if-else rules that are defined by a single explanatory variable. Each split partitions the data into two mutually exclusive groups that are as homogeneous as possible with respect to the response variable. Then for each group, the splitting process is repeated. The goal is to divide the predictor space into homogeneous regions for the response variable, while simultaneously keeping the tree reasonably small. The trees can be represented by graphs. The root node at the top represents the undivided data. There are branches and leaves below it, and each leaf represents one of the final groups (De'ath and Fabricius, 2000).

One advantage of CART is that they can deal with non-linear relationships in the data, and thus they are ideal tools for analysing complex environmental data (De'ath and Fabricius, 2000). However, the disadvantages of CART are its lack of stability and the risk of overfitting (Legasa et al., 2022).

Random forest alleviates these problems by constructing K CART to predict random variables Y (in this study, precipitation amount and temperature) from the predictors X (in this study, large-scale predictors). Each CART is trained with a bootstrap sample (i.e. random samples taken from a data set with replacement) from the original data set. Then, when making splits in each tree, only a random subset of all available predictors is used.

There are two additional ways to reduce the problem of overfitting. First, control the depth of the tree (i.e. the number of splits). The way to control the depth is either directly set the maximum depth or by setting the minimum leaf size (i.e. the minimum number of elements for each leaf). Second, control the total number of trees (Legasa et al., 2022).

According to Legasa et al. (2022), each tree t_k maps a set of predictors $x \in X$ to a leaf that contains a set $\{y \in Y\}_{t_k(x)}$. The split function is used to construct the mappings. It recursively (i.e. the algorithm repeats this process multiple times until it reaches the final groups) selects the predictor variable X^0 and its threshold value $X^0 = x_0^0$ that best divides the observed subset Y^0 from Y into $Y_+ = \{y \in Y_0 \mid X^0 \geq x_0^0\}$ and $Y_- = \{y \in Y_0 \mid X^0 < x_0^0\}$. For the mean observations $\bar{y}_0, \bar{y}_+, \bar{y}_-$ of each subset, $X^0 = x_0^0$ should be appropriately chosen to maximise the reduction of the average error committed in the leaves $Err(\bar{y}_0, y) - \big(Err(\bar{y}_+, y) + Err(\bar{y}_-, y)\big)/2$, where $Err(\bar{y}, y)$ is defined as the split function and represents how accurately $\bar{y}$ represents the observed data for each set.

Different split functions should be chosen for different distributions. For precipitation amount, we assume it follows a gamma distribution, which is the same assumption we made for the GLM. Legasa et al. (2022) found that the best performance split function for precipitation amount is gamma deviance. The formula given by Legasa et al. (2022) is:

$$deviance\big(\bar{y}, \{y\}\big) = 2 \sum_{y \in \{y\}} \left[-\log\left(\frac{y}{\bar{y}}\right) + \frac{y - \bar{y}}{\bar{y}} \right] \tag{3.15}$$

Temperature is assumed to follow a normal distribution. The split function used in this study is mean squared error (MSE), which is $Err(\bar{y}_0, y) = \frac{1}{|Y_0|} \sum_{y \in Y_0} (y - \bar{y}_0)^2$.

After building K trees using the selected split function, for a given predictor value x, the predictions are averaged over all trees to obtain the final prediction $\hat{y}$, that is:

$$\hat{y}(x) = \frac{1}{K} \sum_{k=1}^{K} \bar{y}_{t_k(x)} \tag{3.16}$$

This is defined as the averaging approach (AVG-RF).

3.2.2.2 A Posteriori Random Forests for Weather Generation

Legasa et al. (2022) suggested that compared to AVG-RF, a posteriori random forest (AP-RF) method makes more precise estimations of the shape and better scales parameters of the probability distribution $Y \mid X$ without losing predictive power, thus generating more reliable weather series. Therefore, instead of using AVG-RF which makes a single final prediction (as Equation 3.16), we use AP-RF which merges all predictions into a common set $\bigcup_{k=1\ldots K}\{y \in Y\}_{t_k(x)}$ to estimate the parameters of the predicted probability distribution:

$$\hat{y}(x) = \Theta\left(\bigcup_{k=1}^{K}\{y \in Y\}_{t_k(x)}\right) \tag{3.17}$$

where Θ refers to the particular procedure to estimate the parameters.

For precipitation amount, Legasa et al. (2022) compared different estimators as the procedure Θ to estimate the parameters of the gamma distribution, and found that BC3 estimator is the least biased and has high computational efficiency. It is defined α as follows:

$$\alpha_{\mathrm{BC3}} = \dot{\alpha} - \frac{1}{n}\left(3\dot{\alpha} - \frac{2\dot{\alpha}}{3(1+\dot{\alpha})} - \frac{4\dot{\alpha}}{5(1+\dot{\alpha})^2}\right) \tag{3.18}$$

where $\dot{\alpha} = \bar{y}/\sum_{y\in\{y\}}(y-\bar{y})\log(y)$. The other parameter β can be then calculated by $\beta_{\mathrm{BC3}} = \alpha_{\mathrm{BC3}}/\bar{y}$.

For temperature, the procedure Θ to estimate the parameters of the normal distribution is MLE.

As a result, the AP-RFs produce a gamma distribution and a normal distribution for precipitation amount and temperature respectively. Based on these predicted distributions, we can generate random samples to create predictive simulations that can be used to validate the accuracy of the downscaled probability distribution (Legasa et al., 2022).

3.3 Summary

This chapter introduces the background knowledge of the statistical and machine learning methods used for downscaling. For the statistical method, GLM is used to build the weather generators. Firstly, build a univariate weather generator for different weather variables. Precipitation occurrence and amount are treated separately. Precipitation occurrence is modelled using a logistic regression model, and precipitation amount is modelled using a gamma model. Temperature is fitted by a normal-heteroscedastic model. In the process of model fitting, temporal and spatial dependencies should be considered by incorporating appropriate functions into the model. To compare models, the methods used include likelihood ratio tests, ANOVA tests, and RMSE tests. Diagnostic methods based on Pearson residual can be used to check the performance of models. Moving from a univariate to a bivariate weather generator, a factorisation can be applied to build a joint probability density function. The order of the variables should be considered during factorisation. To generate simulations, the random numbers are produced from different predicted distributions. Different inter-site dependences are constructed based on different distributions.

For the machine learning methods, Bayesian network is used to model precipitation occurrence, and random forest is used to model precipitation amount and temperature. Bayesian network uses CPT to specify the probability of a variable and uses DAG to represent the relationships between variables. The learning of Bayesian network has two aspects. The first aspect is structure learning. The score-based method with AIC as the score and Tabu search as the optimisation algorithm is used. The second aspect is parameter learning. Bayesian estimation with a uniform prior is used. After Bayesian network learns the spatial and temporal structure, it will generate the weather series by ancestral order. Random forest combines multiple CART to predict the response variables. The complexity of the forest should be controlled to avoid the overfitting problem. Gamma deviance is used as the split function for precipitation amount data and MSE is used as the split function for temperature data. In particular, AP-RF is used, which merges all predictions from

trees to estimate the parameters of the distribution. The estimator used is BC3 and MLE for precipitation amount and temperature respectively. Simulations can be generated by creating random samples based on the predicted distributions.

Chapter 4

Model Building

This chapter explains the process of model building. To robustly evaluate the performance of different downscaling methods, cross-validation should be used for any meaningful time series validation (Maraun et al., 2010). The idea behind cross-validation is that the data used for validation are independent of the data used for model calibration. To achieve this, model fitting is typically done on one subset of the data (the training set), and performance is then assessed using another subset (the validation set). In this study, 4-fold cross-validation is used. The study period is from 1959 to 2022, which contains 64 years in total. Divide it into 4 non-overlapping folds, each containing 16 years: $F_1 = 1959 - 1974$, $F_2 = 1975 - 1990$, $F_3 = 1991 - 2006$, $F_4 = 2007 - 2022$. The model is trained 4 times, validated by one fold and trained by the remaining 3 folds each time. Then the out-of-sample predictions that cover the whole period from 1959 to 2022 will be simulated.

For machine learning methods, the models can be trained automatically, so it is easy to carry out cross-validation. However, GLMs are built manually, so cross-validation for GLMs will be more difficult. As suggested by (Maraun et al., 2015), an easier way to do it is to use the complete data set to choose the model structure, and then refit fold.

4.1 Statistical Model

This section explains how to identify the structure of the GLM using all of the data. According to Chandler (2020), the building process of GLM will be first carried out

for precipitation (including both precipitation occurrence and amount) and temperature (including both mean component and dispersion component) separately. Then, link the temperature and precipitation models to build a bivariate model. Finally, incorporate the large-scale atmospheric predictors into the model as covariates.

When fitting the model, start with the simplest models with no covariate. Then gradually add the covariates, with the most significant covariates first to quickly build a reasonable model. The typical covariates include seasonality, regional variation(e.g. different degrees of Legendre polynomials can be used to represent the longitudinal and latitudinal effects), autocorrelation (i.e. the correlation between the weather variables at present and that on previous days), the interactions among these covariates, and so on. To compare different models, the assessment methods include formal model comparisons (e.g. likelihood ratio tests) and diagnostic methods (e.g. Pearson residuals plots). Insignificant covariates should be removed. Simultaneously, the model structures should be mathematically coherent. For instance, when using Fourier representation to show seasonality, sine and cosine components are usually considered in pairs (Chandler, 2020).

4.1.1 First Stage: Univariate Models

4.1.1.1 Precipitation Occurrence

As mentioned in Chapter 3, precipitation is treated separately as precipitation occurrence and amount. The threshold of 0.95 *mm* is defined as a wet day. That is, if the precipitation is denoted by Y then the model is fitted to Y^*, where $Y^* = 0$ if $Y < 0.95$, Y - 0.95 otherwise.

Following the process described above, the final model for first-stage precipitation occurrence is based on the logistic regression model. The seasonal variation is represented by the Fourier series of the annual cycle $\sin(2\pi \times \text{day of year}/366)$ and $\cos(2\pi \times \text{day of year}/366)$ (i.e. annual cycle at a daily timescale that is calculated as though every year is a leap year) and the first harmonic of the daily annual cycle $\sin(2\pi \times \text{day of year}/183)$ and $\cos(2\pi \times \text{day of year}/183)$. The regional variation is represented by Legendre polynomials of degrees up to 2 (i.e. quadratic transformation) in latitude and longitude, along with the interactions between latitude and

longitude. It is also worth adding mean altitude over the domains with a distance of $10 \times 10\ km^2$, east-west slope and north-south slope over the domains with a distance of $30 \times 30\ km^2$ as covariates to the model (these quantities were defined in Section 2.3).

To represent the autocorrelation, a weighted average of the indicators for the precipitation up to the previous three days is used (i.e. indicators $I\left(Y^s_{t-1} > 0\right)$, $I\left(Y^s_{t-2} > 0\right)$ and $I\left(Y^s_{t-3} > 0\right)$, and the indicator will take the value 1 if $Y^s_{t-k} \neq 0$, 0 otherwise), with using weights that decrease exponentially with distance from the current site s (i.e. $\sum_r w_{r,s} I\left(Y^r_{t-1} > 0\right)$, $\sum_r w_{r,s} I\left(Y^r_{t-2} > 0\right)$, $\sum_r w_{r,s} I\left(Y^r_{t-3} > 0\right)$ where the weights $w_{r,s}$ sum to 1 and are proportional to $\exp[-ad_{r,s}]$, a is the unknown parameter, $d_{r,s}$ is the distance between site r and site s). Since the study area is relatively large, the inter-site dependence is constructed based on the latent Gaussian process model (see Chapter 3). It is specified by an exponential correlation function $\rho(i,j) = \exp\left[-\phi d_{ij}\right]$, where d_{ij} denotes the Euclidean distance between sites i and j, parameter ϕ will be estimated by maximum likelihood.

Some interactions are considered. For example, the strength of autocorrelation may vary regionally and seasonally. Therefore, the interactions between seasonal effects and autocorrelation, regional effects and autocorrelation, and regional and seasonal variations are also added to the model.

The performance of the final marginal precipitation occurrence model can be assessed by several diagnostic plots. There is no obvious systematic seasonal structure shown in the monthly and annual residual means and standard deviations plots (Appendix, Figure B.7). Also, the exponential correlation function shows a good fit to represent the spatial variation from inter-site correlation plot (Appendix, Figure B.8). However, from the mean Pearson residual by each site plot (Appendix, Figure B.9), it can be seen that slightly more sites at higher altitude have negative residuals, and more sites at lower altitude have positive residuals, even though mean altitude is already added as a covariate in the model. The other issue is that there are many sites with mean Pearson residuals significantly different from 0 at 5% level, with no obvious systematic structure. According to Chandler (2020), this is a typ-

ical phenomenon when modelling precipitation occurrence. The potential reasons include very local-scale controls on precipitation and variations in the methods or instruments used for recording precipitation.

4.1.1.2 Precipitation Amount

The model for precipitation amount is based on gamma distribution, as described in Chapter 3. Similar to the precipitation occurrence model, seasonality is represented by the Fourier series of the daily annual cycle. Also, the first harmonic is added to enhance the seasonal cycle. To represent the regional variation, Legendre polynomials of degrees up to 2 in latitude and longitude, along with the interactions between latitude and longitude are added to the model. In addition, the mean altitude over the region with a centre distance of 10×10 km^2, east-west slope and north-south slope over the region with a centre distance of 30×30 km^2 are added as covariates. To represent autocorrelation, add the weighted average precipitation of the previous one day using weights that decrease exponentially with distance from the current site s, and then apply a log-transformation of the averaged value (i.e. $\ln\left(1+\sum_r w_{r,s} Y_{t-1}^r\right)$). The spatial structure is represented by a powered exponential correlation function $\rho(i,j) = \lambda \exp\left[-\phi d_{ij}^{\kappa}\right]$, where parameters ϕ, κ, λ are estimated by maximum likelihood. Significant interactions include the interaction between autocorrelation and seasonality, and interactions between autocorrelation and spatial variation.

The performance of the final marginal precipitation amount model is similar to that of the precipitation occurrence model. No obvious systematic seasonal structure is shown in the monthly and annual residual means and standard deviations plots (Appendix, Figure B.10). The Q-Q plot illustrates a good fit of a Gamma distribution. The powered exponential correlation function captures the inter-site structure well (Appendix, Figure B.11). However, slightly more sites at higher altitude have negative residuals, and more sites at lower altitude have positive residuals. Also, there are many sites with mean Pearson residuals significantly different from 0 at 5% level, with no obvious systematic structure.

4.1.1.3 Temperature

The model for temperature is based on normal-heteroscedastic distribution, as described in Chapter 3. Since the variance is not constant, the mean and variance components are modelled separately. Start by building the model for mean components, again following the process outlined at the start of Section 4.1. Seasonality is represented by the Fourier series of the daily annual cycle and the first harmonic of the daily annual cycle. Altitude is expected to affect the temperature, so it is added as a covariate. The regional effect is represented by Legendre polynomials of degrees up to 2 in latitude and longitude, as well as the interactions between latitude and longitude. Autocorrelation is represented by the average temperature value over all sites up to the previous three days (i.e. $S^{-1}\sum_r Y^r_{t-1}$, $S^{-1}\sum_r Y^r_{t-2}$, $S^{-1}\sum_r Y^r_{t-3}$, where S denotes the number of contributing sites ($S = 87$ in this study)). The spatial structure is represented by the correlation function $\rho(i,j) = \alpha + (1-\alpha)\exp\left[-\phi d_{ij}^{\kappa}\right]$, with parameters ϕ, κ, α to be estimated by maximum likelihood. Interaction effects include the interaction between autocorrelation and seasonal variation, autocorrelation and regional variation, and seasonal and regional variation. For the dispersion component, the covariates include altitude, Legendre polynomials of degrees up to 2 in latitude and longitude, along with the interactions between latitude and longitude, daily annual cycles and their first harmonic.

Compared to the marginal precipitation model, the marginal temperature model captures the mean residual by each site more accurately (as shown in Appendix, Figure B.13), and no obvious systematic structure and smaller mean residuals. The correlation function used shows a good fit in the inter-site correlations (Appendix, Figure B.14). However, the lower tail of the residual distribution that is heavier than normal can be seen from Q-Q plot (Appendix, Figure B.14), but this is perhaps not too important because the percentage of variance explained is high (90.2%). From the plot of annual residual means, it can be seen that there is a clear increasing trend of mean residual with years. The large-scale temperature is anticipated to explain this.

4.1.2 Second Stage: Bivariate Models

As mentioned in Chapter 3, a bivariate model can be built using factorisation by adding one of the variables as a covariate to the model of the other variable. This can be done by either condition precipitation on the temperature model or the other way round. The order of variables should be considered during factorisation. The physical relationship between precipitation and temperature is not obvious. Also, since the total number of observations is very large, the number difference between precipitation and temperature data is considered not too big. Therefore, in this stage, we focus on comparing the performance of models. Second-stage models with both two orders are built, and then the performance of both first- and second-stage models are assessed to decide the order.

Firstly, build the precipitation conditioned on the temperature model. For precipitation occurrence, add the average of untransformed temperature over all sites (i.e. $S^{-1}\sum_r Y_t^r$, and Y_t^r is the temperature of site r at day t) as a covariate. The added significant interaction effects include the interaction between temperature and mean altitude, temperature and daily annual cycle, temperature and autocorrelation of precipitation. For precipitation amount, the average of untransformed temperature over all sites is also added to the model. The interactions include interaction between temperature and regional variation, and interaction between temperature and seasonal variation.

Secondly, build the temperature conditioned on the precipitation model. For both mean and variance components of temperature models, incorporating the precipitation by adding the average of untransformed precipitation over all sites as a covariate (i.e. $S^{-1}\sum_r Y_t^r$, where S denotes the number of contributing sites ($S = 87$ in this study), and Y_t^r is the precipitation of site r at day t). The interaction between precipitation and regional variation, and the interaction between precipitation and seasonal variation are added to the model.

Based on the performance of the above first- and second-stage models, temperature is chosen as the primary variable for the following several reasons: First, the first-stage temperature model performs overall better than the precipitation in-

tensity model in terms of capturing spatial variation and mean residuals by site (comparing Appendix, Figure B.12 and B.13). Although Q-Q plot of temperature model (Appendix, Figure B.14) has a heavier lower tail, it is not that important as its percentage of variance explained is high. Second, the second-stage precipitation intensity model has a low percentage of explained variability (5.9%), which means any minor deficiencies in the temperature model will not affect the simulations of precipitation. Third, the annual residual means for the first-stage temperature model have a clear upward trend over years (Appendix, Figure B.15), which possibly can be explained by conditioning on large-scale temperatures. Therefore, the temperature conditioned on the precipitation model is chosen to proceed with the model building. Notably, these results are almost the same as those that Chandler (2020) obtained for a completely different dataset in northern Iberia.

4.1.3 Third Stage: Model including Atmospheric Covariates

In the third stage, the first-stage temperature model and the second-stage precipitation model are used to incorporate large-scale atmospheric predictors.

For temperature model, among all atmospheric predictors considered (see Section 2.4), temperature, specific humidity, u-wind and v-wind, each at 1000 hPa and geopotential at 500 hPa are the most significant ones, and thus add to the model. The 2-way interactions between these atmospheric predictors and seasonality, the interactions between the atmospheric predictors and spatial variation are also added. The 3-way interaction of large-scale temperature, daily annual cycle and autocorrelation is also considered as a covariate. For instance, the correlation between the temperature at a large scale and the temperature at a local scale may vary in strength. They may have more or less consistent temperatures in the warmer months, thus stronger or weaker autocorrelation (Chandler, 2020).

From the diagnostic plots of the final third-stage temperature model (Appendix, Figure B.16), it can be seen that after adding large-scale predictors, the clear upward trend of the annual residual means (which exist in the first-stage temperature model) has now been removed.

For precipitation occurrence model, additional atmospheric covariates include

geopotential, temperature, specific humidity each at 700 hPa, u-wind and v-wind at 850 hPa, interactions between these atmospheric predictors and seasonality, the interactions between the atmospheric predictors and spatial variation. For precipitation amount model, additional atmospheric covariates include geopotential at 1000 hPa, u-wind and v-wind at 850 hPa, specific humidity at 700 hPa, interactions between these atmospheric predictors and seasonality, the interactions between the atmospheric predictors and spatial variation.

The model structures of the final precipitation and temperature models are in Appendix, A.1-A.3. It should be noted that the tables in the Appendix are presented just to emphasise the structures of the models, while the coefficients and parameter estimates will be different for each of the folds in the cross-validation exercise.

4.2 Machine Learning Model

This section describes how to build the Bayesian network and random forest models. Both machine learning methods used need a complete observed dataset to learn from data, but the observed data we use contain missing values (see Section 2.1). The Rglimclim package (which was used to fit the GLMs in Section 4.1) provides a way to impute the missing data by sampling from their distributions conditional on all of the available observations, and makes it possible to characterise the uncertainty related to missing observations from historical weather (Chandler, 2020). Most of the information in the imputation will come from the neighbouring stations and not from the precise model used, so the imputation will not bias the results in favour of the GLM. Therefore, a complete precipitation and temperature dataset is generated by Rglimclim, which is then used in machine learning methods.

Temperature is modelled by random forest alone. Precipitation is separated into two steps: precipitation occurrence is modelled by Bayesian network, and precipitation amount is modelled by random forest. After that, the final precipitation simulation can be calculated by multiplying the simulation results from the two methods. For example, for a certain site s, if the simulated precipitation occurrence on a certain day is 1, and the precipitation amount is 10 mm, then the final simula-

tion result will be 10 mm. If the simulated precipitation occurrence is 0, then this day will be regarded as a dry day, and the final simulation result will be 0 mm.

