# Multivariate postprocessing methods for temperature forecasts and examination of their limitations

Stevens, Ellen

*Award date:*
2021

Technische Universiteit
Eindhoven
University of Technology

5612 AZ Eindhoven
P.O. Box 513, 5600 MB Eindhoven
The Netherlands
www.tue.nl

# Multivariate postprocessing methods for temperature forecasts and examination of their limitations.

**Author**

Ellen Stevens (0951908)

**Supervisors**

Elisa Perrone
Kirien Whan (KNMI)
Edwin van den Heuvel

**Date**

September 7, 2021

# Abstract

Extreme weather events are the source of dangers and economic losses. Skillful and reliable weather forecasts can reduce these impacts. Weather forecasts are often derived from ensemble prediction systems, consisting of multiple members, where each member represents one run of a model with different initial conditions and model physics. Such ensemble forecasts typically reveals biases and dispersion errors. In this work, we apply Ensemble Model Output Statistics (EMOS) to correct for the bias and dispersion in the forecast for temperature at seven stations in the Netherlands for ten different lead times. EMOS works on a single variable, at a single location and for a single lead time. Therefore, the dependence between variables and stations is lost, which is a problem for several applications. To overcome this problem multivariate postprocessing methods that restore the dependence structure can be applied. In this work we compare the methods Ensemble Copula Coupling (ECC), the Schaake Shuffle, the SimSchaake and we introduce a new version of the SimSchaake. ECC restores the dependence based on the raw forecast, while the Schaake Shuffle and both versions of the SimSchaake restore the dependence based on historical observations. The energy and variogram scores show that ECC and the both versions of the SimSchaake perform best for short lead times and ECC performs best for longer lead times.

When applying one of the multivariate postprocessing methods, ties in the raw forecast or historical observations are resolved random. Performing ECC 100 times and comparing the variogram scores shows that forecasts with a large number of ties have more variability in the variogram scores compared to forecasts that barely contain ties. We look into the method 'first' where we assign ranks to ties in order from the first value to the last. We compare this method to the random method and we see that this method often performs better than the random method. This is the case for days where the raw forecasts of the stations have a high correlation and when the raw forecasts contain ties located at the same ensemble member.

# Contents

# 1 Introduction

Extreme weather influence nearly every aspect of our everyday lives. Providing accurate weather forecasts can lead to more effective planning and resource allocation by people and businesses. Weather forecasting is challenging due to the complexity of the atmospheric phenomena involved. Because of large uncertainty in the problem, ensemble prediction systems are often used. Such systems consist of multiple members, where each member represents one run of a numerical weather prediction model with different initial conditions and model physics.

To give an intuition of an ensemble weather forecast, we present a plot of a such an ensemble in Figure 1.1 for the 2-meter temperature. We do this for the initial day October 1 2015 and station Maastricht. We consider 51 ensemble members from the European Centre for Medium-Range Weather Forecasts (ECMWF). These 51 members are initialised at 00:00 UTC and the forecasts are valid at 12:00 for the first 10 days. This means we look at 10 different lead times from 12 up to 228 hours.

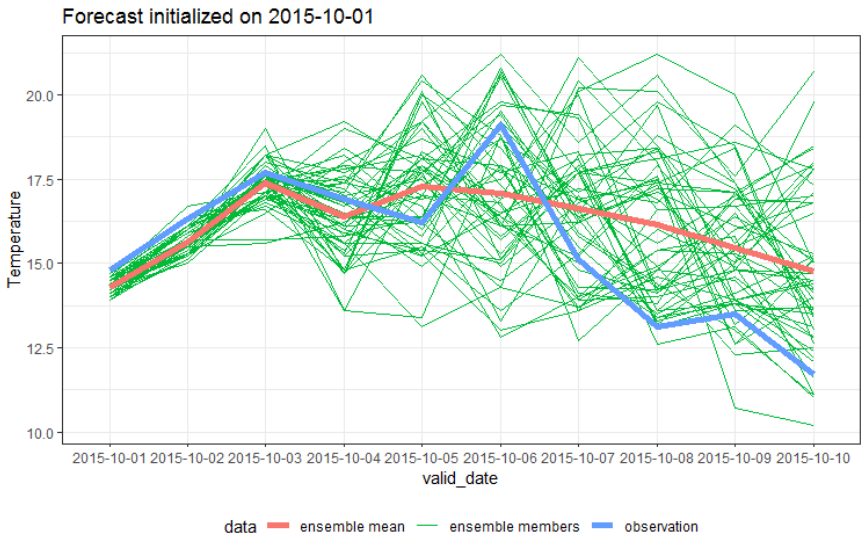


Figure 1.1: *Ensemble forecasts for temperature for various lead times on 2015-10-01 for station Maastricht*

In Figure 1.1 every ensemble member corresponds to a green line and shows the temperature forecast over time. The red line gives the mean of the 51 forecasts and the blue line shows the actual observation. On average, the larger the lead time, the larger the difference between the mean of the forecast and the observation. Thus, forecast skill typically decreases with increasing lead time.

In Figure 1.1, we also notice that the actual observations for October 1 and October 2 are higher than (almost) every ensemble members. This is not surprising as, especially for short lead times, an ensemble forecast typically has biases and dispersion errors. To remove the bias and dispersion errors in these forecasts, different statistical postprocessing methods are used. These methods use historical forecasts and the corresponding observation to obtain corrected forecasts.

Some postprocessing methods that have been proposed are Bayesian model averaging (BMA) [1][2], ensemble model output statistics (EMOS) [3] and machine learning techniques [4][5]. These methods are univariate postprocessing methods as they correct for bias and dispersion in the forecast, but only apply to a single variable, at a single location and for a single lead time.
EMOS uses the summary statistics of the raw forecast and the observations of a specific number of training days to correct the raw forecast. EMOS was first developed for temperature, but it has also been applied to other weather variables such as precipitation [6] and wind speed [7][8].

EMOS is a univariate postprocessing method, and does not take the dependencies between locations, lead times and/or weather variables into account. This is a problem for several applications, such as flood analysis [9], winter road maintenance [10], handling of renewable energy sources [11][12][13] and air traffic management [14], where including dependencies in the model is key. To address this problem, several multivariate postprocessing methods are used. Examples of parametric multivariate postprocessing methods are spatial BMA and spatial EMOS [15][16]. However, these methods are computationally expensive for settings with a large number of variables and they require strong parametric assumptions. To overcome this limitation there are other methods which are rank-based non-parametric and are based on mathematical tools called empirical copulas. Examples of such methods are ensemble copula coupling (ECC) [17], the Schaake Shuffle [18] and the SimSchaake [19].
ECC models dependencies based on the rank structure in the raw forecast, while the Schaake Shuffle and SimSchaake model these dependencies based on the rank structure in historical observations.

ECC, the Schaake Shuffle and the SimSchaake use a discretized version of the corrected forecasts that were found after performing EMOS. The rank structure in the raw forecast or in historical observations is used to restore the dependencies between stations and weather variables in these corrected forecasts. Namely, we reshuffle the corrected forecasts based on the the ranks of the raw forecast or historical observations. This approach works because it preserves the Spearman rank correlation structure between station pairs of the raw forecasts or historical observations.

Each multivariate method has its advantages, but also its limitations. The ability of ECC to represent dependence structures depends on the ability of the raw ensemble to simulate these dependencies, which is it not always the case to a sufficient degree. Another limitation of ECC compared to the Schaake Shuffle and SimSchaake is that the number of members in the multivariate postprocessed forecast always equals the number of members in the raw forecast, because the raw forecast is used as dependence template. For the Schaake Shuffle and SimSchaake we can select as many historical observations as we want which results in more values in the multivariate forecast.
For the Schaake Shuffle the random selection of days is very important. For the SimSchaake the days are selected based on similarities in the raw forecast. For short lead times comparing raw forecasts might give good results, but raw forecasts for longer lead times do not often match the actual observation, as it can be seen in the forecast for October 8, 9 and 10 in Figure 1.1. Therefore, when using the raw forecast with longer lead times the selection of dates will be less meaningful.

The values in the raw forecast are based on different models an initial conditions. These prediction models and initial conditions will possibly change over time. Therefore comparing a raw forecast of the current year with a forecast from a few years ago might be a comparison between two forecasts that are based on different models. We introduce a new method that is an adaption of the SimSchaake. The adapted SimSchaake to solve this problem by using the corrected forecast for comparison instead of the raw forecast.

Another limitation of rank based methods is given by the presence of repeated values, i.e., ties. When ties occur in the raw forecast or in the historical observations, the ranks are typically resolved randomly in all multivariate methods. This might not be a problem when barely any ties occur, but what happens if we have more ties in the raw forecast or in historical observations? We present a case and simulation study to investigate this. This limitation has not been discussed in the literature.

The performance of ECC, the Schaake Shuffle and the SimSchaake has been investigated before [20][21][22]. Perrone [22] compared the performance of ECC and the SimSchaake for the Austrian ensemble system ALADIN-LAEF, which consists of 17 ensemble members. Perrone considered 3 groups of 3 stations and the weather variable temperature. Results show that for this setting the SimSchaake outperforms ECC.
Wilks [20] compared the performance of ECC and the Schaake Shuffle based on the 11-member National Oceanic and Atmospheric Administration (NOAA) reforecasts ensembles. Wilks considered one station and the weather variables temperature and dew point temperature. Results show that the Schaake Shuffle outperforms ECC, but also show that ECC would probably be more competitive if it would be used with larger ensembles such as the ECMWF system.
Whan [21] compared ECC to the Schaake Shuffle for the ECMWF system. The same as in Wilks' setting, the weather variables temperature and dew point temperature are considered, but instead of one station, seven stations are considered. Results show that indeed ECC outperforms the Schaake Shuffle.

In this work, we consider almost the same setting as Whan [21]. However, we will only consider the temperature for the seven stations and do not consider the dew point temperature. Besides ECC and the Schaake Shuffle we will also look at the performance of both version of the SimSchaake and at the performance of ECC when different methods for solving ties are used. We present a case study to answer the following questions:

1. How does EMOS perform for the different stations and lead times?
2. How do the multivariate postprocessing methods ECC, the Schaake Shuffle and the SimSchaake perform for the different lead times?
3. How does a multivariate method that selects historical observations based on similarities in the corrected forecasts perform for the different lead times?
4. What is the effect of ties on the performance of the multivariate postprocessing methods?
5. How do other methods for solving ties perform compared to the random method?

The remainder of the thesis is organized as follows. In Section 2 we discuss the previous mentioned methods in detail and we introduce tools to assess the quality of the raw and postprocessed forecasts. We also introduce the data with 51 ensemble members of the ECMWF system. In Section 3 we present our case study of temperature forecasts for seven stations in the Netherlands. In Section 4 we investigate the issue of ties in the case study setting. Next, in Section 5, we conduct a simulation study and investigate the impact of different dependence

structures in the raw forecast on solving ties. We finish in Section 6 with a conclusion and discussion.

# 2 Methods and Data

As discussed in the introduction, an ensemble forecast typically reveals biases and dispersion errors. Postprocessing of the forecast is performed to correct for such bias and dispersion in the forecast. In this chapter, we describe the methods that are used to statistically postprocess the forecast. Furthermore, we discuss the evaluation of these methods through verification tools. Finally, we introduce the data.

## 2.1 Methods

In this section we present univariate and multivariate postprocessing methods.

### 2.1.1 Univariate Postprocessing

In this work we use a method called ensemble model output statistics (EMOS) to postprocess our raw ensemble forecast [3]. The goal of EMOS is to maximize sharpness subject to calibration, where sharpness refers to the concentration of the predictive distributions and calibration refers to the statistical compatibility between the forecast and the observation [23].

EMOS is a technique that can be used to estimates parameters of the distribution of a weather variable. In particular, EMOS links the parameters of a forecast distribution to the characteristics of an ensemble prediction system. Here we link the parameters to the ensemble mean and variance. This estimation is based on the raw forecast and the matching observations from the previous $n$ days. We use the following regression model where $y$ is the weather variable temperature:

$$\begin{aligned} y &\sim N(\mu, \sigma^2) \\ \mu &= a + b \cdot \bar{x} \\ \sigma^2 &= c + d \cdot s^2 \end{aligned}$$

where $a$ and $c$ are the intercept coefficients, $b$ and $d$ the slope coefficients, $\bar{x}$ the mean of the raw forecast and $s^2$ the variance of the raw forecast. The coefficients $c$ and $d$ are restricted to be nonnegative, because $\sigma^2$ should be positive.