4.2.1 Bayesian Network

Bayesian network is used to model the precipitation occurrence. The first thing to do is convert the original precipitation data to precipitation occurrence (either 0 or 1). The simulated output generated by Rglimclim contains no values between zero and the threshold (0.95 mm in this study). Because as mentioned in Chapter 3, suppose Y is the original precipitation and τ is the threshold, GLM is fitted to Y^*, where $Y^* = 0$ if $Y < 0.95$, Y - 0.95 otherwise. After the simulation, the threshold is added back to any non-zero values. Therefore, set precipitation occurrence $Y^* = 0$ if $Y < 0.95$, $Y^* = 1$ otherwise for Bayesian network. All station data are regarded as predictors, and precipitation occurrence is regarded as predictands in the model.

As discussed in Chapter 3, the score-based method is used as the method for structure learning. The score to be maximised is AIC. The optimisation algorithm used to find the best DAG is Tabu search algorithm. The parameter learning method is Bayesian estimation with a uniform prior distribution. The selections of the above particular methods are all based on Legasa and Gutiérrez (2020), as they have shown the best performance in predicting precipitation occurrence. In particular, the weather generators are built based on the Markov Bayesian network, which learns temporal structure from the values of the previous one day (Equation (3.14)). Markov method is chosen because of its computational efficiency. After Bayesian network learns the spatial structure and temporal structure, it can generate weather series by an iterative process with ancestral order. To begin with, random synthetic initial observations will be generated by Bayesian network using Equation (3.9). After it obtains the initial observations $\mathbf{X}^0 = (X_1^0 = x_1^0, ..., X_n^0 = x_n^0)$, then simulates the next day conditioned on the value on the previous day iteratively, until the weather series for the whole time period is obtained.

4.2.2 Random Forest

Random forest is used to model precipitation amount with the gamma distribution, and model temperature with the normal distribution. Add all large-scale atmospheric variables for each site as predictors to learn from each site every time, and temperature or precipitation amounts are regarded as predictands.

As mentioned in Section 3.2.2, the following additional configurations are set to the forests to avoid overfitting. The depth of the trees is set to 30. Train each tree with a bootstrap sample from the original data set. For each split in each tree, a random subset of all predictors m is selected as split candidates. Legasa et al. (2022) stated that when m equals one-third of the total number of predictands, the forest performances best. In this study, the total number of predictands is 20 (20 atmospheric predictands), so $m = 20/3 \approx 6$.

Legasa et al. (2022) state that the best-performance forest has 200 trees and a minimum leaf size of 5. However, in practice terms, the training process for random forests takes a very long time (about 2 days on a research computer). In addition, Legasa et al. (2022) also state that the best-performance forest has 3-15 minimum leaf sizes. Therefore, we use a minimum leaf size of 10 for each tree and 100 trees in the forests to compromise the computational cost and model performance.

Precipitation amount only considers precipitation on wet days, so extract the precipitation value that is larger than 0 from the original precipitation dataset first (and thus the minimum value is 0.95 because there are no values between 0 and 0.95). Gamma distribution assumes values start at 0 but cannot equal 0. In order to ensure consistency with the gamma distribution, if the original precipitation is Y, then the model is fitted to $Y^* = Y - 0.949$, so that they approximately start at 0. After simulations, 0.949 will be added back to the simulated results. The precipitation amount dataset with only wet days is used to train the model. , According to Legasa et al. (2022), the split function used is gamma deviance, and the method to estimate the parameter of gamma distribution is BC3 estimator. For the temperature model, the split function used is ANOVA. The method to estimate the parameter of normal distribution is MLE.

Finally, multiply the precipitation amount simulations and precipitation occurrence simulations from Bayesian network to get the final precipitation simulations.

4.3 Summary

This chapter describes the model-building process for downscaling methods. A 4-fold cross-validation is used to train and validate the model with different data sets, and get out-of-sample simulations. For statistical method, GLM is built in 3 stages. The first stage is to build univariate models of precipitation and temperature separately. The second stage is to build bivariate models by factorisation, either precipitation conditioned on temperature or temperature conditioned on precipitation. Temperature is chosen as the primary variable based on the performance of first- and second-stage models. The methods to assess the performance include formal model comparisons and diagnostic methods. The third stage is to add the atmospheric covariates.

For machine learning methods, temperature is modelled by random forest. Precipitation is considered separately: precipitation occurrence is modelled by Bayesian network, and precipitation amount is modelled by random forest. Bayeisan network takes station data as predictors, and random forest takes large-scale atmospheric variables for each site as predictors to learn from each site every time. Both methods take precipitation and temperature as predictands. The best configurations of models are set. The final precipitation simulations can be obtained by multiplying the simulations from the two methods.

Chapter 5

Simulation Results

The performance of the fitted models can be assessed by generating the out-of-sample simulation results using cross-validation, and then comparing them with the observations. This chapter describes the simulation results of the statistical model and machine learning models separately, and then compares the two methods.

5.1 Simulations from Statistical Model

100 simulations are generated for each fold. Each simulation consists of a collection of daily time series at each of the 87 stations. Several summary statistics (e.g. mean, standard deviation, etc.) for each month of the year, and for each year of the simulation period, are used to compare the simulated and observed values. However, since the observation data set contains missing data (see Chapter 2), some uncertainty exists in weather properties. To characterise the uncertainty, 39 imputations are generated from GLMs, and the range of summary statistics computed from these 39 imputations forms a 95% uncertainty interval for the actual value. The rationale is that if 39 values are imputed to generate an observation, then there is a 1/40 chance that the observation will be smaller than all the imputed values, and a 1/40 chance of being larger. This means that there is a 5% chance in total that the observation will fall outside the range of the imputed values (Chandler, 2020).

5.1.1 Monthly Summary Statistics

5.1.1.1 Precipitation

To assess the simulations from GLM for precipitation, the monthly summary statistics over the whole region used include mean, standard deviation, extremes (maximum), autocorrelation at lag 1 to 3, the proportion of wet days, conditional mean (i.e. wet-day mean) and conditional standard deviation (i.e. wet-day standard deviation).

Figures 5.1-5.4 show the monthly precipitation summary statistics for folds 1-4 respectively, for the time series obtained by averaging over all 87 stations on each day. It is worth noting the differences between the simulated properties in different folds. For example, the seasonal structure of the simulated mean precipitation in fold 4 is very different from that in the other three folds. In contrast, the pattern of observed mean precipitation is not very different. The model seems ”overreacting” to differences in the training data between folds, which may be a symptom of overfitting.

For all folds, the simulations overestimate both unconditional and conditional means (1st and 8th plots). Especially for fold 1 and 2, the simulations of all months are overestimated. Also, for fold 3 and 4, the simulations of the proportion of wet days differ from the observations (7th plot). For these summary statistics, the simulations failed to capture the seasonality patterns of observations.

Other statistics are generally reproduced well, except for some common issues of simulations occurring in all folds: The unconditional and conditional standard deviations in January are overestimated (2nd and 9th plots); the maximums in January and December are overestimated (3rd plot).

It is also worth assessing the simulation performance for individual stations. For example, comparing the simulations of station S096 which is at a low altitude in the Rhine Valley (Appendix, Figure B.17-B.20 for fold 1-4 respectively), and station S028 which is at a high altitude in the mountains (Appendix, B.21-B.24 for fold 1-4 respectively): The performance of simulated statistics for both stations is similar, except S028 has more underestimated autocorrelations at lag 1 in a few

months. Also, the simulation performance for single stations is similar to that for the entire region.

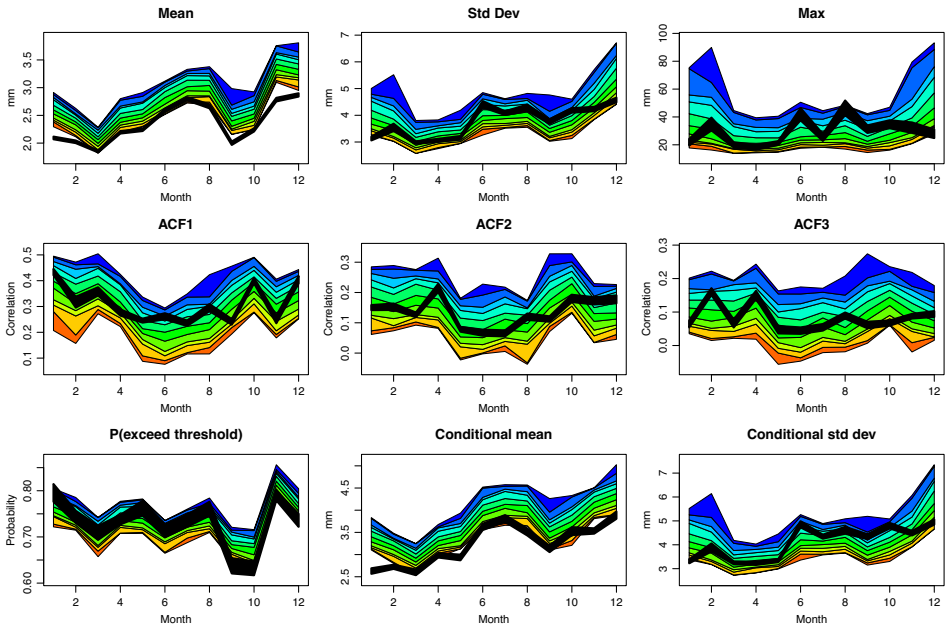

Figure 5.1: Simulated and observed monthly summary statistics for precipitation for fold 1 (1959-1974) over the whole region for the statistical model. The coloured bands show the 1st, 5th, 10th, 25th, 75th, 90th, 95th and 99th percentiles of simulated distributions. The black bands show the 95% uncertainty interval of imputations. The plots show the mean, standard deviation, maximum, auto-correlation at lag 1-3, the proportion of wet days, wet-day mean and wet-day standard deviation in order.

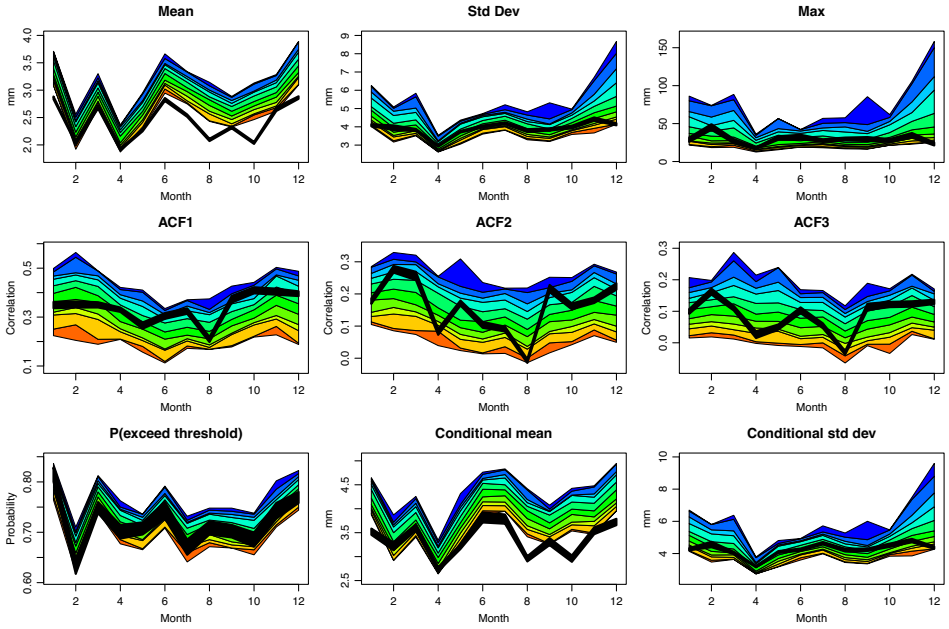

Figure 5.2: As Figure 5.1, but for fold 2 (1975-1990).

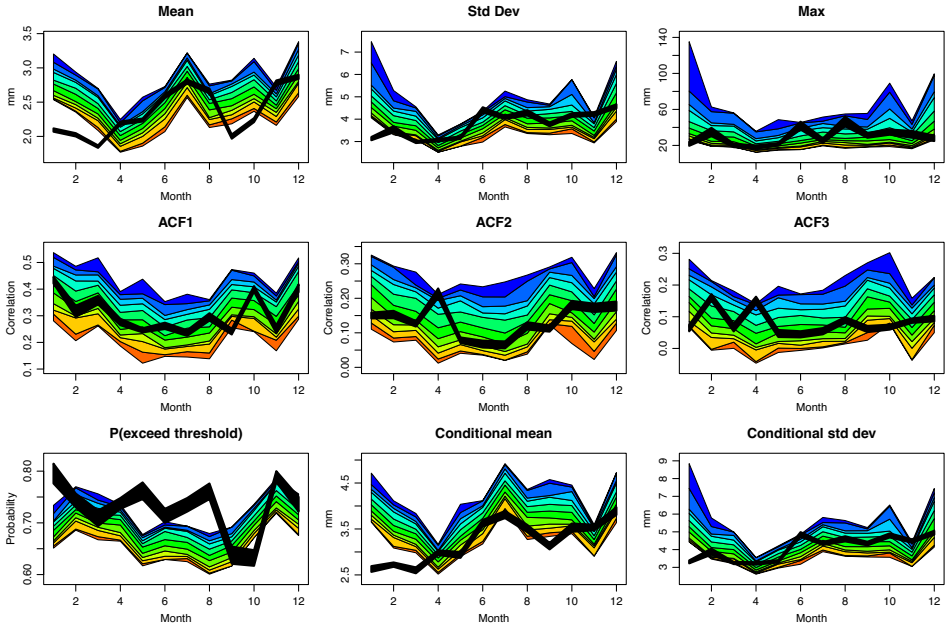

Figure 5.3: As Figure 5.1, but for fold 3 (1991-2006).

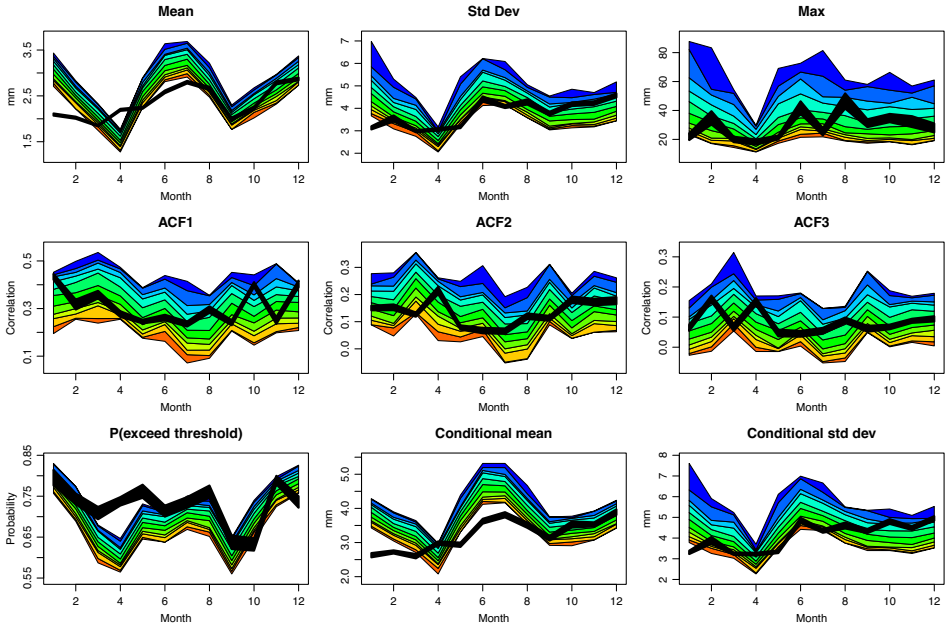

Figure 5.4: As Figure 5.1, but for fold 4 (2007-2021).

5.1.1.2 Temperature

For temperature, the monthly summary statistics used include mean, standard deviation, extremes (maximum and minimum), autocorrelation at lag 1 to 3, and correlation with temperature and precipitation.

Figures 5.5-5.8 show the monthly temperature summary statistics for folds 1-4 respectively. It can be seen that the monthly temperature results are better than the monthly precipitation results, and are more consistent between folds.

The simulated standard deviations (2nd plot) are underestimated in November, December, January and February, except for fold 4, whose simulated values generally match the observed values. The simulated minimums (4th plot) are generally overestimated in December, January and February. Other simulated statistics generally match the observed statistics well.

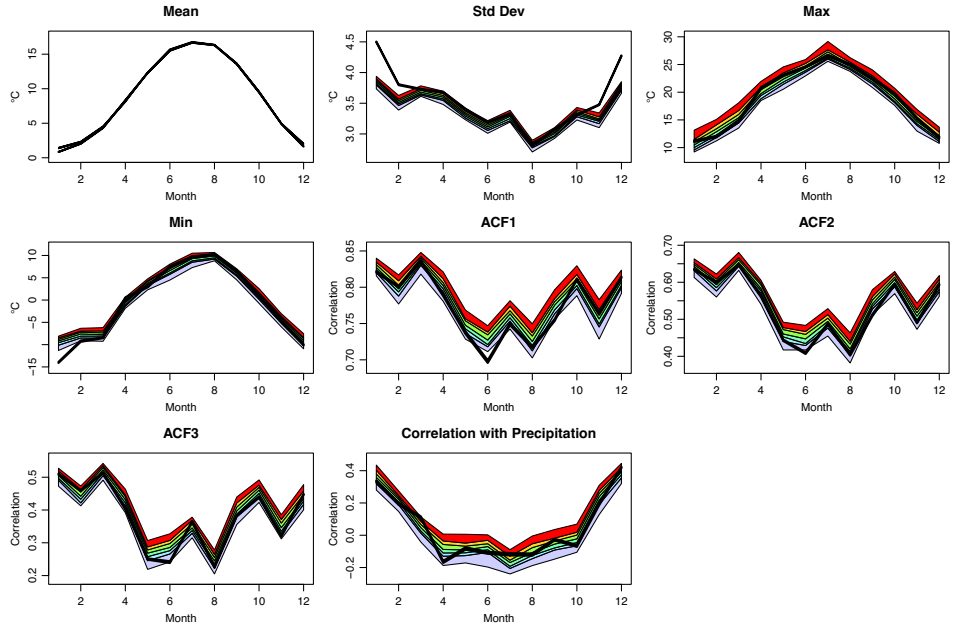

Figure 5.5: Simulated and observed monthly summary statistics for temperature for fold 1 (1959-1974) over the whole region for the statistical model. The coloured bands show the 1st, 5th, 10th, 25th, 75th, 90th, 95th and 99th percentiles of simulated distributions. The black bands show the 95% uncertainty interval of imputations. The plots show the mean, standard deviation, maximum, minimum, autocorrelation at lag 1-3, and correlation with precipitation in order.

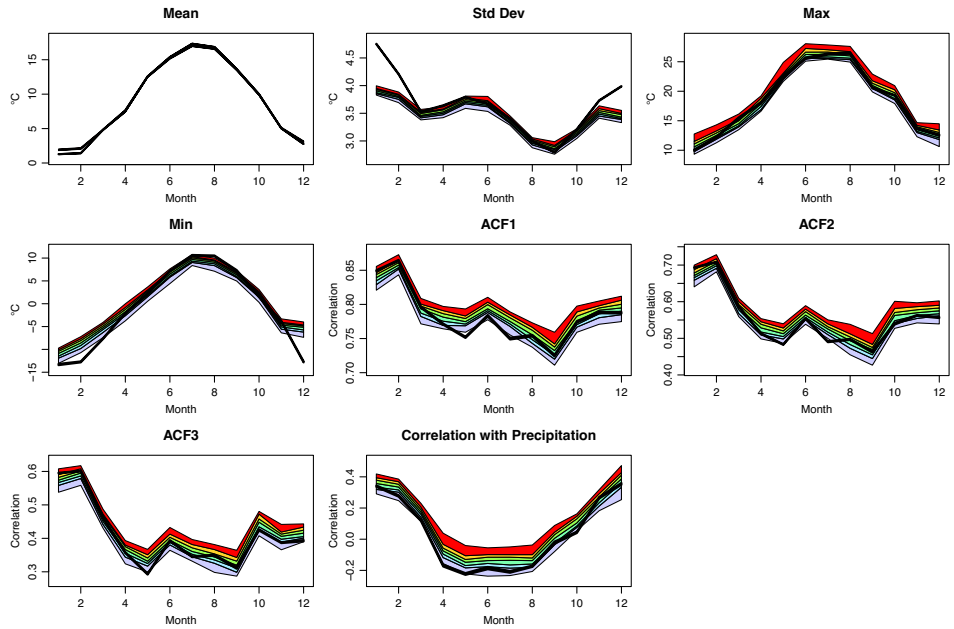

Figure 5.6: As Figure 5.5, but for fold 2 (1975-1990).

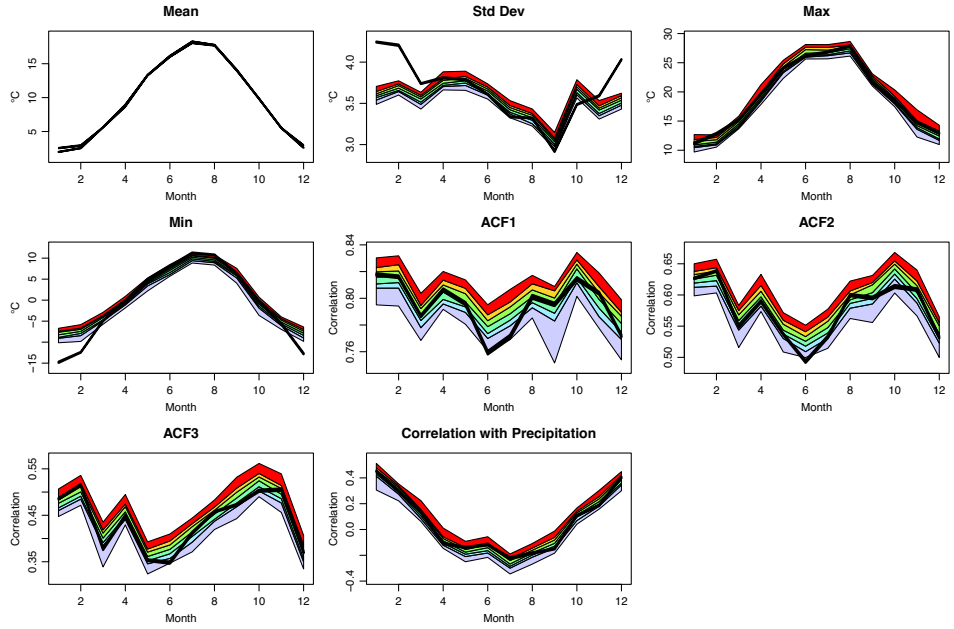

Figure 5.7: As Figure 5.5, but for fold 3 (1991-2006).