One could use various approaches to estimate the parameters. A technique that can be used to estimate them is maximum likelihood. The log-likelihood function for the statistical model is

$$l(a, b, c, d) = -\frac{1}{2}\left(n\log(2\pi) + \sum_{i=1}^{n}\frac{(y_i - \mu_i)^2}{\sigma_i} + \sum_{i=1}^{n}\log(\sigma_i)\right)$$

where we sum over the $n$ training days [24].

Studies found that, especially for temperature, the continuous ranked probability score (CRPS) tends to give sharper results than maximum likelihood [25][26]. Therefore, in this work the four

coefficients are estimated by minimizing the mean CRPS [3]. We do this over a number of training days. To calculate this CRPS for one day we use the formula

$$\text{CRPS} = \sigma \left( \frac{y-\mu}{\sigma} \left( 2\Phi \left( \frac{y-\mu}{\sigma} \right) - 1 \right) + 2\phi \left( \frac{y-\mu}{\sigma} \right) - \frac{1}{\sqrt{\pi}} \right)$$

where $\Phi$ denotes the CDF of the standard normal distribution and $\phi$ denotes the PDF of the standard normal distribution.

Minimizing the CRPS gives the four coefficients $a, b, c$ and $d$. This results in a predictive normally distributed PDF per day, per lead time and per station.

### 2.1.2 Multivariate Postprocessing

As we mentioned in the introduction, EMOS works on a single station and does not preserve the spatial and temporal covariability between stations. To model these dependencies between stations and obtain multivariate postprocessed distributions, we can apply various multivariate postprocessing methods.

We now introduce the notation. Let $j \in \{1, ..., J\}$ be a station and let $k \in \{1..., K\}$ be a lead time, $J, K \in \mathbb{N}$. This summarizes to the multi-index $l = (j, k)$. We have the raw forecast $x_1^l, ..., x_M^l$ of length $M \in \mathbb{N}$.

In the Section 2.1.1 we have seen that by performing EMOS on the raw forecasts $x_1^l, ..., x_M^l$, we find a predictive normally distributed PDF per day. Because we use rank-based methods to restore the dependence, we need to discretize the PDFs obtained from EMOS. The most common method to discretize the PDFs is by taking equidistant quantiles $\frac{1}{N+1}, ..., \frac{N}{N+1}$. If we do this for every PDF obtained after performing EMOS, we find the corrected forecast $\tilde{x}_1^l, ..., \tilde{x}_N^l$ of length $N$.

Applying ECC, Schaake Shuffle or SimSchaake gives the multivariate forecast $\hat{x}_1^l, ..., \hat{x}_N^l$. We will now discuss these three multivariate postprocessing methods.

#### Ensemble Copula Coupling

In the ensemble copula coupling (ECC) approach, the ranks of the raw forecast are used to perform a multivariate correction of the forecast distribution. Because we use the raw forecast as dependence template, the discrete corrected forecast should have the same length as the raw forecast. Therefore we have that $N$ should be equal to $M$, which results in the corrected forecast $\tilde{x}_1^l, ..., \tilde{x}_M^l$.

To find the multivariate forecast we rearrange the corrected forecast in the rank order structure of the raw ensemble. We use the function $\sigma_l$ to denote the ranks of the raw forecast. We have $\sigma_l(i) = rank(x_i^l)$. Then if we apply ECC to the discrete sample we find that

$$\hat{x}_1^l = \tilde{x}_{\sigma_l(1)}^l, ..., \hat{x}_M^l = \tilde{x}_{\sigma_l(M)}^l$$

**Example 2.1**
We will give an example of ECC where we consider 5 members. Let the raw forecast of one day, one lead time at three stations be given as

$$S_{raw} = \begin{pmatrix} 4.8 & 5 & 4.9 & 5.0 & 5.4 \\ 4.8 & 5.2 & 5.3 & 4.5 & 5.0 \\ 5.5 & 5.6 & 5.1 & 5.0 & 4.8 \end{pmatrix}$$

Ranking the 5 ensemble members separately by station gives the rank matrix

$$R = \begin{pmatrix} 1 & 4 & 2 & 3 & 5 \\ 2 & 4 & 5 & 1 & 3 \\ 4 & 5 & 3 & 2 & 1 \end{pmatrix}$$

Univariate post-processing produces the following forecasts

$$S_{EMOS} = \begin{pmatrix} 5.1 & 5.2 & 5.3 & 5.4 & 5.5 \\ 5.2 & 5.25 & 5.3 & 5.35 & 5.4 \\ 4.8 & 5.0 & 5.2 & 5.4 & 5.6 \end{pmatrix}$$

If we reshuffle the post-processed forecasts according to the dependence template we find the multivariate forecast

$$S_{ECC} = \begin{pmatrix} 5.1 & 5.4 & 5.2 & 5.3 & 5.5 \\ 5.25 & 5.35 & 5.4 & 5.2 & 5.3 \\ 5.4 & 5.6 & 5.2 & 5.0 & 4.8 \end{pmatrix}$$

This gives the multivariate forecast where dependencies between stations are restored.

## Schaake Shuffle

In the Schaake Shuffle approach we use historical observations instead of the raw forecast. The size of the postprocessed ensemble is not restricted to be equal to the size $M$ of the raw forecast.

To find the multivariate forecast we rearrange the corrected forecast in the rank order structure of the historical observations. Let $o_1^l, ..., o_N^l$ denote the observations of $N$ randomly selected days. Now, we use the function $\sigma_l$ to denote the ranks of the historical observations. We have $\sigma_l(i) = rank(o_i^l)$. The Schaake Shuffle consists of

$$\hat{x}_1^l = \tilde{x}_{\sigma_l(1)}^l, ..., \hat{x}_N^l = \tilde{x}_{\sigma_l(N)}^l$$

These $N$ historical observations are not selected randomly over all available historical days, but they are randomly selected from the training days within a range of a number of days before or after the target day. By doing this we select days in the past with the same climatology as the target day, if the climate is stationary.