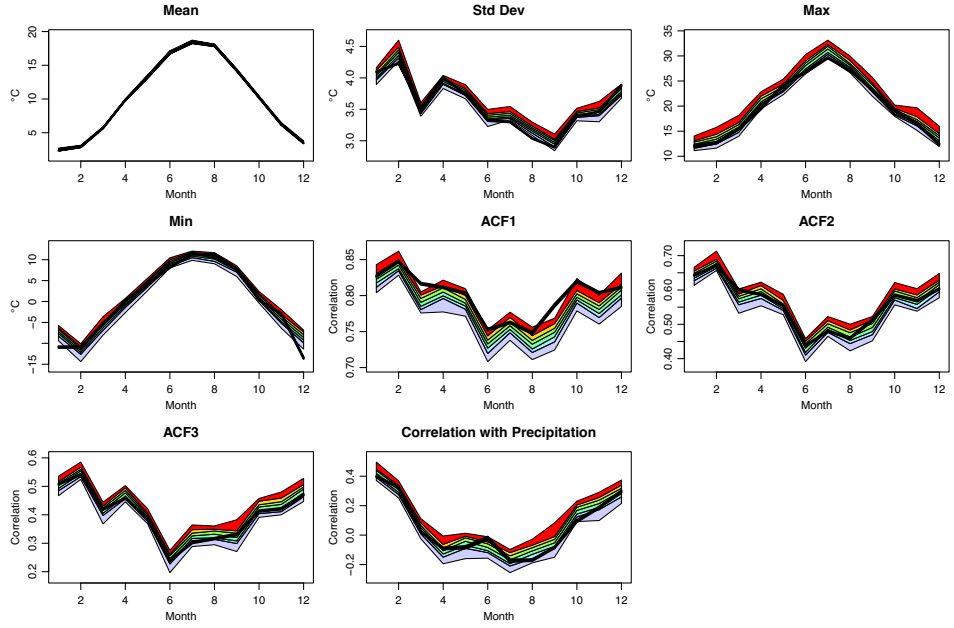

Figure 5.8: As Figure 5.5, but for fold 4 (2007-2021).

5.1.2 Seasonal Summary Statistics

5.1.2.1 Precipitation

Figure 5.9-5.12 show the distributions of seasonal precipitation totals each year, averaged over the whole region and for folds 1-4 respectively. The significant variation

of mean precipitation with years illustrates the effects of the large-scale variables on climate change. Although the simulated and observed means have a similar annual trend in general, simulated means are overestimated for all folds, which is the same as we found earlier.

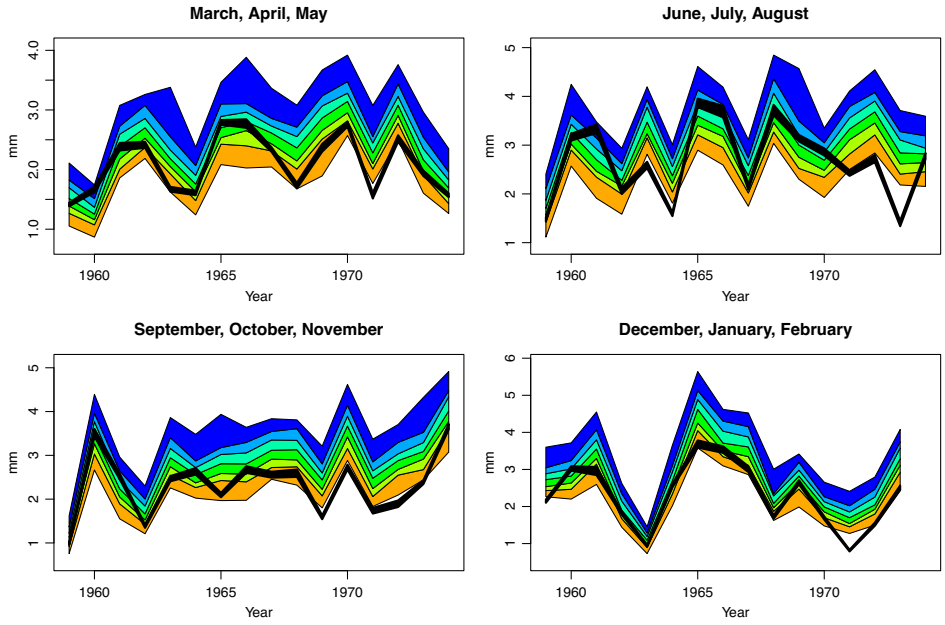

Figure 5.9: Simulated and observed annual mean for precipitation for fold 1 (1959-1974) over the whole region separated by seasons for the statistical model. The coloured bands show the 1st, 5th, 10th, 25th, 75th, 90th, 95th and 99th percentiles of simulated distributions. The black bands show the 95% uncertainty interval of imputations.

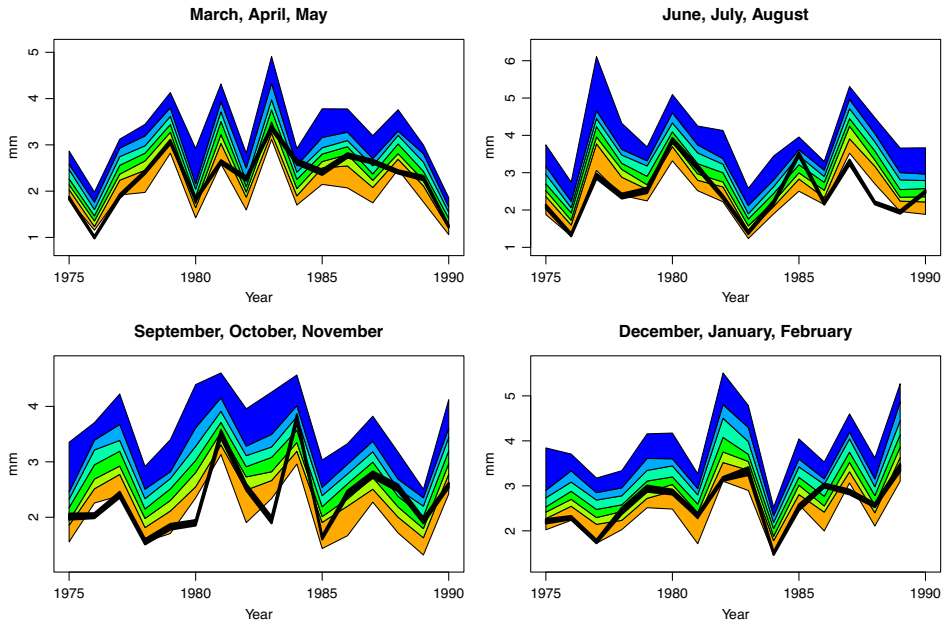

Figure 5.10: As Figure 5.9, but for fold 2 (1975-1990).

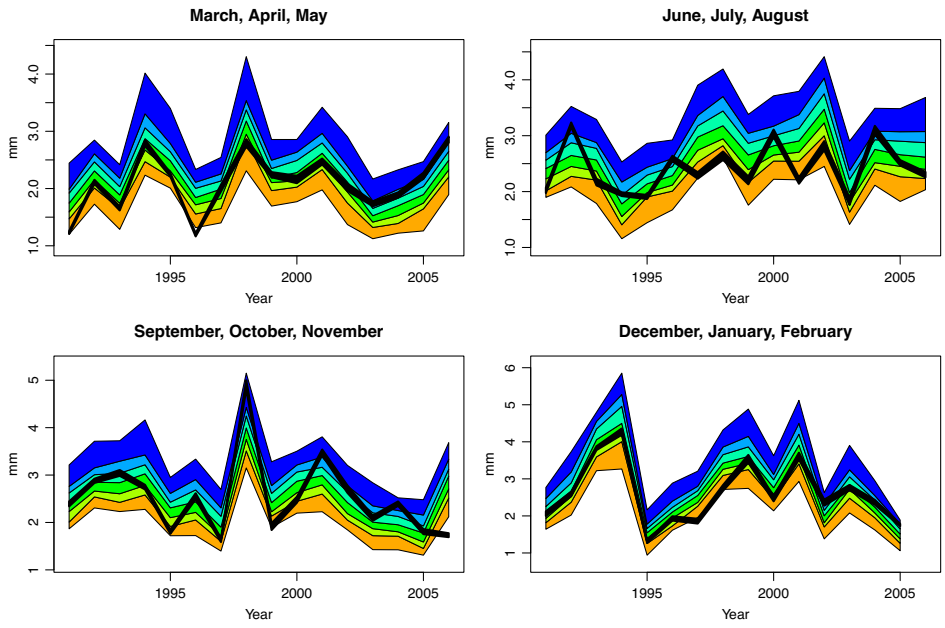

Figure 5.11: As Figure 5.9, but for fold 3 (1991-2006).

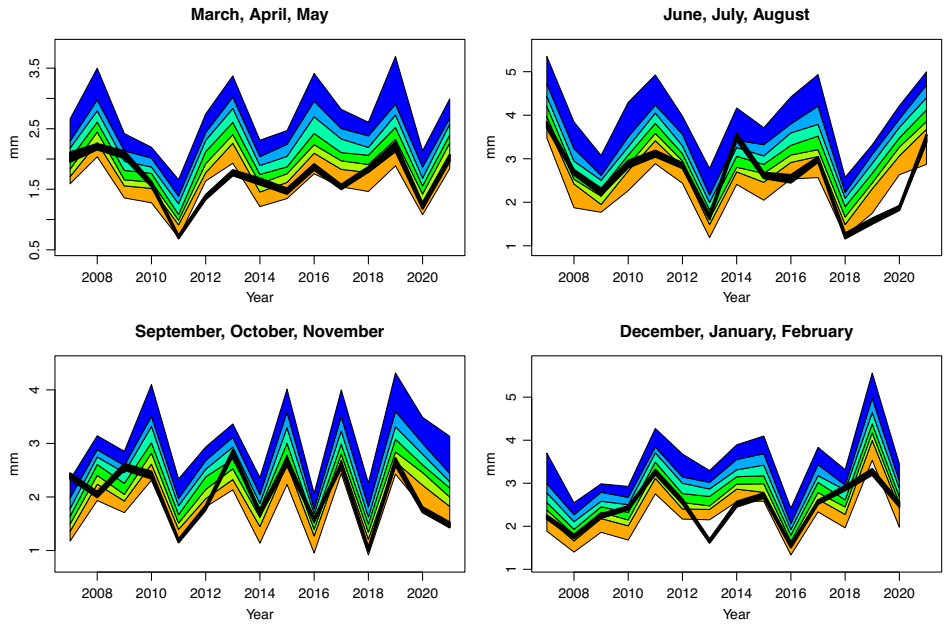

Figure 5.12: As Figure 5.9, but for fold 4 (2007-2021).

5.1.2.2 Temperature

Figures 5.13-5.16 show the distributions of seasonal temperature totals in each year, averaged over the whole region and for folds 1-4 respectively. Again, there is much variation between years in the simulations, which shows the effects of the large-scale covariates.

The simulated mean temperature in winter (December, January and February) is generally overestimated, especially in folds 1-3. Other statistics are reproduced reasonably well.

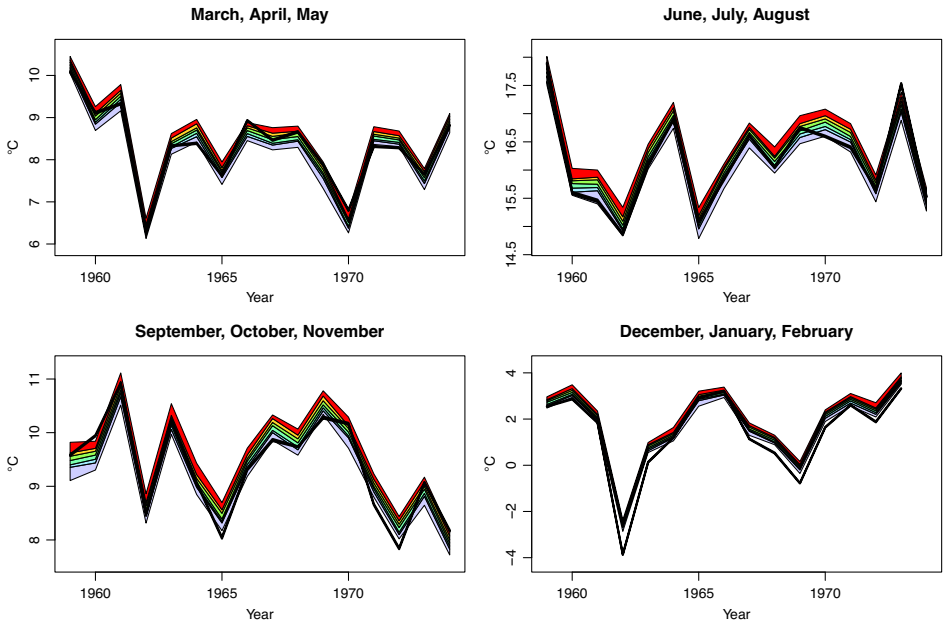

Figure 5.13: Simulated and observed annual mean for temperature for fold 1 (1959-1974) over the whole region separated by seasons for the statistical model. The coloured bands show the 1st, 5th, 10th, 25th, 75th, 90th, 95th and 99th percentiles of simulated distributions. The black bands show the 95% uncertainty interval of imputations.

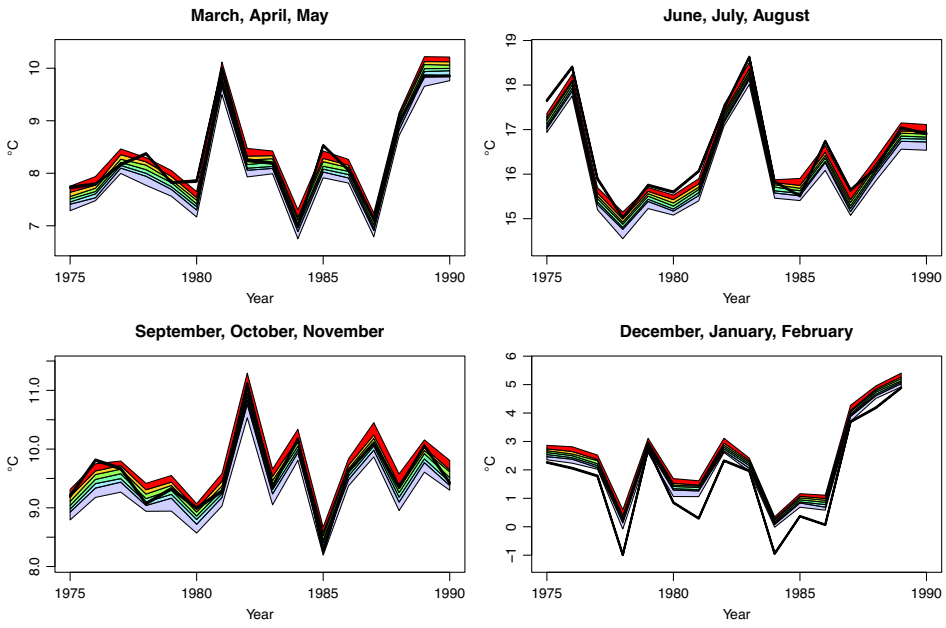

Figure 5.14: As Figure 5.13, but for fold 2 (1975-1990).

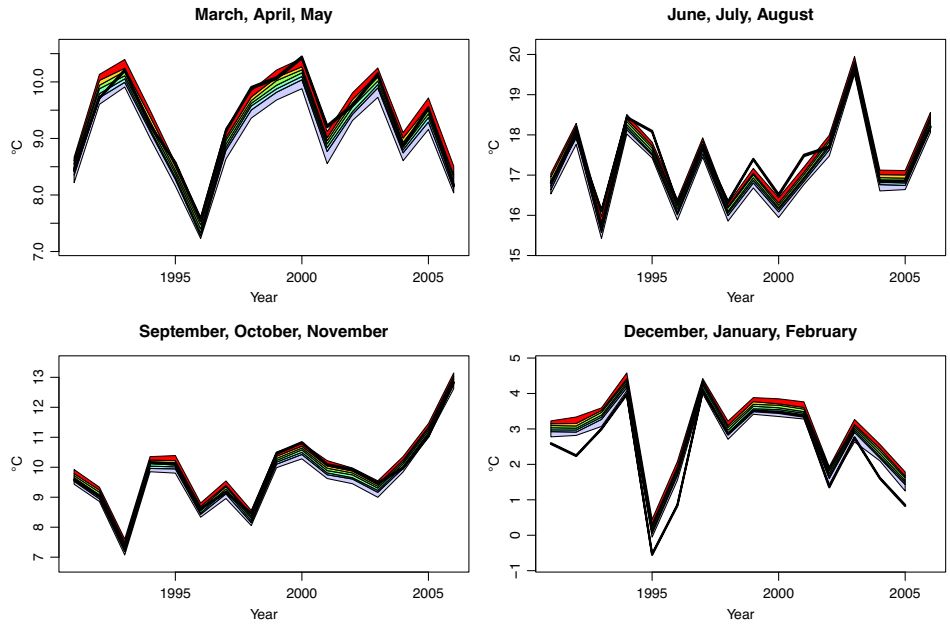

Figure 5.15: As Figure 5.13, but for fold 3 (1991-2006).

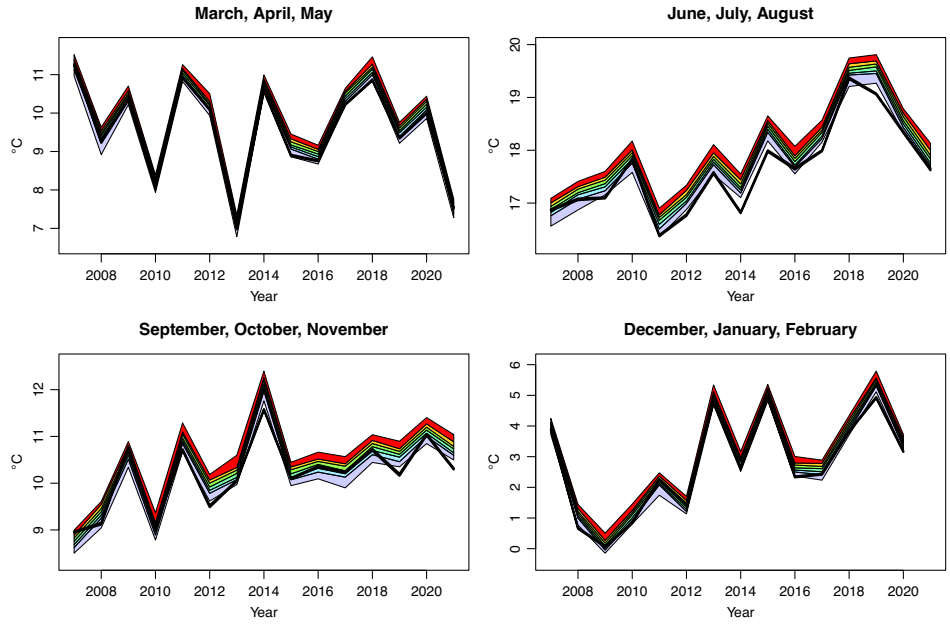

Figure 5.16: As Figure 5.13, but for fold 4 (2007-2021).

5.2 Simulations from Machine Learning Model

In practice terms, the training and simulation process of machine learning methods takes a very long time (about 4 days on a research computer), so it is impractical

to simulate 100 times. Alternatively, only 1 simulation is generated, and then compared with 1 observed weather series (which is a complete dataset obtained from the imputation using GLM as mentioned in Chapter 3, and the properties of the imputed data are unlikely to be sensitive to the GLM formulation because most of the information will come from the neighbouring stations and not from the model).

5.2.1 Monthly Summary Statistics

5.2.1.1 Precipitation

To assess the simulations from the machine learning model for precipitation, the monthly summary statistics over the whole region used include mean, standard deviation, extremes (maximum), autocorrelation at lag 1-3, the proportion of wet days, conditional mean and conditional standard deviation.

Figures 5.17-5.20 show the monthly precipitation summary statistics for folds 1-4 respectively. Some differences exist between the simulated and observed statistics. The simulated means, conditional means, standard deviations, and conditional standard deviations (1st, 2nd, 5th and 6th plots) are generally underestimated. In particular, simulated conditional means and conditional standard deviations are substantially underestimated. However, the seasonal patterns of these 4 simulated statistics and observed statistics are similar in general. Notably, the machine learning methods do not contain any explicit representation of seasonality (e.g. Fourier series in GLM). Therefore, if they are producing seasonal patterns, this must be learned from the seasonality in the large-scale predictors.

However, the seasonal patterns of simulated maximum (3rd plot), autocorrelations (4-6th plots), and proportion of wet days (7th plot) differ from the corresponding observed statistics.