## SimSchaake

By applying the Schaake Shuffle it is not guaranteed that the past observations that are randomly selected resemble the current forecast. The SimSchaake overcomes this limitation by selecting the historical observations based on similarities within the raw forecast [19].
We compare the current raw forecast $x^t$ to all raw forecasts in the training days $x^1, ..., x^{t_D}$, and select the historical observations of the day where the raw forecasts are most similar to the current raw forecast. We use the similarity criterion introduced by Schefzik [19] that compares the means and standard deviations of the raw forecasts.

For every $x^{t_d} \in \{x^1, ..., x^{t_D}\}$ we use the similarity criterion

$$\Delta(x^t, x^{t_d}) = \sqrt{\frac{1}{L}\sum_{l=1}^{L}(\bar{x}^{l,t} - \bar{x}^{l,t_d})^2 + \frac{1}{L}\sum_{l=1}^{L}(s^{l,t} - s^{l,t_d})^2}$$

where

- $\bar{x}^{l,\tau}$ is the mean of the raw forecast at day $\tau$ and multi-index $l$
- $s^{l,\tau}$ is the standard deviation of the raw forecast at day $\tau$ and multi-index $l$

Let $o_1^l, ..., o_N^l$ denote the observations of the $N$ days with the lowest value for $\Delta$. Then the Schaake Shuffle is performed as described in Section 2.1.2.

**Example 2.2**
To calculate this similarity criterion $\Delta$ we compare the raw forecast of the day we are correcting for to the raw forecast of all training days. Let us cover one example of this calculation. We have the raw forecast at two stations

$$S_{raw,obs} = \begin{pmatrix} 2.1 & 1.0 & 3.7 & 5.4 & 4.4 \\ 2.5 & 2.6 & 4.0 & 4.3 & 5.9 \end{pmatrix}$$

On this day station 1 has a mean of $\bar{x}_{raw,1} = 3.32$ and a standard deviation of $s_{raw,1} = 1.768$ and station 2 has a mean of $\bar{x}_{raw,2} = 3.86$ and a standard deviation of $s_{raw,2} = 1.397$.

We want to compare this to all forecasts in the training days. One of these training days has the raw forecast

$$S_{raw,histobs} = \begin{pmatrix} 6.0 & 8.0 & 5.9 & 4.7 & 8.4 \\ 6.7 & 9.1 & 9.2 & 5.4 & 7.8 \end{pmatrix}$$

On this day station 1 has a mean of $\bar{x}_{hist,1} = 6.60$ and a standard deviation of $s_{hist,1} = 1.554$ and station 2 has a mean of $\bar{x}_{hist,2} = 7.64$ and a standard deviation of $s_{hist,2} = 1.620$.

The similarity criterion $\Delta$ can be calculated as

$$\begin{aligned} \Delta &= \sqrt{\frac{(\bar{x}_{hist,1} - \bar{x}_{raw,1})^2 + (\bar{x}_{hist,2} - \bar{x}_{raw,2})^2}{2} + \frac{(s_{hist,1} - s_{raw,1})^2 + (s_{hist,2} - s_{raw,2})^2}{2}} \\ &= \sqrt{\frac{(6.60 - 3.32)^2 + (7.64 - 3.86)^2}{2} + \frac{(1.554 - 1.768)^2 + (1.620 - 1.397)^2}{2}} \\ &= 3.546 \end{aligned}$$

We perform this calculation for every day in the training days and we select the $N$ days with the smallest $\Delta$ as historical observations.

## SimSchaake using the corrected forecast

The values in the raw forecast are based on different models an initial conditions. These prediction models and initial conditions will possibly change over time. Therefore comparing a raw forecast of the current year with a forecast from a few years ago might be a comparison between two forecasts that are based on different models.

We try to solve this problem by using the corrected forecast for comparison instead of the raw forecast.

For every $\tilde{x}^{t_d} \in \{\tilde{x}^1, ..., \tilde{x}^{t_D}\}$ we use the same similarity criterion

$$\Delta(\tilde{x}^t, \tilde{x}^{t_d}) = \sqrt{\frac{1}{L}\sum_{l=1}^{L}(\mu^{l,t} - \mu^{l,t_d})^2 + \frac{1}{L}\sum_{l=1}^{L}(\sigma^{l,t} - \sigma^{l,t_d})^2}$$

where

- $\mu^{l,\tau}$ is the mean of the PDF after EMOS at day $\tau$ and multi-index $l$
- $\sigma^{l,\tau}$ is the standard deviation of the PDF after EMOS at day $\tau$ and multi-index $l$

Let $o_1^l, ..., o_N^l$ denote the observations of the $N$ days with the lowest value for $\Delta$. Then the Schaake Shuffle is performed as described in Section 2.1.2.

## 2.2 Verification and evaluation of the methods

As discussed in Section 2.1, the goal in probabilistic forecasting is to maximize the sharpness of the predictive distributions subject to calibration. In this Section we discuss scoring rules that are able to evaluate calibration and sharpness simultaneously, and we discuss various rank histograms to assess the calibration of all univariate and multivariate methods.

### 2.2.1 Scoring rules

We use scoring rules to quantify the skill of the forecasts. In the univariate setting, one of the most commonly used scoring rules is the CRPS that we discussed in Section 2.1.1. In the multivariate setting, we use the energy and variogram scores [27]. Both scores assign a numerical score $S(F, y)$ to each pair $(F, y)$ with $F$ the multivariate forecast and $y$ the observation. The values represent a distance between the forecast and the observations, so the lower, the better.

**Energy Score**
The energy score is a multivariate version of the CRPS and can be calculated with the formula

$$S_{en}(F, y) = \mathbb{E} \mid\mid X - y \mid\mid -\frac{1}{2}\mathbb{E} \mid\mid X - X' \mid\mid$$

where $X$ and $X'$ are independent random vectors that are distributed according to $F$, $y$ is the vector of observations and $|| \cdot ||$ is the Euclidean norm [24][27].

**Variogram Score**
The variogram score of order $p$ can be calculated with the formula

$$S_{vs_p}(F, y) = \sum_{i,j=1}^{d} w_{ij}(\mid y_i - y_j \mid^p - \mathbb{E} \mid X_i - X_j \mid^p)^2$$

where $X_i$ and $X_j$ are the $i$th and $j$th component of a random vectors $X$ that is distributed according to $F$, $y$ is the vector of observations and $w_{ij}$ are nonnegative weights [27]. By using different weights for pairs we can change the variogram score. In spatial context, it is possible to add weight based on the distance between the different stations [28]. However, previous research showed that adding weights in our case study will not improve the variogram scores [21]. In general, $p = 0.5$ is used [27].

Both scores are calculated using the 'ScoringRules' package in $\mathrm{R}$ [29].

### 2.2.2 Calibration

Besides evaluating the forecast skill, we apply visual verification tools to quantify the calibration, which refers to the statistical compatibility between the forecast and the observation. A forecast is calibrated if the observation can be seen as a random draw from the forecast distribution. Therefore, the ranks of the observations should be uniformly distributed. We use rank histograms to show if the forecast distributions are uniformly distributed. A rank histogram gives the ranks for all the days in the data set.
We discuss how to calculate these ranks for the univariate and multivariate calibration.

#### Univariate calibration

The rank for one specific day is calculated as follows. Let $S = \{y, x_1, ..., x_M\}$ be a vector that contains the observation $y$ at the specific day, and the corrected forecast for the $M$ ensemble

members. To calculate the rank for the observation we use the formula

$$rank(y) = \sum_{u \in S} \mathbb{1}\{u \leq y\}$$

If the observation is lower than (almost) all ensemble values, it will result in a low rank and if the observation is higher than (almost) all ensemble values, it will result in a high rank. As said, for a perfect calibrated forecast we would see that every rank between $1$ and $M+1$ occurs equally, and this would result in a histogram with a flat line. An underdispersed forecast will result in a histogram with an $\cup$-shape because the ranks of the observations will often be very low or very high, while an overdispersed forecast will result in a histogram with an $\cap$-shape because the ranks of the observations will not often be very low or very high.

**Example 2.3**
We will calculate the rank of the observation for one day. We have

$$S = \begin{pmatrix} 4.1 & 3.8 & 4.1 & 8.6 & 5.5 \end{pmatrix}$$

The observation is $4.1$ and we have a corrected forecast of $4$ members. Then we have that $rank(y) = 3$, because the first three values are all less or equal than the value of the observation, which was $4$.

## Multivariate calibration

For assessing the calibration of a multivariate forecast, we use a two-step method [30][31]. Let $S = \{y, x_1, ..., x_M\}$ denote a set of vectors, with $y$ the vector of observations at the different stations, and $x_1, ..., x_M$ the corresponding multivariate forecast for the $M$ ensemble members. All these vectors have a length of $d$, namely the number of stations we are considering.
We calculate the rank of $y$ in two steps:

- Apply a prerank function $\rho$ to calculate $\rho(u)$ for every $u \in S$
- Set the rank of the observation $y$ equal to the rank of $\rho(y)$ in $\{\rho(y), \rho(x_1), ..., \rho(x_M)\}$
  If there are ties between the values of $\rho(u)$, we rank these random.

We will discuss three different approaches to calculate the prerank function $\rho$. All three methods give different shapes of rank histograms with different interpretation of the multivariate forecast.

## Multivariate ranking

For multivariate ranking we have the prerank function

$$\rho(u) = \sum_{v \in S} \mathbb{1}\{v \preceq u\}$$

where $v \preceq u$ is only true when $v_i \leq u_i$ for all $i = 1, ..., d$.

**Example 2.4**

$$S = \begin{pmatrix} 4.1 & 3.8 & 4.1 & 8.6 & 5.5 \\ 6.6 & 5.5 & 5.6 & 6.7 & 5.9 \end{pmatrix}$$

We know from the first column that we have observations $y_1 = 4.1$ at the first station, and observation $y_2 = 6.6$ at the second station. The other four columns represent the multivariate forecast.
If we calculate the ranks of all vectors $u \in S$ we find $\rho(y) = 3$, $\rho(x_1) = 1$, $\rho(x_2) = 2$, $\rho(x_3) = 5$, $\rho(x_4) = 3$, so we have

$$s = (3, 1, 2, 5, 3)$$

This means that by solving the ties random, we find that $rank(y) = 3$ or $rank(y) = 4$.

## Average ranking

For average ranking we have the prerank function

$$\rho(u) = \frac{1}{d}\sum_{i=1}^{d} rank(u,i)$$

where $rank(u,i)$ is calculated by ranking all values $u \in S$ per row $i$, with $i = 1, ..., d$. If we find ties in a row, we use the average of the ranks as the rank for all ties.

**Example 2.5**

$$S = \begin{pmatrix} 4.1 & 3.8 & 4.1 & 8.6 & 5.5 \\ 6.6 & 5.5 & 5.6 & 6.7 & 5.9 \end{pmatrix}$$

If we rank all values per row we find

$$R = \begin{pmatrix} 2.5 & 1 & 2.5 & 5 & 4 \\ 4 & 1 & 2 & 5 & 3 \end{pmatrix}$$

By taking the average of these ranks per column we find

$$s = (3.25, 1, 2.25, 5, 3.5)$$

This means that we find that $rank(y) = 3$.

## Band-depth ranking

For band-depth ranking we have the prerank function

$$\rho(u) = \frac{1}{d}\sum_{i=1}^{d}((K+1) - rank(u,i))(rank(u,i) - 1)$$

where $rank(u,i)$ is again calculated by ranking all values $u \in S$ per row $i$, with $i = 1, ..., d$. If we find ties in a row, we use the average of the ranks as the rank for all ties.

When this method for ranking is used, we see that values in the center get a high $\rho$, while values that are low or high get a very low $\rho$. This is very different from the multivariate and the average ranking, where low values have a low $\rho$ and high values have a high $\rho$.

**Example 2.6**

$$S = \begin{pmatrix} 4.1 & 3.8 & 4.1 & 8.6 & 5.5 \\ 6.6 & 5.5 & 5.6 & 6.7 & 5.9 \end{pmatrix}$$

If we rank all values per row we find

$$R = \begin{pmatrix} 2.5 & 1 & 2.5 & 5 & 4 \\ 4 & 1 & 2 & 5 & 3 \end{pmatrix}$$

By using the formula $(5 - x) \cdot (x - 1)$ for every $x \in R$ we find that

$$\tilde{R} = \begin{pmatrix} 3.75 & 0 & 3.75 & 0 & 3 \\ 3 & 0 & 3 & 0 & 4 \end{pmatrix}$$

By taking the average of these ranks per column we find

$$s = (3.375, 0, 3.375, 0, 3.5)$$

This means that by solving the ties random, we find that $rank(y) = 4$ or $rank(y) = 5$.