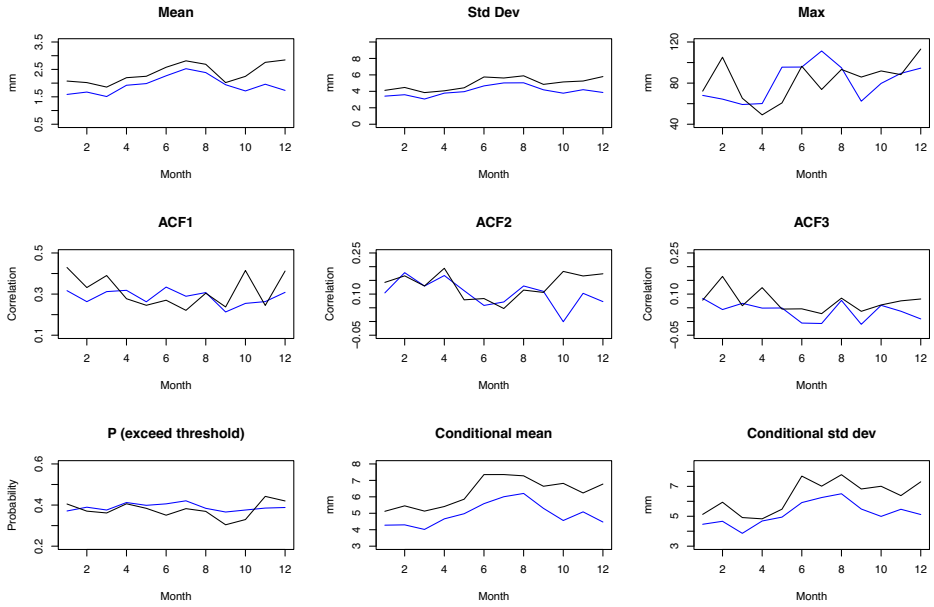

Figure 5.17: Simulated and observed monthly summary statistics for precipitation for fold 1 (1959-1974) over the whole region for the machine learning model. The blue line represents the simulated precipitation. The black line represents the observed precipitation. The plots show the mean, standard deviation, maximum, proportion of wet days, wet-day mean and wet-day standard deviation in order.

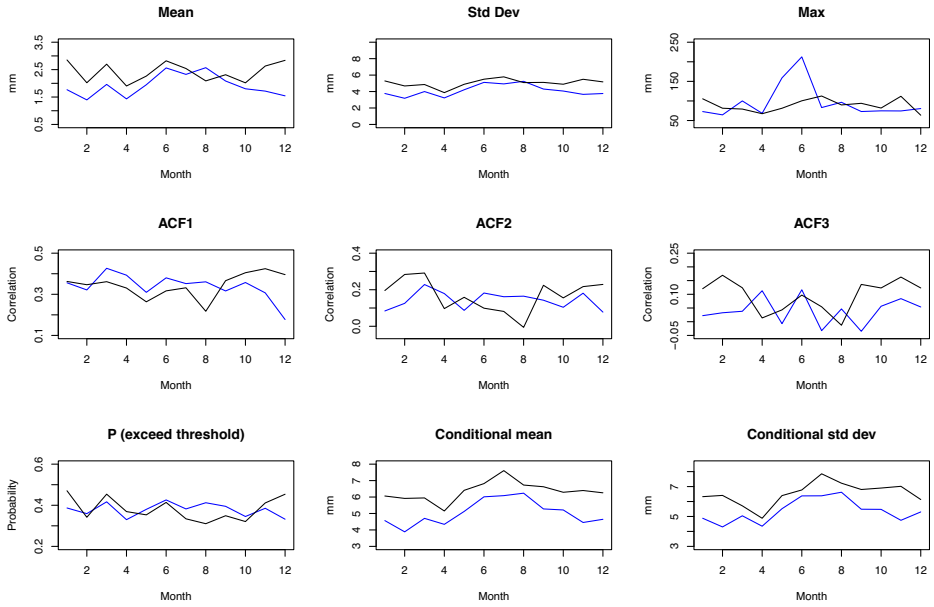

Figure 5.18: As Figure 5.17, but for fold 2 (1975-1990).

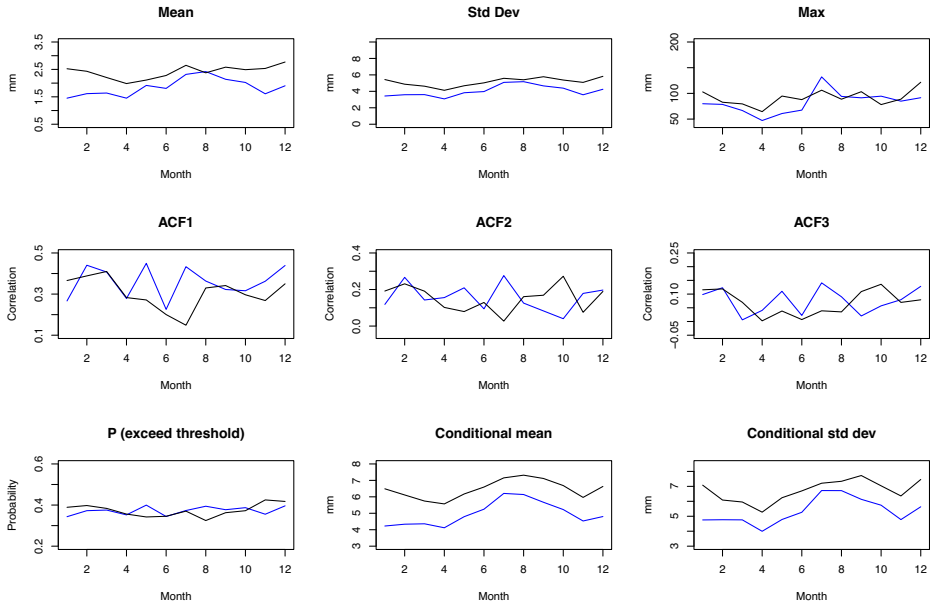

Figure 5.19: As Figure 5.17, but for fold 3 (1991-2006).

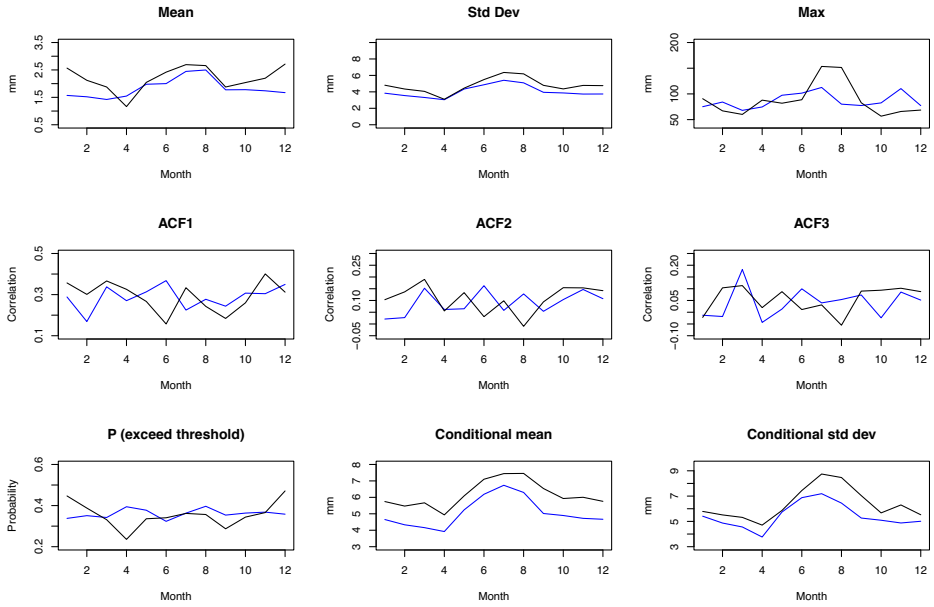

Figure 5.20: As Figure 5.17, but for fold 4 (2007-2021).

5.2.1.2 Temperature

For temperature, the monthly summary statistics used include mean, standard deviation, and extremes (maximum and minimum), autocorrelation at lag 1-3, correlation between temperature and precipitation. Figures 5.21-5.24 show the monthly

temperature summary statistics for folds 1-4 respectively.

In general, the seasonal patterns for all simulated and observed statistics are similar except for correlation with precipitation. Again, the seasonal patterns must be learned from the seasonality in the large-scale predictors. For means, the simulated distributions are generally similar to observed distributions. For standard deviations, the simulated values are slightly overestimated in March, September and October. The simulated maximums are generally overestimated, while the simulated minimums are slightly underestimated in general. For autocorrelations, the simulated autocorrelations at lag 1-3 are generally overestimated. From lag 1 to lag 3, the simulation values are getting closer to the observation values. For correlation with precipitation, the simulated values are about 0.05 for most of the months and generally remain unchanged. In contrast, the observed correlation has a clear decreasing trend in the first half of the year, and an increasing trend in the second half of the year. It shows that the dependence of both trees on the large-scale predictors is not enough to induce the correlation between the simulated variables.

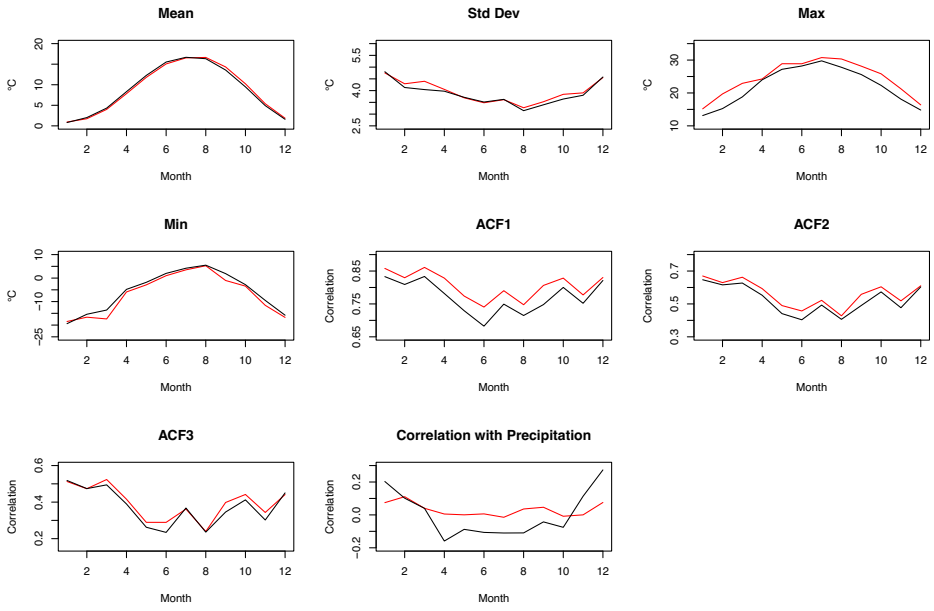

Figure 5.21: Simulated and observed monthly summary statistics for temperature for fold 1 (1959-1974) over the whole region for the machine learning model. The red line represents the simulated temperature. The black line represents the observed temperature. The plots show the mean, standard deviation, maximum and minimum in order.

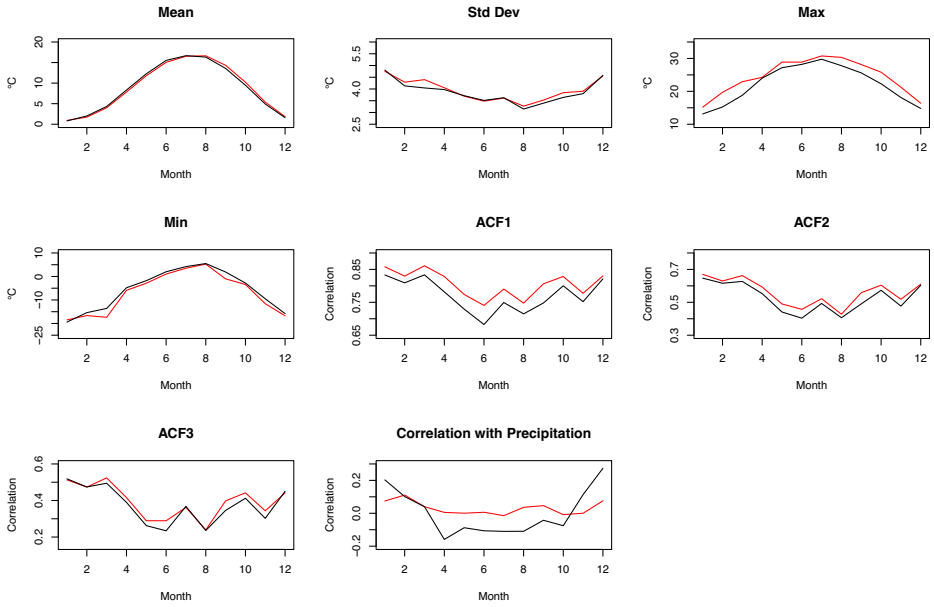

Figure 5.22: As Figure 5.21, but for fold 2 (1975-1990).

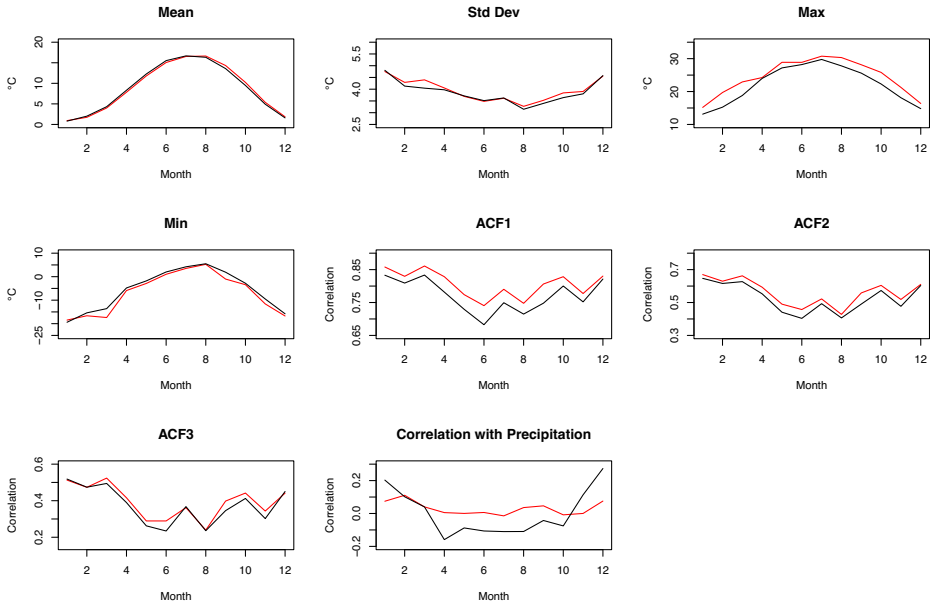

Figure 5.23: As Figure 5.21, but for fold 3 (1991-2006).

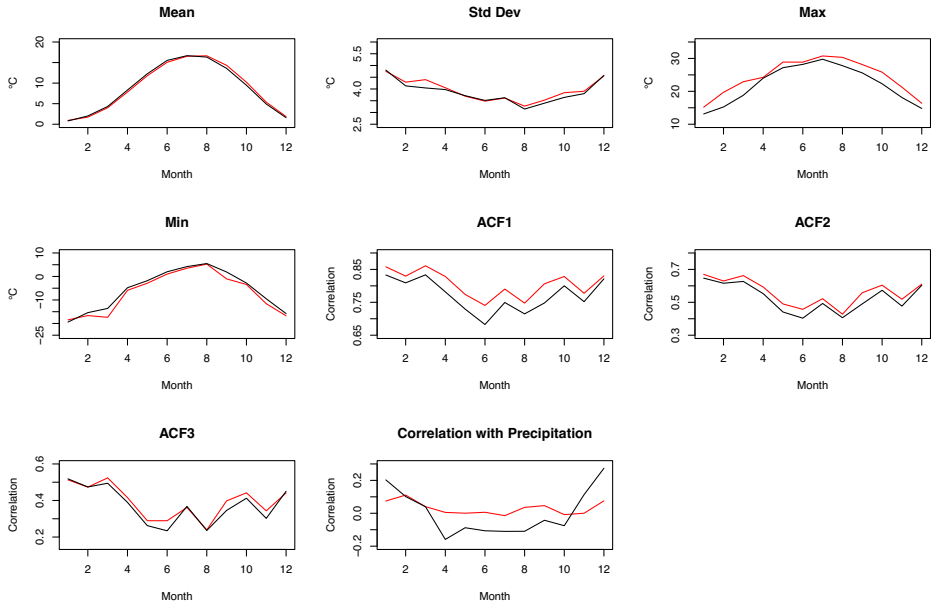

Figure 5.24: As Figure 5.21, but for fold 4 (2007-2021).

5.2.2 Seasonal Summary Statistics

5.2.2.1 Precipitation

Figures 5.25-5.28 show the distributions of seasonal precipitation totals each year, averaged over the whole region and for folds 1-4 respectively. Again, the simulated and observed distributions look different. The simulated means are generally underestimated in most years in each season.

The seasonal patterns of simulations and observations have some differences as well. The variation of simulations is generally small, while observations have more significant variations over the years. Notably, large-scale predictors are not incorporated in Bayesian network (see Section 3.2.1). However, they are incorporated in the random forest for precipitation amounts, so some interannual variation in the simulations is expected. Comparing these plots with those from the GLM, for example, in the MAM plot for fold 1, there are high simulated values in 1961, 1965 and 1969; and a low value in 1962. These do not coincide with the peaks and troughs in the GLM-simulated distributions from Figure 5.9 (which follow the observations fairly well). Therefore, the random forest alone cannot capture the interannual variation in precipitation.

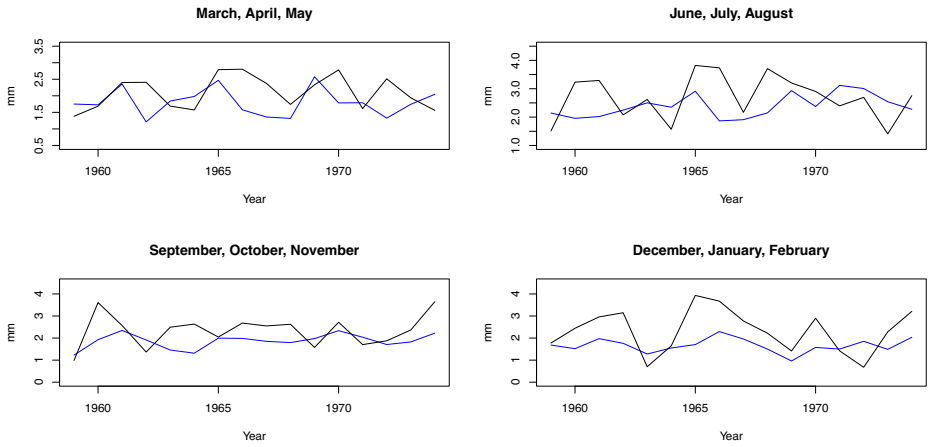

Figure 5.25: Simulated and observed annual mean for precipitation for fold 1 (1959-1974) over the whole region separated by seasons for the machine learning model. The blue line represents the simulated precipitation. The black line represents the observed precipitation.

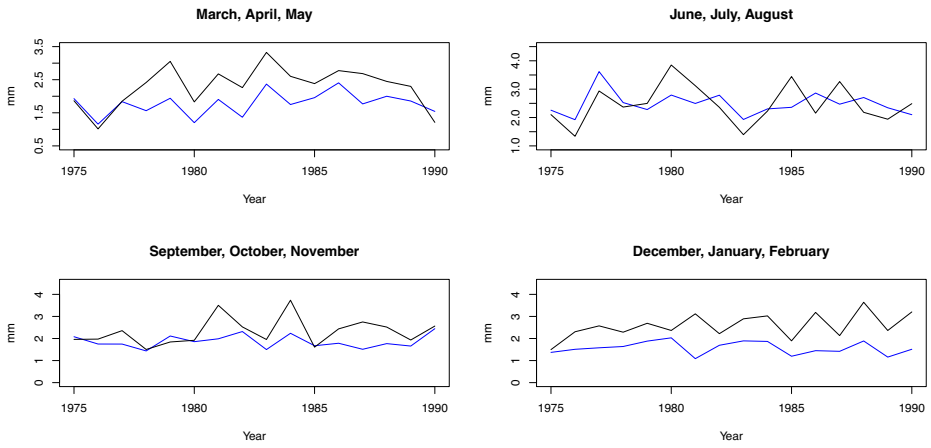

Figure 5.26: As Figure 5.25, but for fold 2 (1975-1990).

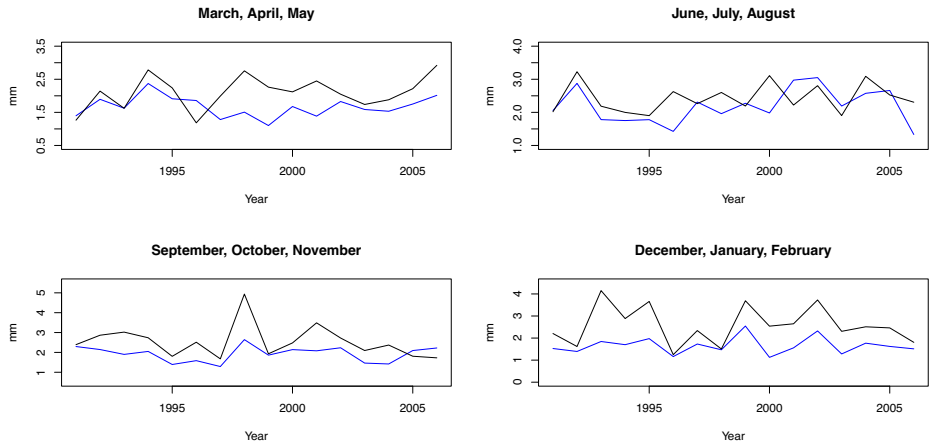

Figure 5.27: As Figure 5.25, but for fold 3 (1991-2006).

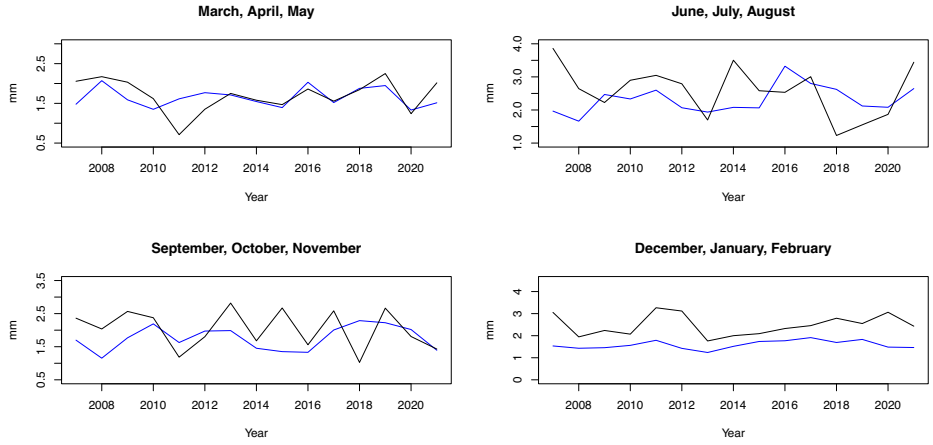

Figure 5.28: As Figure 5.25, but for fold 4 (2007-2021).