## Interpretation of the multivariate rank histograms

When the forecast is underdispersed or overdispersed, we can still recognize a $\cup$- or $\cap$-shape respectively when multivariate or average ranking is used. When the band-depth ranking is used, and the forecast is under- or overdispersed, we will get a triangular shape in the histogram. When the forecast is underdispersed, we will see low ranks occur very often, while high ranks barely occur. On the other hand, when the forecast is overdispersed, we see high ranks occur very often and low ranks barely occur.

In the multivariate case we can also say something about the correlation between stations. This can clearly be seen from the rank histograms when band-depth ranking is used [32][31]. Let us consider the case where we have fully dependent observations and an independent forecast across the different stations. By calculating the variance of the prerank for the observation curve and the variance of the forecast curve, we see that the latter is much smaller. Therefore, it is more likely to observe a very low or very high prerank when we have these dependent observations but independent forecast. This results in an $\cup$- shaped histogram. On the other hand, if we overestimate the correlation and we have an ensemble with too high correlations, we will see an $\cap$-shaped histogram.

## 2.3 Exploratory Data Analysis

In this work, we use the data provided by the Royal Netherlands Meteorological Institute (KNMI). For the days between April 1 and October 31 from 2011 until 2018 we have the 51 ensemble members of the European Centre for Medium-Range Weather Forecasts (ECMWF) that give predictions for the 2 meter temperature. Besides the raw forecast we also have the observations of the 2 meter temperature.
As already described in Section 1, the 51 members are initialised at 00:00 UTC and the forecasts are valid at 12:00 for the first 10 days. This means we look at 10 different lead times from 12 to 228 hours. We use forecasts at seven different stations throughout the Netherlands (Figure 2.1). The raw forecast with initial day 2011-08-28 is missing.

In Table 2.1 the seven different stations are listed. For each station we calculate the mean of the observed temperature over all days in the data set. We notice that the average temperature is lower at the coast (stations De Kooy (235) and Vlissingen (310)) and higher inland.

Table 2.1: *The seven station names and corresponding numbers. In the third column the mean of the observations of temperature is given. The fourth and fifth column give the corresponding Latitude and Longitude of the stations*

| Station Name | Station Number | Mean T2m | Latitude (°N) | Longitude (°E) |
|---|---|---|---|---|
| De Kooy | 235 | 16.35 | 52.93 | 4.78 |
| Schiphol | 240 | 17.61 | 52.32 | 4.78 |
| De Bilt | 260 | 17.86 | 52.10 | 5.18 |
| Eelde | 280 | 17.39 | 53.12 | 6.58 |
| Twenthe | 290 | 18.06 | 52.27 | 6.88 |
| Vlissingen | 310 | 17.07 | 51.45 | 3.60 |
| Maastricht | 380 | 18.23 | 50.90 | 5.77 |

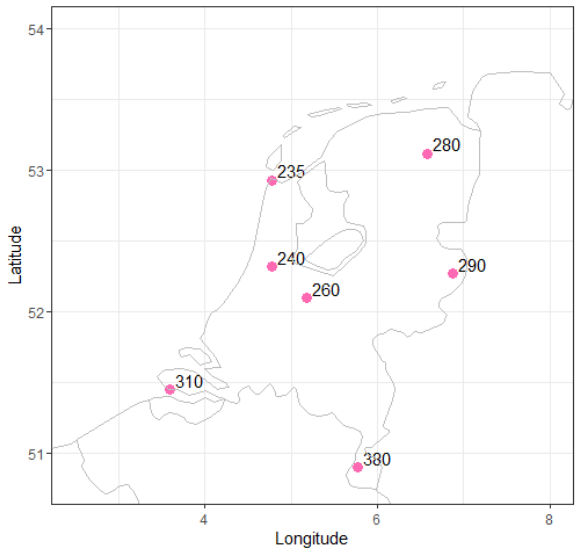


Figure 2.1: *The location and number of the seven stations in the Netherlands*

### 2.3.1 Correlation between stations

The correlation between the observations at the stations is given in the table below. The correlation will be of great importance in Section 4 where we investigate the ties in the raw forecasts.

From Table 2.2 we find that station Schiphol (240) and De Bilt (260) have the highest correlation, and stations De Kooy (235) and Maastricht (380) have the lowest correlation. This can be explained by the location of the stations: stations Schiphol (240) and De Bilt (260) are only approximately 50 km from each other, while stations De Kooy (235) and Maastricht (380) are approximately 250 km from each other. Besides this, stations Schiphol (240) and De Bilt (260) are both inland, while station De Kooy (235) is at the coast, and station Maastricht (380) inland.

Table 2.2: *Correlations between the observed temperature for every pair of stations*

| | **235** | **240** | **260** | **280** | **290** | **310** | **380** |
|---|---|---|---|---|---|---|---|
| **235** | 1 | 0.94 | 0.92 | 0.92 | 0.88 | 0.92 | 0.87 |
| **240** | | 1 | 0.98 | 0.94 | 0.93 | 0.93 | 0.92 |
| **260** | | | 1 | 0.94 | 0.95 | 0.92 | 0.94 |
| **280** | | | | 1 | 0.95 | 0.88 | 0.90 |
| **290** | | | | | 1 | 0.88 | 0.93 |
| **310** | | | | | | 1 | 0.90 |
| **380** | | | | | | | 1 |

### 2.3.2 Average annual cycle

We plot the average monthly temperature over the years for station De Bilt. We do this to compare the climatology in every month for different years and to make decisions on how many training days we need for EMOS. Results are in Figure 2.2. As can be expected, we see a clear trend that shows a higher average temperature in July compared to the average temperatures in April and October.

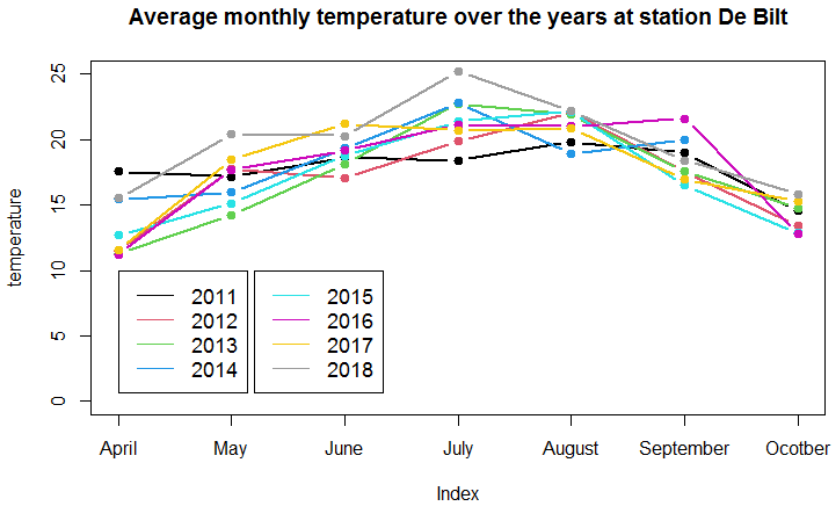


Figure 2.2: *Average monthly temperature over the years at station De Bilt*

### 2.3.3 Details about postprocessing

In this work we use $n = 50$ training days for EMOS. These 50 days are the previous 50 days in the data set. From Figure 2.2 we see that the climatology in April and October seem similar. Therefore, also for days in April the 50 previous days in the data set are used which gives training days in October and September.

We postprocess the forecasts from 2015-2018, and the data from May 21 2011-2014 is used to select historical observations in the Schaake Shuffle and SimSchaake method.
We start from May 21 instead of April 1 in 2011 because EMOS uses 50 training days and therefore no corrected forecast is available for the first 50 days of the data set. Because the adaptation of SimSchaake uses the corrected forecast instead of the raw forecast, the first 50 days cannot be selected as historical observations. Therefore, for comparison we also remove the first 50 days from the set of potential historical observations.

## 2.4 Summary

We summarize this Section to give an overview of the most important points that were discussed in Section 2.1, Section 2.2 and Section 2.3.

1. Methods
    - We use EMOS for univariate postprocessing. We minimize the CRPS over $n$ training days to estimate the parameters of the regression model.
    - We use ECC, Schaake Shuffle, SimSchaake and an adaptation of SimSchaake for multivariate postprocessing.
      ECC models dependencies based on the ranks in the raw forecast. The Schaake Shuffle and both versions of the SimSchaake model dependencies based on the ranks in historical observations.
    - The Schaake Shuffle uses historical observations that are randomly selected within a window around the target day. The SimSchaake selects historical observations based on similarities in the raw forecasts. The adapted SimSchaake selects historical observations based on similarities in the corrected forecasts.
2. Verification
    - We use scoring rules to quantify the skill of the forecasts. In the univariate case we use the CRPS, in the multivariate case we use the energy and variogram score.
    - We use rank histograms as visualisation tools to quantify the calibration of the forecasts.
    - For the multivariate forecasts we have 3 types of ranking: multivariate, average and band-depth ranking. From all three types of ranking we can see under- and overdispersion of the forecast. In average and band-depth ranking we can also draw conclusions about the correlation between stations.
3. Exploratory Data Analysis
    - We have 51 ensemble members of the ECMWF data that give predictions for the 2 meter temperature between April 1 2011 and October 31 2018 at seven different stations in the Netherlands.
      The 51 members are initialised at 00:00 UTC and the forecasts are valid at 12:00 for the first 10 days.
    - For EMOS we use $n = 50$ training days.
    - In the multivariate methods we forecast on the data from 2015-2018. The data from May 21 2011-2014 is used as historical observations for the Schaake Shuffle and SimSchaake.

# 3 Case Study

In this Section we apply the techniques discussed in this section.

We will first discuss the skill and calibration of the univariate forecasts. Thereafter we discuss the performance of all multivariate postprocessed forecasts.

## 3.1 Univariate Postprocessing

To look at how skillful EMOS is, we consider the CRPS of the raw forecasts and the corrected forecasts at every station. In Figure 3.1 we show the mean of the CRPS at stations De Kooy and Schiphol for different lead times.
We see that the corrected forecasts have a lower CRPS for every lead time. However, the CRPS of the forecasts at longer lead times is higher than the CRPS of the forecasts for short lead times. This shows that the forecast at longer lead times is not as skillful as the forecast for shorter lead times.

We can plot the continuous ranked probability skill score (CRPSS) for every station and lead time to show the skill of the corrected forecast relative to the raw forecast. We use the formula

$$CRPSS = 1 - \frac{CRPS_{corrected}}{CRPS_{raw}}$$

Because the CRPS of the corrected forecast is lower than the CRPS of the raw forecast, the CRPSS gives a score between 0 and 1. A higher score indicates that EMOS has more effect, while a lower score indicated that EMOS is less effective for that station and/or lead time.

The results are presented in Figure 3.2. These results are typical for all the stations.

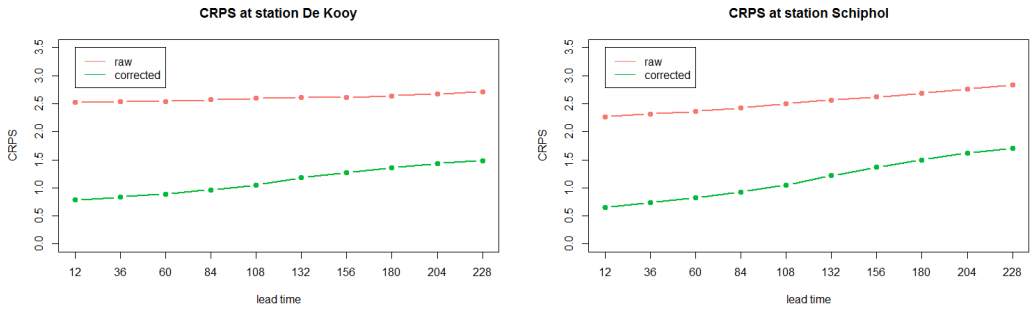


Figure 3.1: *CRPS for the raw forecast and CRPS for the corrected forecast at station De Kooy (235) and station Schiphol (240)*