5.2.2.2 Temperature

Figures 5.29-5.32 are the distributions of the annual temperature time series for folds 1-4 respectively, over the whole region separated by seasons. It can be seen that the simulations and observations have the same seasonal patterns. Therefore, it illustrates that the random forest is picking up relevant information from the large-scale predictors (whereas for precipitation it was not). However, the simulations for the first half of the year are generally underestimated, while the simulations for the second half of the year are generally overestimated. The simulations in summer and winter are reproduced better than the simulations in spring and autumn, only slightly underestimated and overestimated respectively.

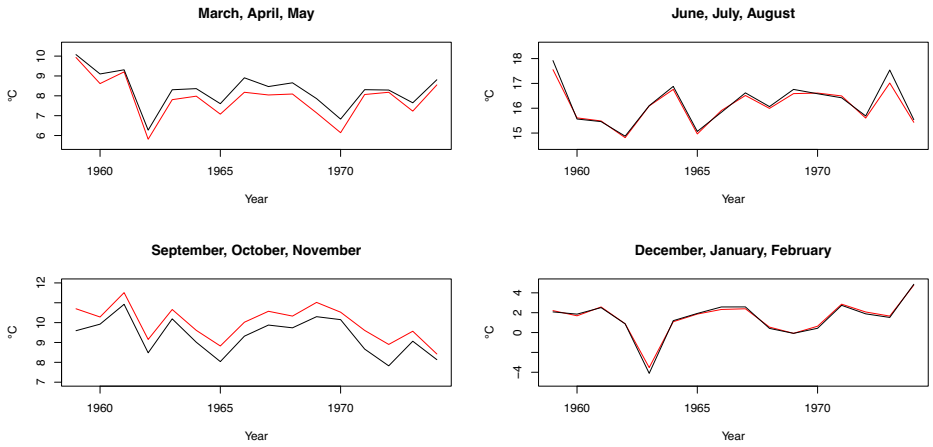

Figure 5.29: Simulated and observed annual mean for temperature for fold 1 (1959-1974), over the whole region separated by seasons for the machine learning model. The red line represents the simulated temperature. The black line represents the observed temperature.

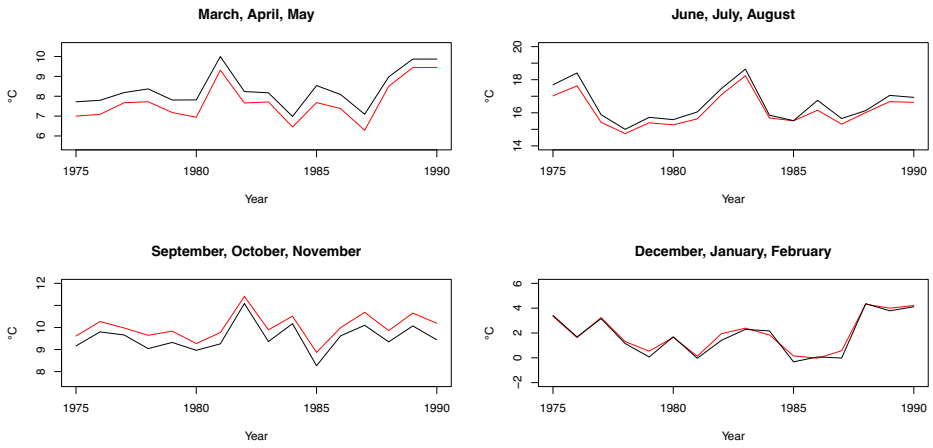

Figure 5.30: As Figure 5.29, but for fold 2 (1975-1990).

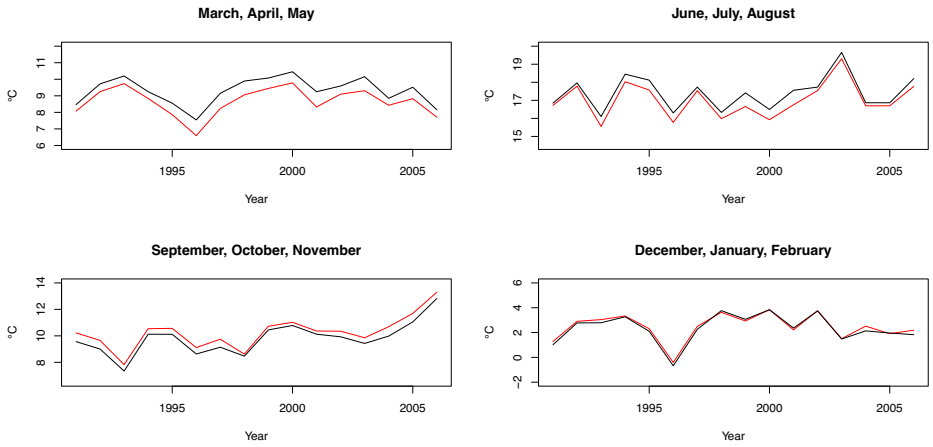

Figure 5.31: As Figure 5.29, but for fold 3 (1991-2006).

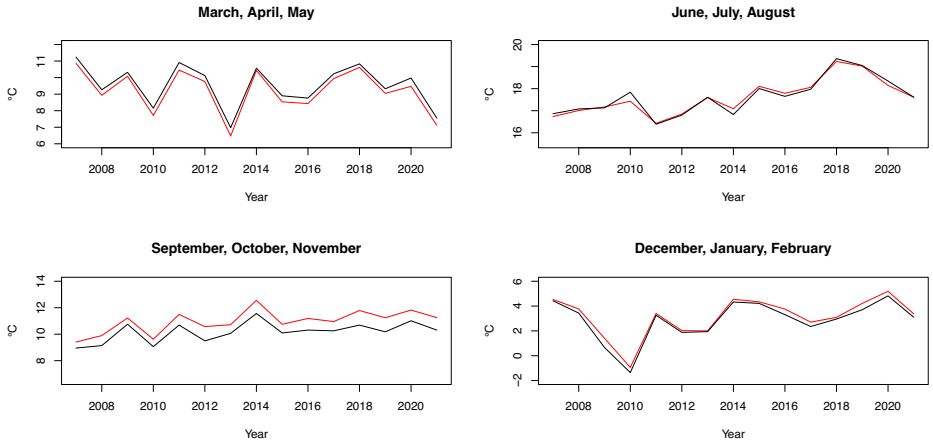

Figure 5.32: As Figure 5.29, but for fold 4 (2007-2021).

5.3 Comparison between Two Methods

Based on the simulation results from the two methods, GLM performs better in general, especially for predicting precipitation. Although some uncertainties exist because only 1 simulation is generated from the machine learning model, the width of the GLM-simulated distributions can be taken as an informal indication of the expected range of simulations from the machine learning models if more simulations can be generated. Also, GLM simulates the extremes of precipitation distributions that are more similar to the observed distributions. It suggests that GLM may be more useful for investigating climate change on flooding in particular.

Apart from the model performance, other differences between the two downscaling methods include the following:

- Computational efficiency: To generate 1 simulation for all folds and for both precipitation and temperature, GLM takes about 1.5 minutes on a normal computer. However, the machine learning method takes about 4 days in total on a research computer. All of the results that have been produced in this study using the software packages provided by the developers of the respective methodologies. Hence these conclusions are expected to give a reasonable indication of what other users of these packages might experience. The huge difference in simulation time indicates that the machine learning method is not practical in use.

- Requirement for a complete dataset: GLM does not need a complete dataset to train the model, but the machine learning method needs a complete dataset. If there are missing data in the observed dataset, they need to be imputed to be used for the machine learning method. This may cause the problems such as the possibility of the filled values having different statistical features from the rest of the data, and thus affect the predictions of the machine learning model.

- Model Building: GLM is built manually and many different models should be compared. This may affect the model's performance if the best-performance model is not chosen and the model-building process is also time-consuming (about 2 days). However, human intervention is unnecessary for machine learning methods as they can learn from data automatically, and the learning process only takes a few seconds.

- Relationship between weather variables: GLM builds a bivariate model by adding one weather variable as a covariate to the model of the other variable, so the model can capture the relationship between variables. However, random forest builds trees for each variable separately and the dependence of both trees on the large-scale predictors is not enough to induce the correla-

tion between the simulated variables. The seasonal variations for correlation between temperature and precipitation are not well captured (see Subsection 5.2.1).

5.4 Summary

In this chapter, the simulation results of the two methods are assessed. Both monthly and seasonal summary statistics are used to compare the simulated and observed values. For precipitation, simulations from GLM overestimate the means and conditional means, while other statistics of simulations generally match the observations well. For machine learning methods, some differences exist between the simulated and observed statistics. For example, the simulations generally underestimate means, standard deviations, conditional means, and conditional standard deviations. Also, random forest learns from the seasonality in the large-scale predictors for some statistics, but it alone cannot capture the interannual variation.

For temperature, both models perform reasonably well. Exceptions include: For GLM, the simulated standard deviations were underestimated in some months and simulated minimums were overestimated in some months. For machine learning model, the simulated maximums are overestimated, and simulated minimums are underestimated. The simulations for the first half of the year are generally underestimated, while the simulations for the second half of the year are generally overestimated, but the overall simulated means show a good agreement with observed means. The simulated correlation with precipitation generally remains unchanged and does not capture the seasonal properties.

Then the comparisons are carried out for the two methods. The differences include: predicting performance, computational efficiency for simulation, the requirement for a complete dataset, model building, and the relationship between weather variables.

Chapter 6

Conclusion

The main objectives of the study are to compare the performance of the statistical method including GLM and machine learning methods including Bayesian network and random forest, and to identify the strengths and weaknesses of each method. To achieve these, the following steps are carried out:

1. First, collect relevant weather variables (daily precipitation and temperature) at all sites, topographical data, geographical data for stations, and large-scale atmospheric data.

2. The second step is to build models. For GLM, precipitation and temperature are fitted to different models to construct univariate models. Temporal, spatial dependencies and autocorrelation are represented in the models. The performance of the models is assessed by formal model comparisons (e.g. likelihood ratio test) and diagnostic plots (e.g. residual plots). Decide the primary variable and then construct a bivariate model. Add large-scale predictors to link the local-scale and large-scale weather. For machine learning models: Bayesian network takes geographical data for stations as predictands, precipitation and temperature data as predictors; random forest takes large-scale atmospheric data as predictands, and precipitation and temperature data as predictors. The best configurations are set to train the models.

3. Third, generate the simulations of daily weather series and compare simulations with the observations by several monthly and annual summary statistics.

Based on the simulation results from each method, GLM performs better in general. Most of the simulated values match the observed value. Exceptions include the simulated mean and conditional mean of precipitation being overestimated. However, for machine learning method, differences exist between the simulated and observed statistics, especially for the precipitation model. The performance of the machine learning model for temperature is reasonably well. Although some statistics are slightly overestimated and underestimated, the seasonal patterns and interannual variations of simulations and observations are roughly the same.

Other aspects of comparisons are carried out. Some advantages of GLM include: a higher computational efficiency for simulation; it does not need a complete dataset; the ability to capture the relationship between weather variables. The advantages of the machine learning method include the auto-selection of the best model, and thus no human intervention of the model and time-saving.

Some limitations of the study include:

1. In other applications of the machine learning methods, the large-scale predictors are often incorporated at a finer spatial resolution so that the predictors used for each station correspond more closely to the local conditions at that station (e.g. Legasa et al. (2022) used predictors as 0.25-degree spatial resolution). In this work, only a single time series of atmospheric predictors were used to represent average conditions over the entire study area so that some detailed information would be lost. Therefore, this issue may be considered as a possible explanation for why the results from the machine learning methods in this work are not as good as those reported in the literature.

2. One possible reason for the relatively poorer performance of the machine learning model for precipitation is that random forests are underfitted. To cut the computational costs, the number of trees is limited and the number of leaf sizes is insufficient (see Chapter 3). To solve this problem, the complexity of the model needs to be increased. Therefore, increase the number of trees and decrease the number of leaf sizes to see if there will be better performance for the model.

3. The particular type of Bayesian network used is Markov Bayesian network. It only allows the temporal relationships between a node and its past. Legasa et al. (2022) also introduced other types, such as augmented Bayesian network, whose directions of the arcs are less restrictive, and thus it may better capture the temporal structure to improve the model. Therefore, use an augmented Bayesian network to see if there is any improvement in the model.

4. The large-scale predictors are currently not added to train Bayesian network. This approach is consistent with that from Legasa and Gutiérrez (2020) as they did not use large-scale predictors either. However, BNWeatherGen package allows the inclusion of large-scale predictors in principle, but large-scale predictors need a grid format to be used as input in the package. Due to the time limit, the large-scale predictors are not added to the model. Therefore, add the large-scale predictors to Bayesian network may improve the model performance.

Appendices

Appendix A

Model Structure

The following tables are for models fitted to the complete dataset (and the results reported in Chapter 5 are obtained from models that have been refitted separately to the four training sets and then validated independently on the corresponding test sets).

A.1 Precipitation Occurrence Model

Main effects:

	Coefficient	Std Err	Z-stat	Pr(|Z|>z)
Constant	11.6432	0.9192	12.6670	< 2.2e-16
1 Legendre polynomial 1 for	1.8393	0.3803	4.8365	1.321e-06
2 Legendre polynomial 2 for	-0.0969	0.0093	-10.4757	< 2.2e-16
3 Legendre polynomial 1 for	-2.2757	0.4464	-5.0977	3.439e-07
4 Legendre polynomial 2 for	-0.1279	0.0099	-12.8889	< 2.2e-16
5 100km^2 mean altitude	0.1743	0.0040	43.4451	< 2.2e-16
6 1000km^2 E-W slope (100m /	1.9328	0.0611	31.6180	< 2.2e-16
7 1000km^2 N-S slope (100m /	1.2543	0.0598	20.9887	< 2.2e-16
8 Daily annual cycle, cosine	-2.8521	1.2427	-2.2951	0.021726
9 Daily annual cycle, sine c	-7.5342	1.1547	-6.5247	6.812e-11
10 First harmonic of daily an	0.3835	0.0393	9.7675	< 2.2e-16
11 First harmonic of daily an	0.2018	0.0406	4.9656	6.849e-07
12 Distance-weighted mean of	0.4291	0.0580	7.4032	1.330e-13
13 Distance-weighted mean of	-0.1421	0.0295	-4.8199	1.436e-06
14 Mean of Temperature[t]	-0.0326	0.0075	-4.3437	1.401e-05
15 GeoPot_700	-5.8034	0.3002	-19.3346	< 2.2e-16
16 Temp_700	-0.2190	0.0079	-27.6267	< 2.2e-16
17 SpecHum_700	1.6706	0.0319	52.3374	< 2.2e-16
18 UWind_850	0.1127	0.0026	42.7525	< 2.2e-16
19 VWind_850	-0.0109	0.0035	-3.1168	0.001829

Two-way interactions:

	Coefficient	Std Err	Z-stat	Pr(|Z|>z)
Legendre polynomial 1 with Legendre polynomial 1	0.1004	0.0164	6.1136	9.744e-10
First harmonic of daily with Distance-weighted mean	-0.2678	0.0403	-6.6406	3.124e-11
First harmonic of daily with Distance-weighted mean	0.0437	0.0394	1.1085	0.2676654
Legendre polynomial 1 with Daily annual cycle, cos	-0.0422	0.0212	-1.9905	0.0465332
Legendre polynomial 1 with Daily annual cycle, sin	-0.0409	0.0167	-2.4436	0.0145403
Legendre polynomial 1 with Distance-weighted mean	0.0964	0.0263	3.6626	0.0002496
Legendre polynomial 1 with Distance-weighted mean	-0.1466	0.0287	-5.1077	3.260e-07
100km^2 mean altitude with Mean of Temperature[t]	-0.0037	0.0003	-11.5277	< 2.2e-16
Daily annual cycle, cos with Mean of Temperature[t]	0.0462	0.0095	4.8511	1.228e-06
Daily annual cycle, sin with Mean of Temperature[t]	0.0062	0.0100	0.6168	0.5373789
Distance-weighted mean with Mean of Temperature[t]	-0.0192	0.0047	-4.0602	4.903e-05
Legendre polynomial 1 with GeoPot_700	-0.7981	0.1331	-5.9947	2.039e-09
Legendre polynomial 1 with GeoPot_700	0.9041	0.1536	5.8858	3.962e-09
Daily annual cycle, cos with GeoPot_700	0.8016	0.4059	1.9749	0.0482793
Daily annual cycle, sin with GeoPot_700	2.0849	0.3733	5.5850	2.337e-08
Daily annual cycle, cos with Temp_700	0.0768	0.0103	7.4735	7.808e-14
Daily annual cycle, sin with Temp_700	-0.0446	0.0100	-4.4424	8.897e-06
Legendre polynomial 1 with SpecHum_700	0.1178	0.0117	10.0610	< 2.2e-16
Legendre polynomial 1 with SpecHum_700	-0.1138	0.0165	-6.9177	4.591e-12
Daily annual cycle, cos with SpecHum_700	0.0030	0.0396	0.0766	0.9389097
Daily annual cycle, sin with SpecHum_700	0.4298	0.0417	10.2998	< 2.2e-16
Legendre polynomial 1 with UWind_850	0.0283	0.0016	17.4068	< 2.2e-16
Legendre polynomial 1 with UWind_850	0.0518	0.0020	25.4124	< 2.2e-16

```
      Daily annual cycle, cos      0.0455   0.0037  12.1654 < 2.2e-16
               with UWind_850
      Daily annual cycle, sin      0.0112   0.0033   3.4446 0.0005719
               with UWind_850
      Legendre polynomial  1      -0.0525   0.0023 -23.3229 < 2.2e-16
               with VWind_850
      Legendre polynomial  1      -0.0079   0.0023  -3.3662 0.0007621
               with VWind_850
      Daily annual cycle, cos     -0.0704   0.0048 -14.8039 < 2.2e-16
               with VWind_850
      Daily annual cycle, sin      0.0054   0.0046   1.1765 0.2394070
               with VWind_850

Parameters in nonlinear transformations:
----------------------------------------
Legendre polynomial  1 for Longitude:
     Lower limit for polynomial rep:      5.6000 (prespecified)
     Upper limit for polynomial rep:      8.9000 (prespecified)
Legendre polynomial  1 for Latitude:
     Lower limit for polynomial rep:     50.1000 (prespecified)
     Upper limit for polynomial rep:     52.0000 (prespecified)
Distance-weighted mean of I(Precipitation[t-1]>0):
             Exponential decay rate:      3.9251 (Std Err:    0.4035)

Global quantities:
------------------
     'Soft' threshold for +ve values:      0.9500

No dispersion parameters defined

Spatial dependence structure:
-----------------------------
Structure used: Exponential correlation function: exp[-phi*d] fo
     Correlation decay rate, phi:      0.3438
```

A.2 Precipitation Amount Model

Main effects:

	Coefficient	Std Err	T-stat	Pr(|T|>t)
Constant	0.2740	0.0421	6.5120	7.426e-11
1 Legendre polynomial 1 for	-0.2651	0.0220	-12.0698	< 2.2e-16
2 Legendre polynomial 2 for	-0.0395	0.0068	-5.8327	5.459e-09
3 Legendre polynomial 1 for	0.0696	0.0270	2.5840	0.0097664
4 Legendre polynomial 2 for	-0.0913	0.0077	-11.9182	< 2.2e-16
5 100km^2 mean altitude	0.1265	0.0033	37.8718	< 2.2e-16
6 1000km^2 E-W slope (100m /	1.7493	0.1004	17.4180	< 2.2e-16
7 1000km^2 N-S slope (100m /	1.5503	0.0998	15.5326	< 2.2e-16
8 Daily annual cycle, cosine	-0.5660	0.0560	-10.1009	< 2.2e-16
9 Daily annual cycle, sine c	-0.1738	0.0493	-3.5230	0.0004267
10 First harmonic of daily an	-0.0782	0.0169	-4.6221	3.801e-06
11 First harmonic of daily an	-0.0270	0.0187	-1.4425	0.1491567
12 Ln(1+Mean of Precipitation	0.0511	0.0094	5.4425	5.256e-08
13 Mean of Temperature[t]	-0.0532	0.0040	-13.2156	< 2.2e-16
14 GeoPot_1000	-2.1499	0.1755	-12.2504	< 2.2e-16
15 SpecHum_700	0.5902	0.0151	39.0813	< 2.2e-16
16 UWind_850	0.0233	0.0017	14.1260	< 2.2e-16
17 VWind_850	-0.0116	0.0021	-5.5600	2.699e-08

Two-way interactions:

	Coefficient	Std Err	T-stat	Pr(|T|>t)
Legendre polynomial 1 with Legendre polynomial 1	-0.0388	0.0124	-3.1305	0.0017451
Daily annual cycle, cos with Ln(1+Mean of Precipitat	0.0508	0.0128	3.9651	7.338e-05
Daily annual cycle, sin with Ln(1+Mean of Precipitat	0.0047	0.0136	0.3500	0.7263421
100km^2 mean altitude with Mean of Temperature[t]	-0.0044	0.0003	-15.2212	< 2.2e-16
1000km^2 E-W slope (100 with Mean of Temperature[t]	-0.0189	0.0098	-1.9278	0.0538798
1000km^2 N-S slope (100 with Mean of Temperature[t]	-0.0537	0.0093	-5.7921	6.957e-09
Daily annual cycle, cos with Mean of Temperature[t]	-0.0188	0.0048	-3.8867	0.0001016
Daily annual cycle, sin with Mean of Temperature[t]	-0.0005	0.0053	-0.0955	0.9239178
Daily annual cycle, cos with GeoPot_1000	0.0317	0.2236	0.1417	0.8872780
Daily annual cycle, sin with GeoPot_1000	0.6531	0.2228	2.9311	0.0033783
Legendre polynomial 1 with SpecHum_700	0.0517	0.0071	7.2705	3.589e-13
Legendre polynomial 1 with SpecHum_700	-0.0189	0.0084	-2.2582	0.0239325
Daily annual cycle, cos with SpecHum_700	0.0948	0.0202	4.6921	2.705e-06
Daily annual cycle, sin with SpecHum_700	0.0454	0.0208	2.1796	0.0292869
Legendre polynomial 1 with UWind_850	0.0139	0.0011	12.7962	< 2.2e-16
Legendre polynomial 1 with UWind_850	0.0167	0.0014	11.5423	< 2.2e-16
Daily annual cycle, cos with UWind_850	0.0440	0.0023	19.2695	< 2.2e-16
Daily annual cycle, sin with UWind_850	0.0057	0.0022	2.6041	0.0092119
Legendre polynomial 1 with VWind_850	-0.0148	0.0013	-11.5220	< 2.2e-16
Legendre polynomial 1 with VWind_850	-0.0026	0.0014	-1.8905	0.0586954
Daily annual cycle, cos with VWind_850	-0.0023	0.0027	-0.8412	0.4002368
Daily annual cycle, sin with VWind_850	0.0086	0.0029	3.0043	0.0026623

Parameters in nonlinear transformations:

Legendre polynomial 1 for Longitude:

```
     Lower limit for polynomial rep:      5.6000 (prespecified)
     Upper limit for polynomial rep:      8.9000 (prespecified)
Legendre polynomial  1 for Latitude:
     Lower limit for polynomial rep:     50.1000 (prespecified)
     Upper limit for polynomial rep:     52.0000 (prespecified)

Global quantities:
------------------
     'Soft' threshold for +ve values:      0.9500

Dispersion parameter:      1.2418

Spatial dependence structure:
-----------------------------
Structure used: Pow. exp. corr. fn with nugget: L*exp[-phi*(d^k)
     Correlation decay rate, phi:      0.8975
     Power of distance, d:      0.4026
     Limiting correlation at zero distance, L:      1.0000
```

A.3 Temperature Model

```
Main effects:
-------------
                              Coefficient  Std Err    Z-stat Pr(|Z|>z)
                     Constant    -22.3117   0.4075  -54.7473 < 2.2e-16
1             Altitude (100m)     -0.6686   0.0056 -119.3145 < 2.2e-16
2  Legendre polynomial  1 for      6.7102   0.2879   23.3039 < 2.2e-16
3  Legendre polynomial  2 for     -0.1035   0.0042  -24.4858 < 2.2e-16
4  Legendre polynomial  1 for     -1.1388   0.3071   -3.7085 0.0002085
5  Legendre polynomial  2 for     -0.1589   0.0047  -34.1713 < 2.2e-16
6  Daily annual cycle, cosine     12.2336   0.5737   21.3228 < 2.2e-16
7  Daily annual cycle, sine c     -5.5932   0.4770  -11.7263 < 2.2e-16
8  First harmonic of daily an      0.0747   0.0159    4.7130 2.441e-06
9  First harmonic of daily an      0.0366   0.0160    2.2924 0.0218853
10 Distance-weighted mean of       0.1830   0.0049   37.2773 < 2.2e-16
11 Distance-weighted mean of      -0.0543   0.0037  -14.4954 < 2.2e-16
12 Distance-weighted mean of       0.0186   0.0027    6.9090 4.881e-12
13                  Temp_1000      0.5600   0.0067   83.2094 < 2.2e-16
14                 GeoPot_500      4.0792   0.0775   52.6346 < 2.2e-16
15               SpecHum_1000      0.5926   0.0135   44.0116 < 2.2e-16
16                 UWind_1000     -0.0287   0.0023  -12.3275 < 2.2e-16
17                 VWind_1000      0.0479   0.0030   15.8288 < 2.2e-16

Two-way interactions:
---------------------
                              Coefficient  Std Err   Z-stat Pr(|Z|>z)
      Legendre polynomial  1       0.2168   0.0092  23.4506 < 2.2e-16
 with Legendre polynomial  1
      Daily annual cycle, cos      0.0226   0.0057   3.9884 6.652e-05
 with Distance-weighted mean
      Daily annual cycle, sin     -0.0042   0.0052  -0.8059  0.420313
 with Distance-weighted mean
             Altitude (100m)     -0.0340   0.0006 -61.0070 < 2.2e-16
 with Distance-weighted mean
      Legendre polynomial  1     -0.0621   0.0151  -4.1066 4.015e-05
 with Daily annual cycle, cos
      Legendre polynomial  1       0.1326   0.0105  12.6129 < 2.2e-16
 with Daily annual cycle, sin
      Legendre polynomial  1       0.0032   0.0157   0.2025  0.839544
 with Daily annual cycle, cos
      Legendre polynomial  1     -0.0529   0.0100  -5.2819 1.278e-07
 with Daily annual cycle, sin
      Legendre polynomial  1       0.1217   0.0029  41.5329 < 2.2e-16
 with Distance-weighted mean
      Legendre polynomial  1     -0.0859   0.0030 -29.0011 < 2.2e-16
 with Distance-weighted mean
             Altitude (100m)      0.0356   0.0006  54.9401 < 2.2e-16
              with Temp_1000
      Legendre polynomial  1     -0.1784   0.0042 -42.4436 < 2.2e-16
              with Temp_1000
      Legendre polynomial  1      0.0585   0.0042  13.8286 < 2.2e-16
              with Temp_1000
      Daily annual cycle, cos     -0.1725   0.0083 -20.7508 < 2.2e-16
              with Temp_1000
      Daily annual cycle, sin      0.0397   0.0071   5.5877 2.301e-08
              with Temp_1000
      Legendre polynomial  1     -1.4681   0.0542 -27.0679 < 2.2e-16
             with GeoPot_500
      Legendre polynomial  1      0.1890   0.0578   3.2686  0.001081
             with GeoPot_500
      Daily annual cycle, cos     -2.8019   0.1095 -25.5981 < 2.2e-16
             with GeoPot_500
      Daily annual cycle, sin      1.1620   0.0890  13.0536 < 2.2e-16
             with GeoPot_500
      Legendre polynomial  1      0.3095   0.0078  39.7540 < 2.2e-16
           with SpecHum_1000
      Legendre polynomial  1     -0.0003   0.0083  -0.0325  0.974109
           with SpecHum_1000
      Daily annual cycle, cos      0.7126   0.0181  39.3962 < 2.2e-16
           with SpecHum_1000
      Daily annual cycle, sin     -0.0703   0.0133  -5.2702 1.363e-07
           with SpecHum_1000
      Legendre polynomial  1      0.0165   0.0017   9.5448 < 2.2e-16
             with UWind_1000
```

```
      Legendre polynomial  1      -0.0151   0.0019  -7.8924 2.965e-15
              with UWind_1000
      Daily annual cycle, cos       0.0735   0.0034  21.5314 < 2.2e-16
              with UWind_1000
      Daily annual cycle, sin      -0.0361   0.0029 -12.3368 < 2.2e-16
              with UWind_1000
      Legendre polynomial  1      -0.0390   0.0023 -16.8642 < 2.2e-16
              with VWind_1000
      Legendre polynomial  1       0.0592   0.0024  24.4833 < 2.2e-16
              with VWind_1000
      Daily annual cycle, cos       0.0061   0.0041   1.4779  0.139447
              with VWind_1000
      Daily annual cycle, sin       0.0187   0.0039   4.8068 1.534e-06
              with VWind_1000

Three-way interactions:
-----------------------
                              Coefficient  Std Err   Z-stat Pr(|Z|>z)
      Daily annual cycle, cos     -0.0020   0.0003  -6.1752 6.609e-10
 with Distance-weighted mean
                and Temp_1000
      Daily annual cycle, sin     -0.0003   0.0003  -1.1572    0.2472
 with Distance-weighted mean
                and Temp_1000

Parameters in nonlinear transformations:
----------------------------------------
Legendre polynomial  1 for Longitude:
     Lower limit for polynomial rep:      5.6000 (prespecified)
     Upper limit for polynomial rep:      8.9000 (prespecified)
Legendre polynomial  1 for Latitude:
     Lower limit for polynomial rep:     50.1000 (prespecified)
     Upper limit for polynomial rep:     52.0000 (prespecified)
Distance-weighted mean of Temperature[t-1]:
             Exponential decay rate:      7.0650 (prespecified)

<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<<
Dispersion structure:
---------------------
TEMPERATURE DISPERSION MODEL, CONDITIONED ON LARGE-SCALE DAILY TEMPERATURE
==========================================================================
Main effects:
-------------
                              Coefficient  Std Err   T-stat Pr(|T|>t)
                     Constant      0.4442   0.0079  56.5047 < 2.2e-16
1             Altitude (100m)     -0.0125   0.0022  -5.5934 2.227e-08
2 Legendre polynomial  1 for       0.0699   0.0066  10.5613 < 2.2e-16
3 Legendre polynomial  2 for       0.0976   0.0063  15.5928 < 2.2e-16
4 Legendre polynomial  1 for      -0.0578   0.0075  -7.7233 1.134e-14
5 Legendre polynomial  2 for       0.0414   0.0067   6.1889 6.061e-10
6 Daily annual cycle, cosine       0.0579   0.0096   6.0569 1.388e-09
7 Daily annual cycle, sine co     -0.0083   0.0087  -0.9591  0.337499
8 First harmonic of daily ann     -0.0276   0.0087  -3.1618  0.001568
9 First harmonic of daily ann     -0.0846   0.0092  -9.1687 < 2.2e-16

Two-way interactions:
---------------------
                              Coefficient  Std Err   T-stat Pr(|T|>t)
      Legendre polynomial  1       0.0196   0.0119   1.6421    0.1006
 with Legendre polynomial  1

Parameters in nonlinear transformations:
----------------------------------------
Legendre polynomial  1 for Longitude:
     Lower limit for polynomial rep:      5.6000 (prespecified)
     Upper limit for polynomial rep:      8.9000 (prespecified)
Legendre polynomial  1 for Latitude:
     Lower limit for polynomial rep:     50.1000 (prespecified)
     Upper limit for polynomial rep:     52.0000 (prespecified)
>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>

Spatial dependence structure:
-----------------------------
```

```
Structure used: Thresholded pow. exp. corr. fn: a + (1-a)exp[-ph
     Correlation decay rate, phi:      0.6674
     Power of distance, d:      0.5971
     Limiting correlation at large distances, a:     -0.0540
```

Appendix B

Figures

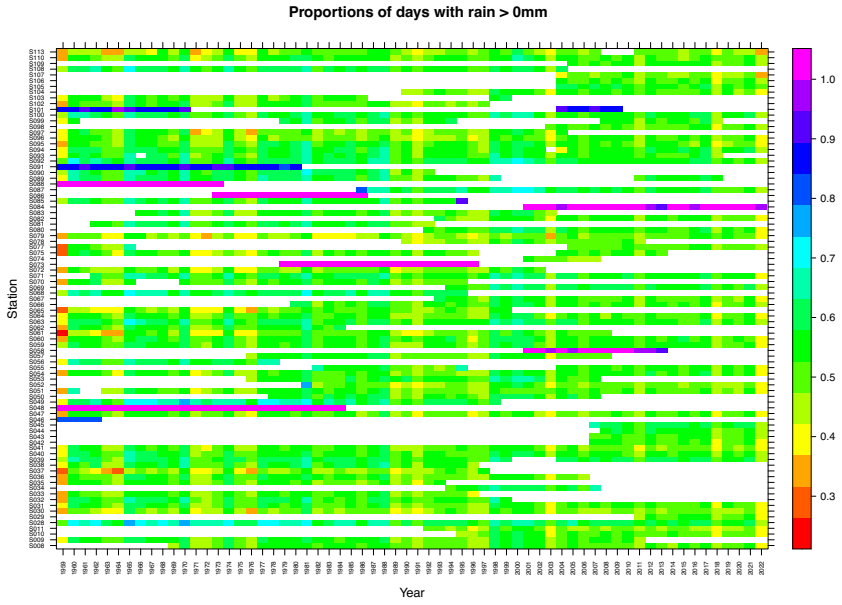

Figure B.1: Annual proportions of wet days at each site for the threshold of 0 mm. It can be seen that some sites that show anomalies have unusually wet or dry in certain periods.

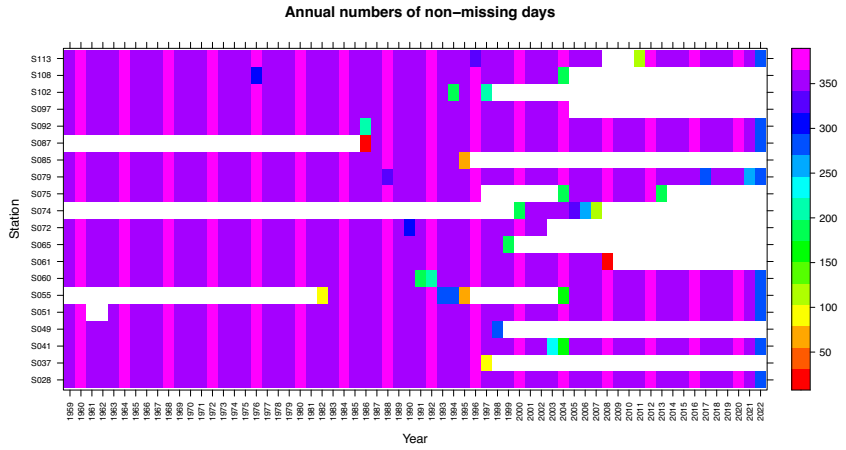

Figure B.2: Frequency of observations per year for these unusual sites to speculate on the reasons for the unusual precipitation at these sites.

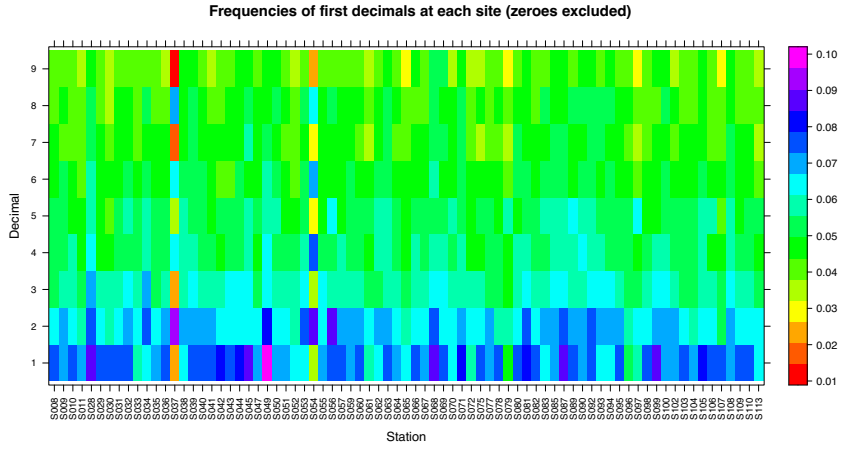

Figure B.3: Frequencies of first decimals at each site to check differences in recording resolution between sites. Most of the recording resolution is fairly homogeneous except for sites S037 and S054.

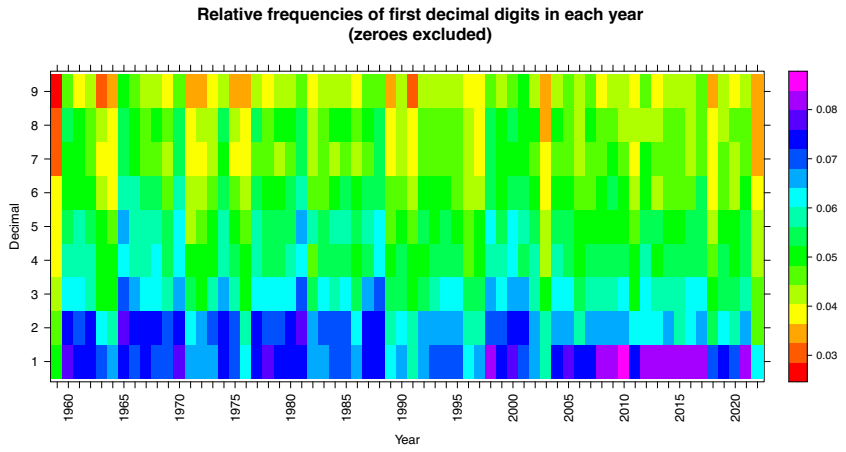

Figure B.4: Relative frequencies of first decimal digits in each year to check whether the recording resolution has changed over time.

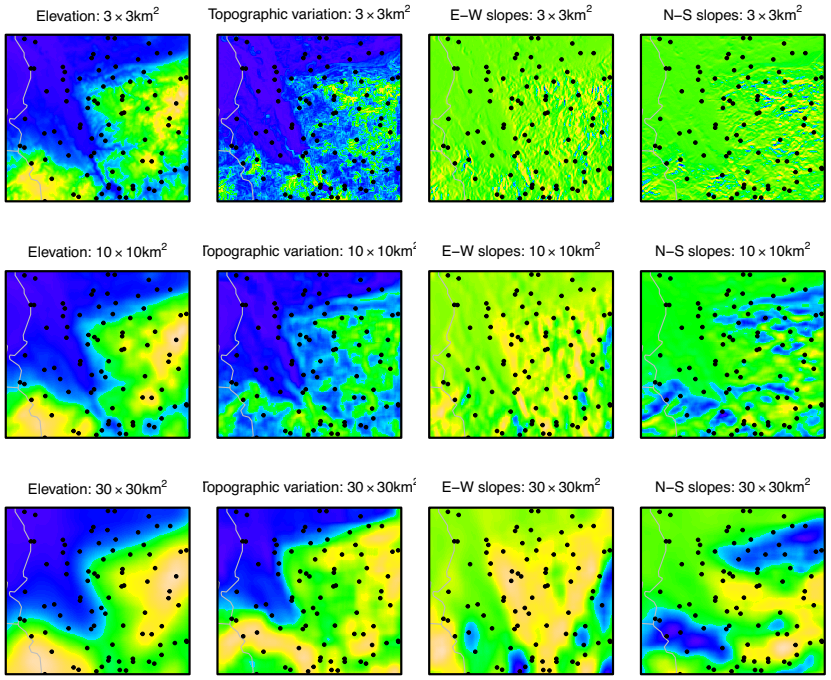

Figure B.5: Maps of different aspects of topographic variation across the region: altitude, altitude variation, east-west and north-south average slopes over domains with a centre distance of 3 × 3, 10 × 10, and 30 × 30 km2 for each site. Colour scales are indicative only and range from blue (0.06 m) to beige (8.17 m).

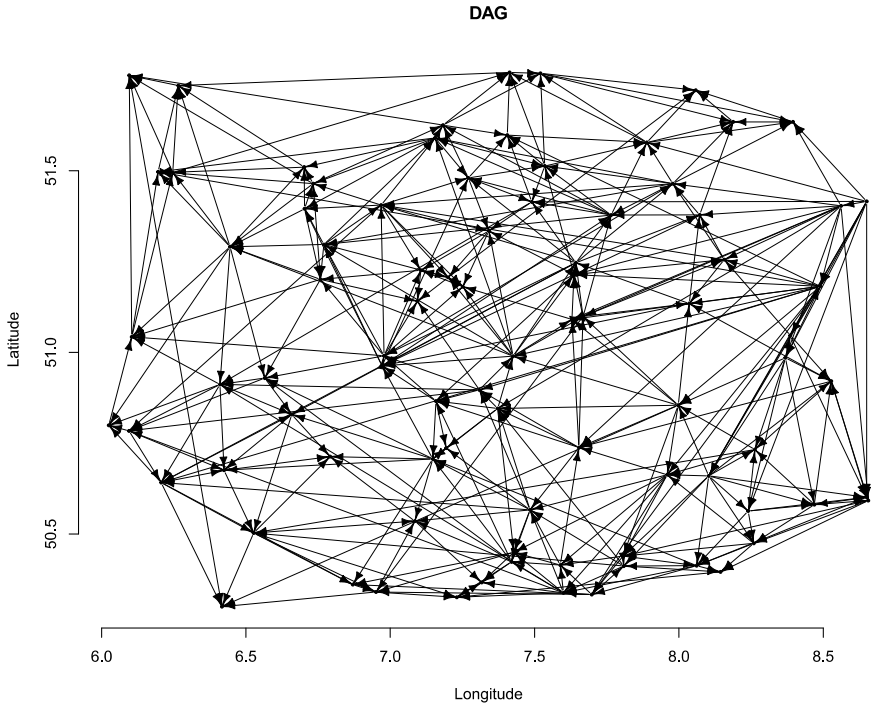

Figure B.6: DAG that shows the spatial dependence structure among 87 stations. The nodes represent stations, and the arcs show the dependence relationship among stations.

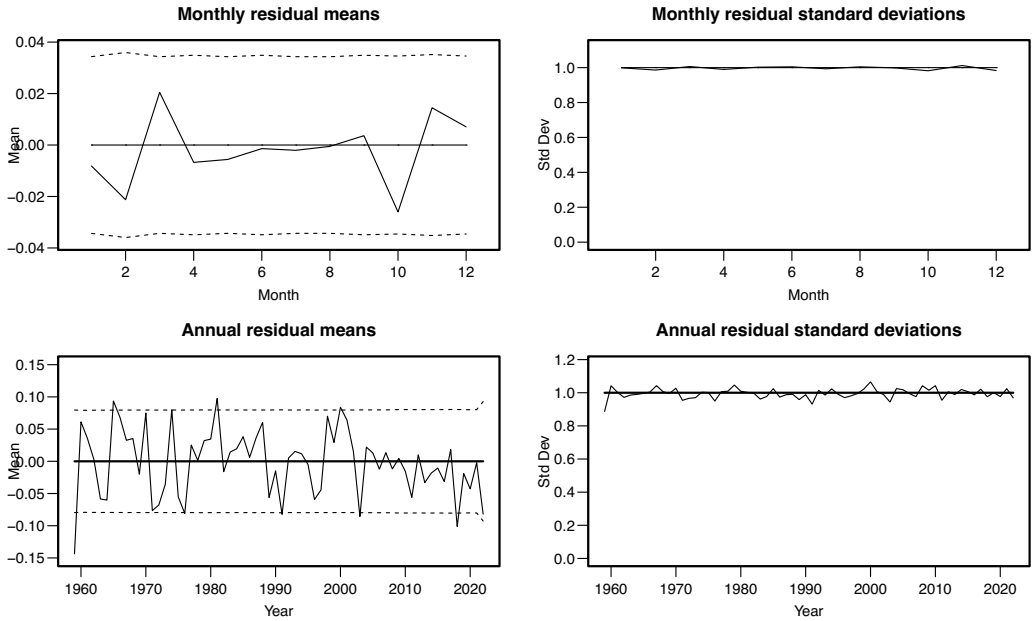

Figure B.7: Diagnostic plots of monthly and annual mean and standard deviation of Pearson residuals for first-stage precipitation occurrence model. The solid lines represent the theoretical means and standard deviations. The dashed lines in the left-hand plots show the 95% confidence interval of mean residuals.

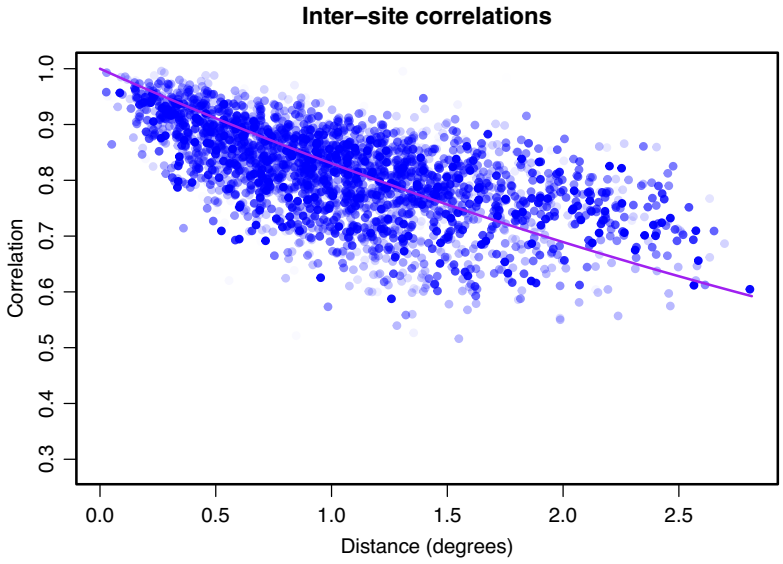

Figure B.8: Diagnostic plots of inter-site correlations for first-stage precipitation occurrence model. The darkness of the shading on each point indicates the number of days available that are used to calculate the correlation. The curve is fitted to each pair of inter-site correlations by the distance between sites using the exponential correlation function.

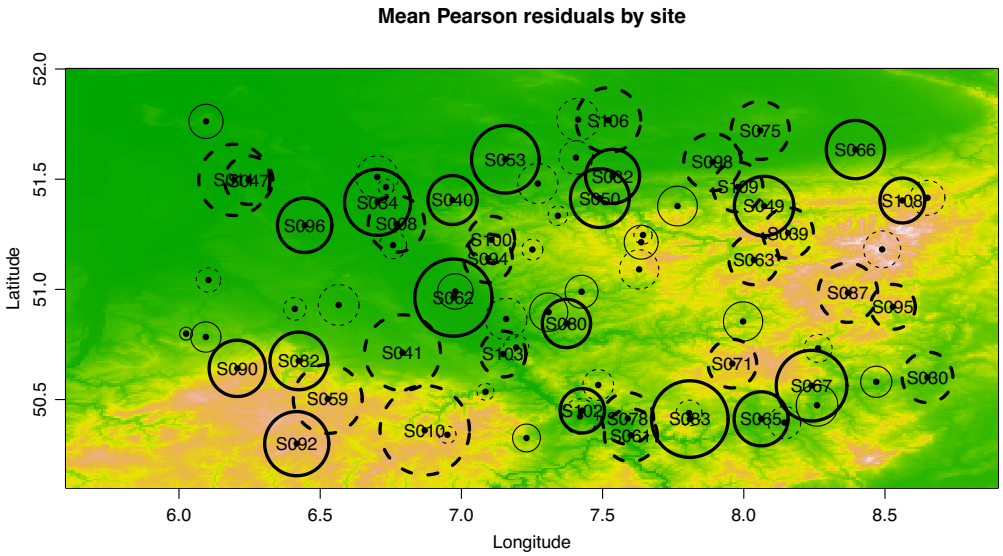

Figure B.9: Diagnostic plots of mean Pearson residuals by site for first-stage precipitation occurrence model. The solid circles represent positive values (i.e. observed proportion of wet days is higher than expected) and the dashed circles represent negative values. The circles that contain the name of the sites represent the mean residuals of those sites are significantly different from 0 at the 5% level.

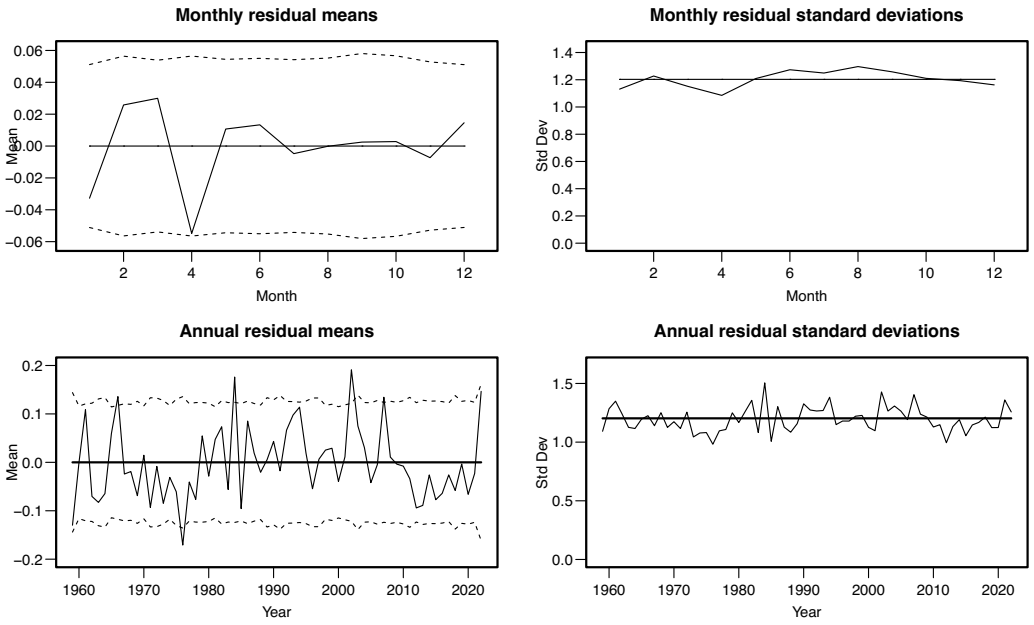

Figure B.10: Diagnostic plots of monthly and annual mean and standard deviation of Pearson residuals for first-stage precipitation amount model. The solid lines represent the theoretical means and standard deviations. The dashed lines in the left-hand plots show the 95% confidence interval of mean residuals.

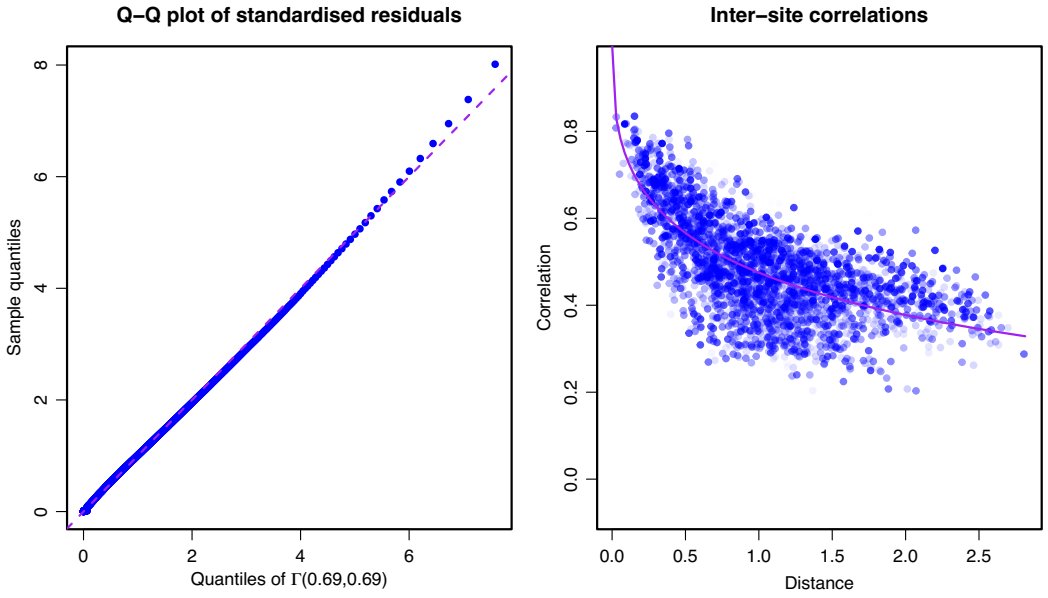

Figure B.11: For first-stage precipitation amount model, left: Q-Q plot of standardised residuals. The curve is fitted to a Gamma distribution with parameters (0.69, 0.69). Right: Diagnostic plots of inter-site correlations. The darkness of the shading on each point indicates the number of days available that are used to calculate the correlation. The curve is fitted to each pair of inter-site correlations by the distance between sites using the powered exponential correlation function.

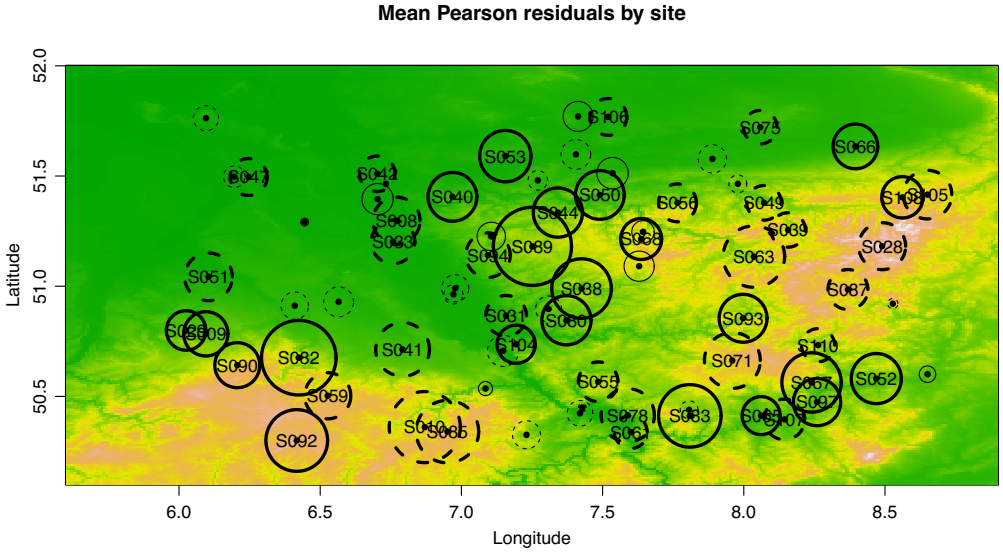

Figure B.12: Diagnostic plots of mean Pearson residuals by site for first-stage precipitation amount model. The solid circles represent positive values (i.e. observed precipitation amount in wet days is higher than expected) and the dashed circles represent negative values. The circles that contain the name of the sites represent the mean residuals of those sites are significantly different from 0 at the 5% level.

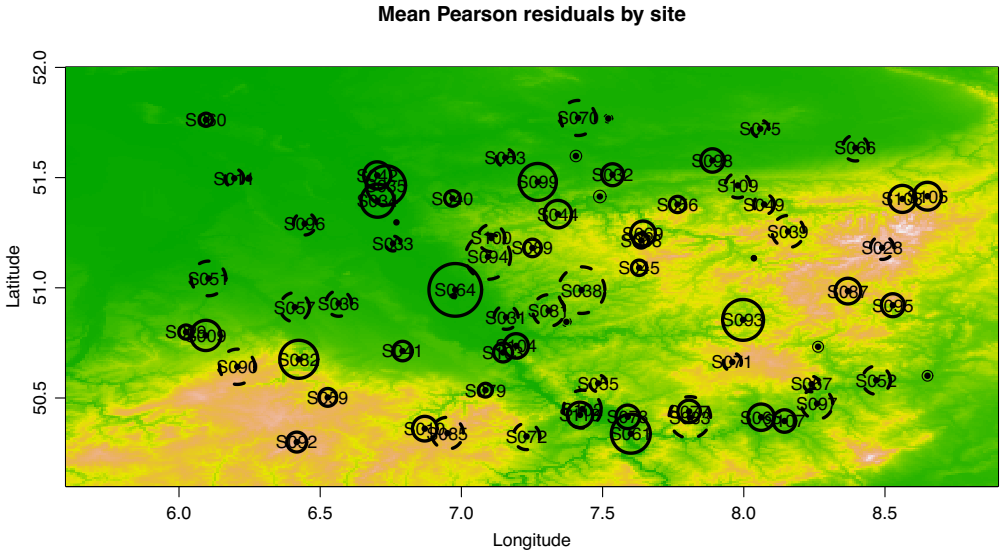

Figure B.13: Diagnostic plots of mean Pearson residuals by site for first-stage temperature model. The solid circles represent positive values (i.e. observed temperature is higher than expected) and the dashed circles represent negative values. The circles that contain the name of the sites represent the mean residuals of those sites are significantly different from 0 at the 5% level.

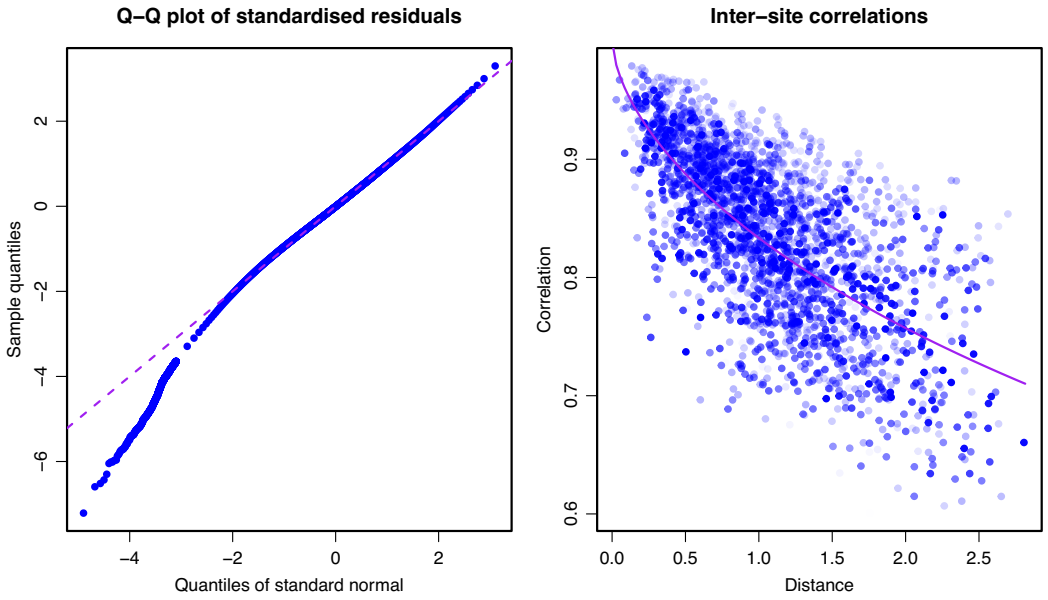

Figure B.14: For first-stage temperature model, left: Q-Q plot of standardised residuals. The curve is fitted to a standard normal distribution. Right: Diagnostic plots of inter-site correlations. The darkness of the shading on each point indicates the number of days available that are used to calculate the correlation. The curve is fitted to each pair of inter-site correlations by the distance between sites using the correlation function.

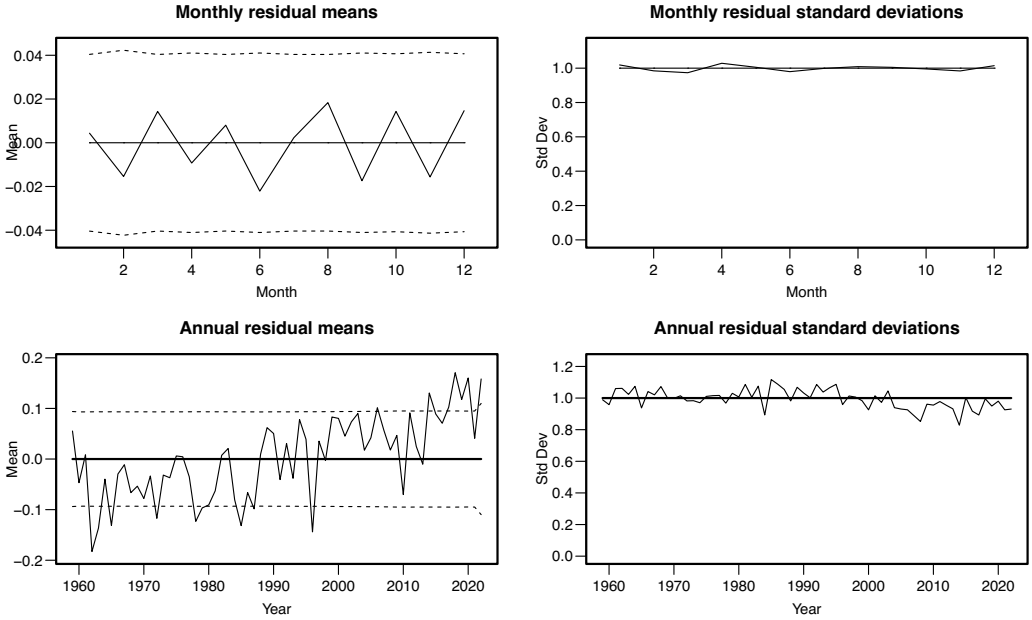

Figure B.15: Diagnostic plots of monthly and annual mean and standard deviation of Pearson residuals for first-stage temperature model. The solid lines represent the theoretical means and standard deviations. The dashed lines in the left-hand plots show the 95% confidence interval of mean residuals.

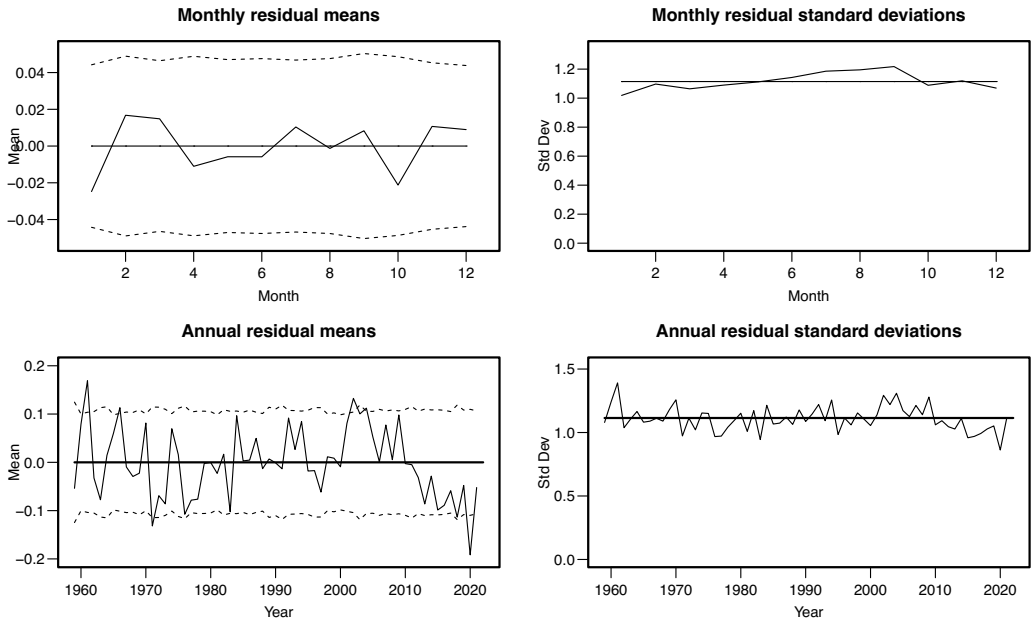

Figure B.16: Diagnostic plots of monthly and annual mean and standard deviation of Pearson residuals for third-stage temperature model. The solid lines represent the theoretical means and standard deviations. The dashed lines in the left-hand plots show the 95% confidence interval of mean residuals.

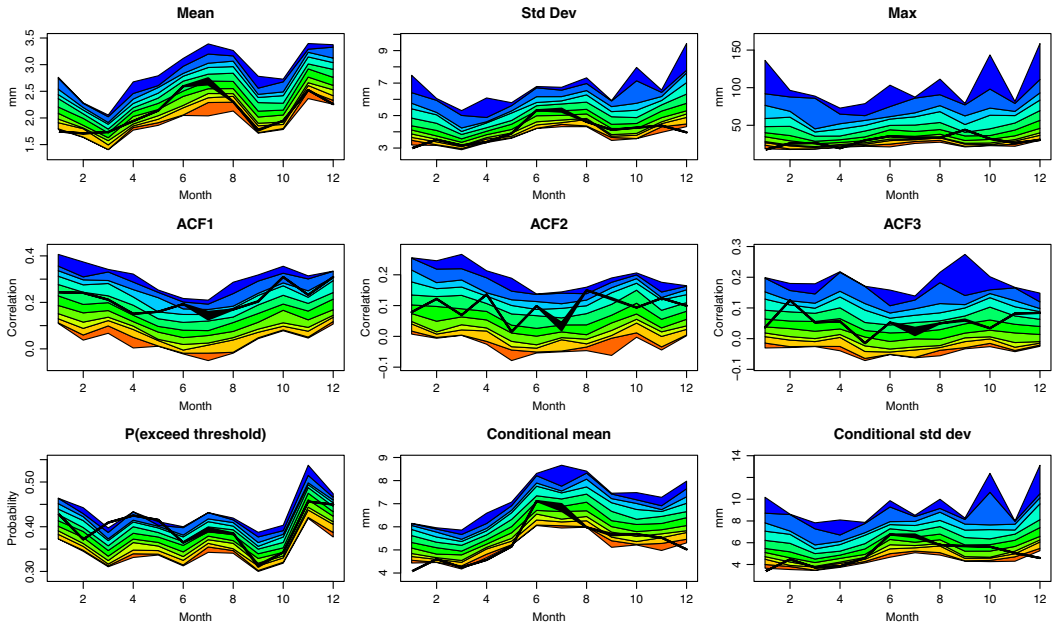

Figure B.17: Simulated and observed monthly summary statistics for precipitation at station S096 for fold 1 (1959-1974) for the statistical model. The coloured bands show the 1st, 5th, 10th, 25th, 75th, 90th, 95th and 99th percentiles of simulated distributions. The black bands show the 95% uncertainty interval of imputations. The plots show the mean, standard deviation, maximum, autocorrelation at lag 1-3, proportion of wet days, wet-day mean and wet-day standard deviation in order.

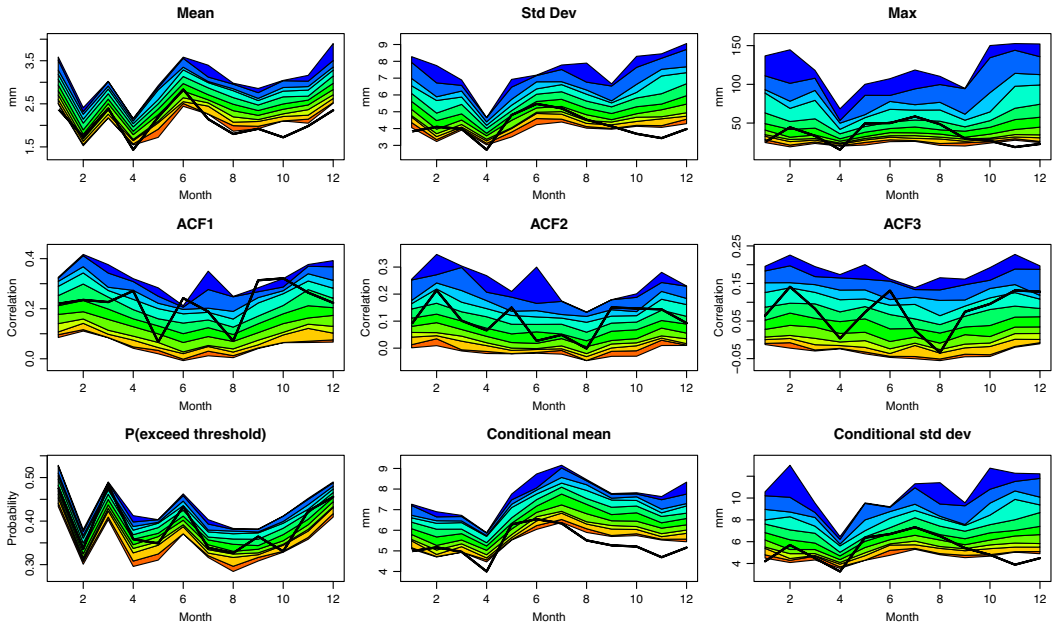

Figure B.18: As Figure B.17, but for fold 2 (1975-1990).

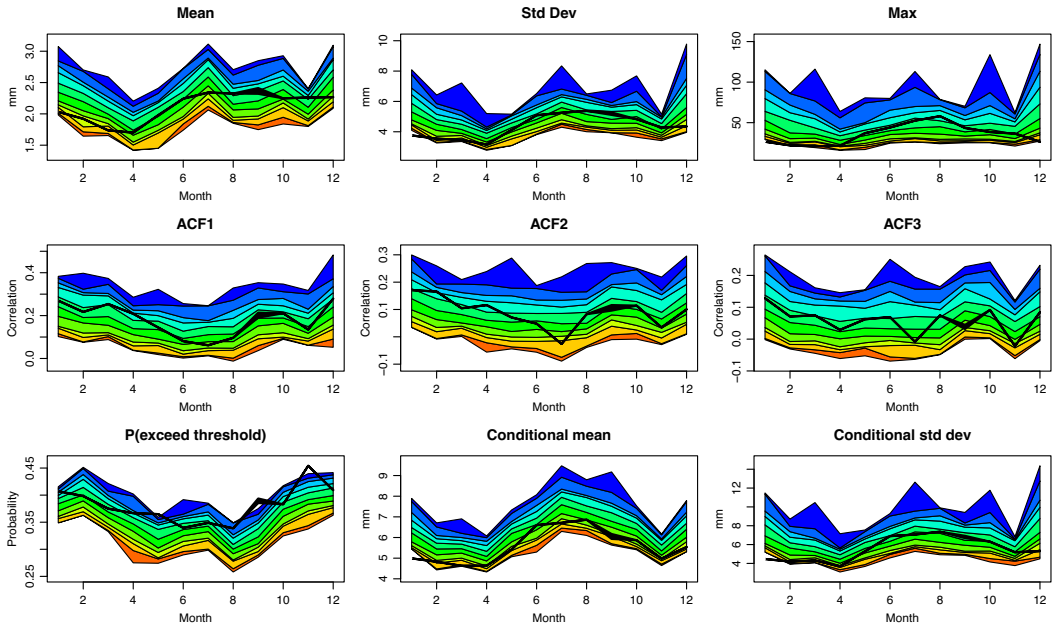

Figure B.19: As Figure B.17, but for fold 3 (1991-2006).

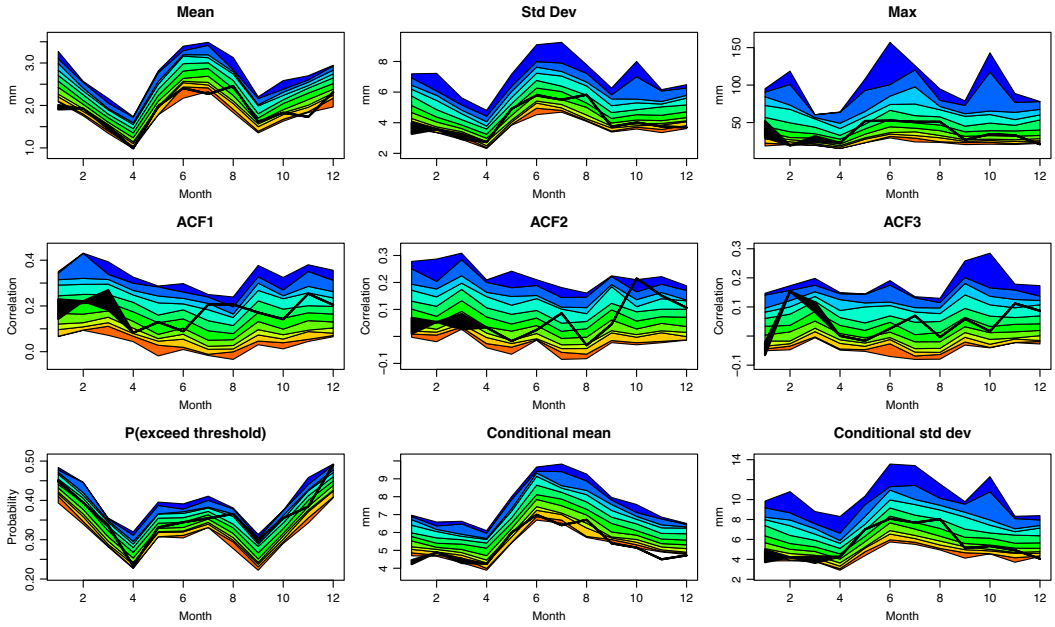

Figure B.20: As Figure B.17, but for fold 4 (2007-2021).

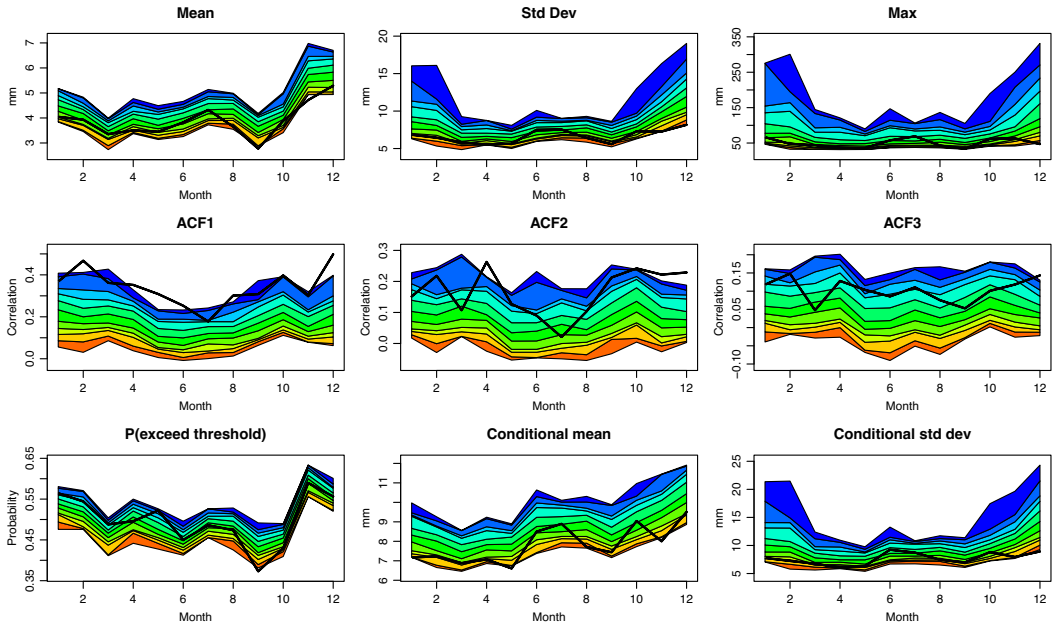

Figure B.21: Simulated and observed monthly summary statistics for precipitation at station S028 for fold 1 (1959-1974) for the statistical model. The coloured bands show the 1st, 5th, 10th, 25th, 75th, 90th, 95th and 99th percentiles of simulated distributions. The black bands show the 95% uncertainty interval of imputations. The plots show the mean, standard deviation, maximum, autocorrelation at lag 1-3, proportion of wet days, wet-day mean and wet-day standard deviation in order.

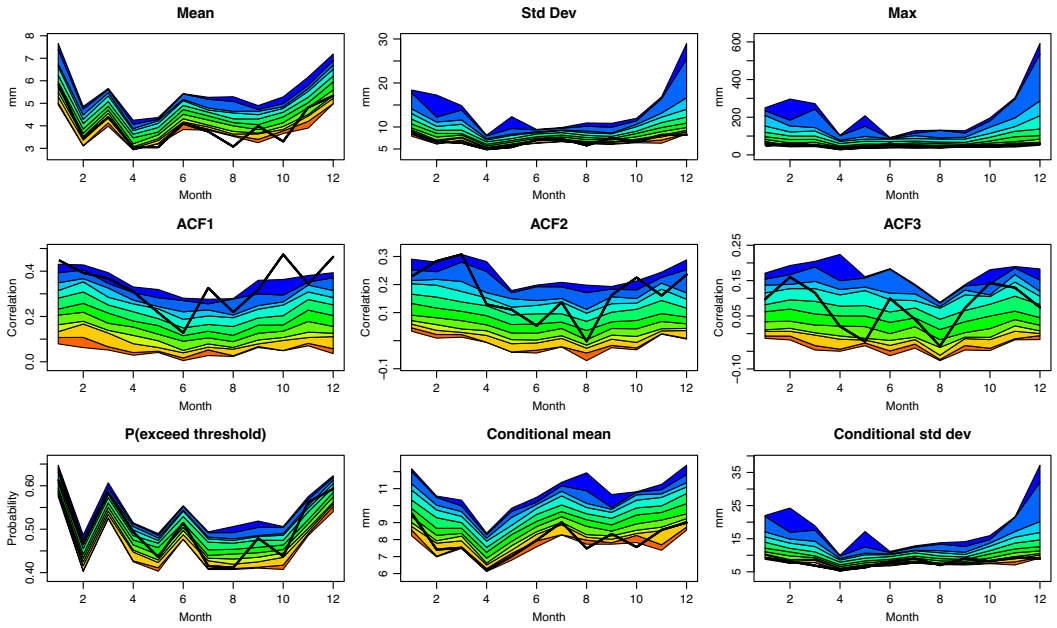

Figure B.22: As Figure B.21, but for fold 2 (1975-1990).

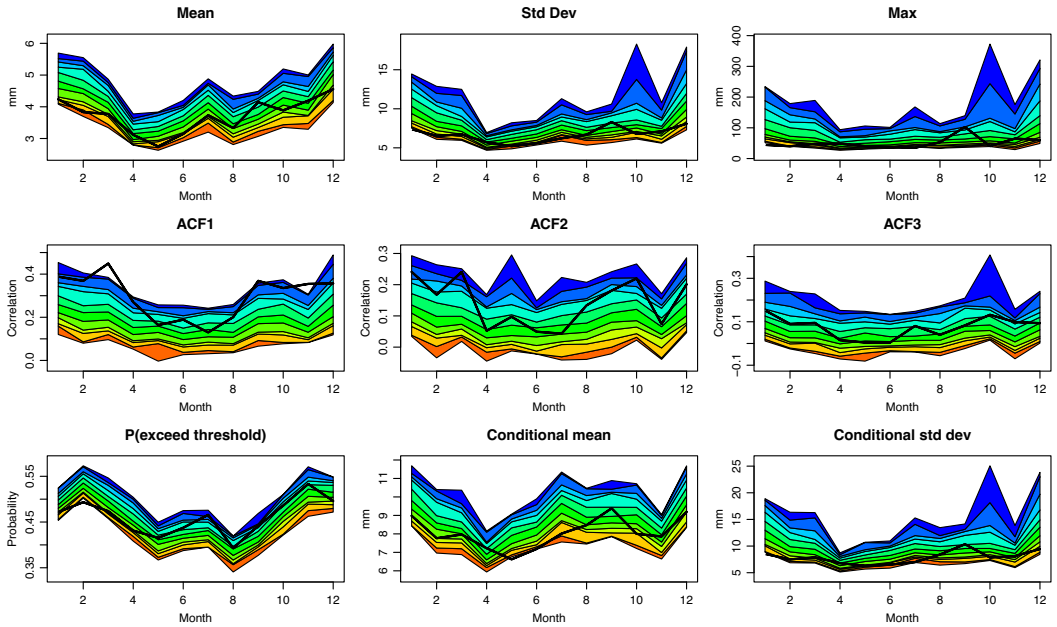

Figure B.23: As Figure B.21, but for fold 3 (1991-2006).

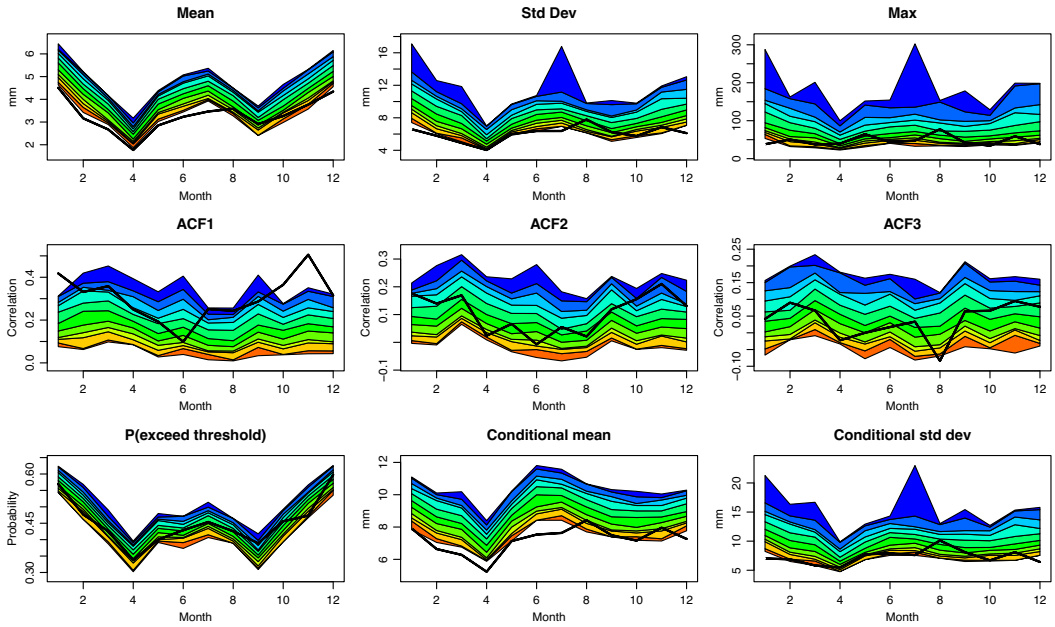

Figure B.24: As Figure B.21, but for fold 4 (2007-2021).

Bibliography

Chandler, R. E. (2018). *Rglimclim: A Multisite, Multivariate Weather Generator Based on Generalised Linear Models*. R package version 1.4-0.

Chandler, R. E. (2020). Multisite, multivariate weather generation based on generalised linear models. *Environmental Modelling & Software*, 134:104867.

De'ath, G. and Fabricius, K. E. (2000). Classification and regression trees: A powerful yet simple technique. *Ecology*, 81(11):3178–3192.

Do Hoai, N., Udo, K., and Mano, A. (2011). Downscaling global weather forecast outputs using ann for flood prediction. *Journal of Applied Mathematics*, 2011.

Fakhri, M., Farzaneh, M. R., Eslamian, S., and Khordadi, M. J. (2012). Uncertainty assessment of downscaled rainfall: impact of climate change on the probability of flood. *Journal of Flood Engineering*, 3(1):19–28.

Fekete, A. and Sandholz, S. (2021). Here comes the flood, but not failure? lessons to learn after the heavy rain and pluvial floods in germany 2021. *Water*, 13(21):3016.

Jang, S. and Kavvas, M. (2015). Downscaling global climate simulations to regional scales: statistical downscaling versus dynamical downscaling. *Journal of Hydrologic Engineering*, 20(1):A4014006.

Klein Tank, A., Wijngaard, J., Können, G., Böhm, R., Demarée, G., Gocheva, A., Mileta, M., Pashiardis, S., Hejkrlik, L., Kern-Hansen, C., et al. (2002). Daily dataset of 20th-century surface air temperature and precipitation series for the

european climate assessment. *International Journal of Climatology: A Journal of the Royal Meteorological Society*, 22(12):1441–1453.

Legasa, M. and Gutiérrez, J. M. (2020). Multisite weather generators using bayesian networks: An illustrative case study for precipitation occurrence. *Water Resources Research*, 56(7):e2019WR026416.

Legasa, M., Manzanas, R., Calviño, A., and Gutiérrez, J. M. (2022). A posteriori random forests for stochastic downscaling of precipitation by predicting probability distributions. *Water Resources Research*, 58(4):e2021WR030272.

Lehmkuhl, F., Hänsel, S., Knapp, J. L., Szabo, S., Zsoter, E., Staudinger, M., Zsebehazy, T., Unger-Shayesteh, K., Singh, A., Merchel, S., et al. (2022). Assessment of the 2021 summer flood in central europe. *Environmental Sciences Europe*, 34(1):1–14.

Maraun, D., Wetterhall, F., Ireson, A., Chandler, R., Kendon, E., Widmann, M., Brienen, S., Rust, H., Sauter, T., Themeßl, M., et al. (2010). Precipitation downscaling under climate change: Recent developments to bridge the gap between dynamical models and the end user. *Reviews of geophysics*, 48(3).

Maraun, D., Widmann, M., Gutiérrez, J. M., Kotlarski, S., Chandler, R. E., Hertig, E., Wibig, J., Huth, R., and Wilcke, R. A. (2015). Value: A framework to validate downscaling approaches for climate change studies. *Earth's Future*, 3(1):1–14.

Mcilveen, R. (1992). *Fundamentals of Weather and Climate*. Stanley Thornes, Cheltenham, UK.

Meehl, G. A., Stocker, T. F., Collins, W. D., Friedlingstein, P., Gaye, T., Gregory, J. M., Kitoh, A., Knutti, R., Murphy, J. M., Noda, A., et al. (2007). Global climate projections.

Nam, D. H., Mai, D. T., Udo, K., and Mano, A. (2014). Short-term flood inundation

prediction using hydrologic-hydraulic models forced with downscaled rainfall from global nwp. *Hydrological Processes*, 28(24):5844–5859.

Quesada-Chacón, D., Barfus, K., and Bernhofer, C. (2022). Repeatable high-resolution statistical downscaling through deep learning. *Geoscientific Model Development*, 15(19):7353–7370.

Rucker, C., Tull, N., Dietrich, J., Langan, T., Mitasova, H., Blanton, B., Fleming, J., and Luettich, R. (2021). Downscaling of real-time coastal flooding predictions for decision support. *Natural Hazards*, 107:1341–1369.

Scutari, M. and Denis, J.-B. (2014). *Bayesian Networks: With Examples in R*. Chapman and Hall/CRC.

Scutari, M., Graafland, C. E., and Gutiérrez, J. M. (2019). Who learns better bayesian network structures: Accuracy and speed of structure learning algorithms. *International Journal of Approximate Reasoning*, 115:235–253.

Tofiq, F. and Güven, A. (2015). Potential changes in inflow design flood under future climate projections for darbandikhan dam. *Journal of Hydrology*, 528:45–51.

Zimmermann, E. (2022). One year after the german flood disaster, angry victims complain: ”we have received nothing”. *World Socialist Web Site*.