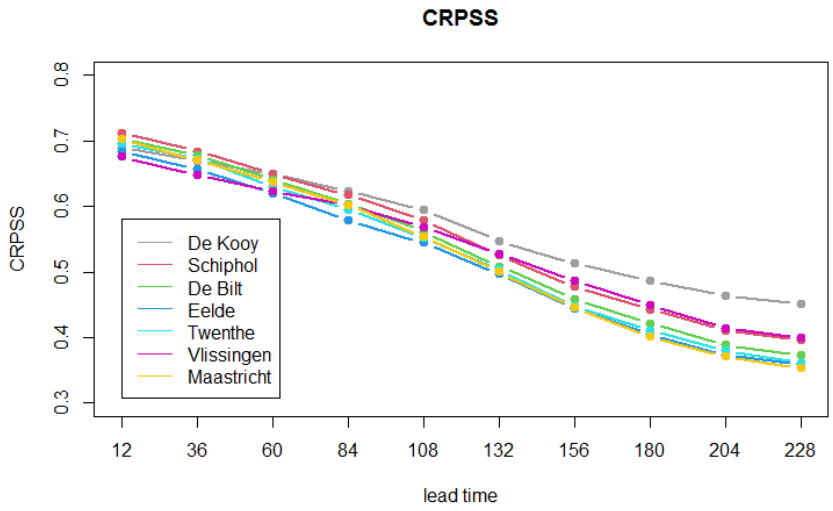


Figure 3.2: *CRPSS for every station and lead time*

We see that the CRPSS is higher for short lead times, which means EMOS improves more on the forecasts at short lead times. Also from this plot we see that EMOS does improve the forecast for longer lead times, but not as much as for short lead times as we already noticed in Figure 3.1.

We can look at rank histograms to see if the forecasts are well calibrated. In Section 2.2 we discussed that a forecast is calibrated if the ranks of the observations are uniformly distributed. The rank histogram should be flat. We compare the rank histogram of the corrected forecast to the rank histograms of the raw forecasts.
The rank histograms for the raw forecast and the corrected forecast give more or less the same shape for the seven different station and therefore, we can draw the same conclusions for every station. We show the rank histograms for the forecasts at station De Bilt.

As expected, for short lead times the observation of the temperature is often less than the lowest value of the raw forecast or higher than the highest value of the raw forecast. Therefore, we see a very underdispersed raw forecast for the shorter lead times.
If we look at the raw forecast for the longer lead times, we see that the histograms look more like a flat line. This indicated that the raw forecasts are well calibrated.

EMOS corrects for biases and dispersion errors in the raw forecast, so the range of the temperature forecast for short lead times becomes a bit wider. This has an influence on the rank histogram. In the histogram for the corrected forecast we see almost a flat line.
For large lead times we do not see that much of a difference between the histogram of the raw forecast and the histogram of the corrected forecast.

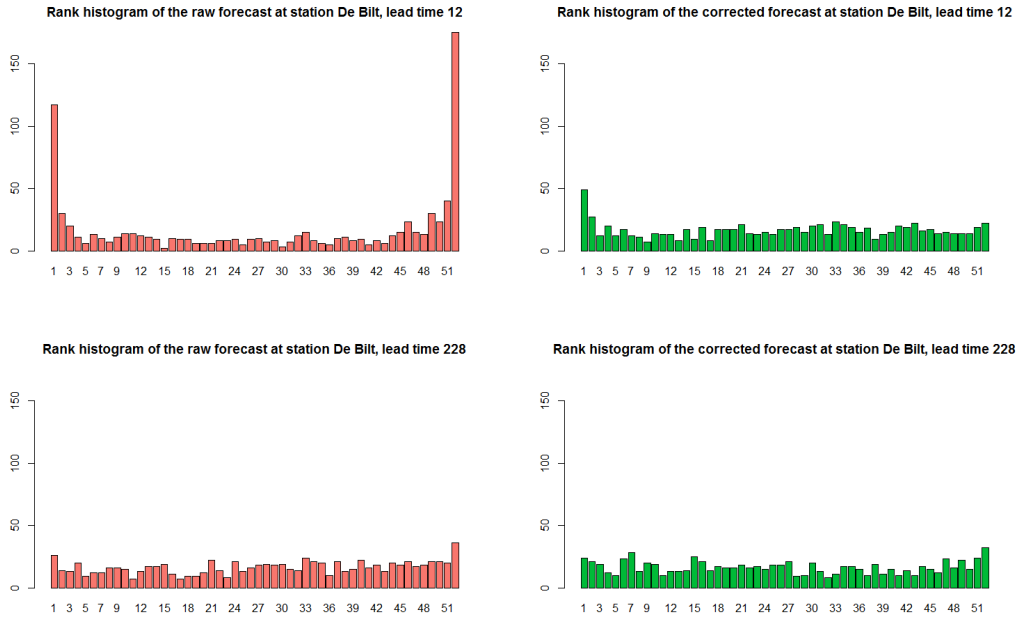


Figure 3.3: *Rank histograms for different lead times of the raw forecast for temperature on the left and rank histograms for different lead times of the corrected forecast for temperature on the right*

## 3.2 Multivariate Postprocessing

After removing most of the bias and dispersion errors with univariate postprocessing, we continue with the multivariate postprocessing to restore the dependence structure between stations. As described in Section 2.3 we postprocess forecasts from 2015-2018, and use the data from May 21 2011-2014 for historical observations in the Schaake Shuffle and SimSchaake methods.

We see the results for the energy score in Figure 3.4. As we can expect, forecasts reshuffled with ECC, Schaake Shuffle, SimSchaake and the adapted SimSchaake perform better than the raw forecast. Furthermore, we conclude that ECC, SimSchaake and adapted SimSchaake are the best methods for short lead times, although the differences in energy scores are very small. For longer lead times all energy scores are very close and all methods seem to perform almost equally well.

We can see a clearer difference between the variogram scores for the methods in Figure 3.5. Only the scores for the SimSchaake and adapted SimSchaake barely differ. We do see that for longer lead times the adapted SimSchaake performs slightly better than the SimSchaake. Again, ECC, SimSchaake and adapted SimSchaake seem the best methods for short lead times.

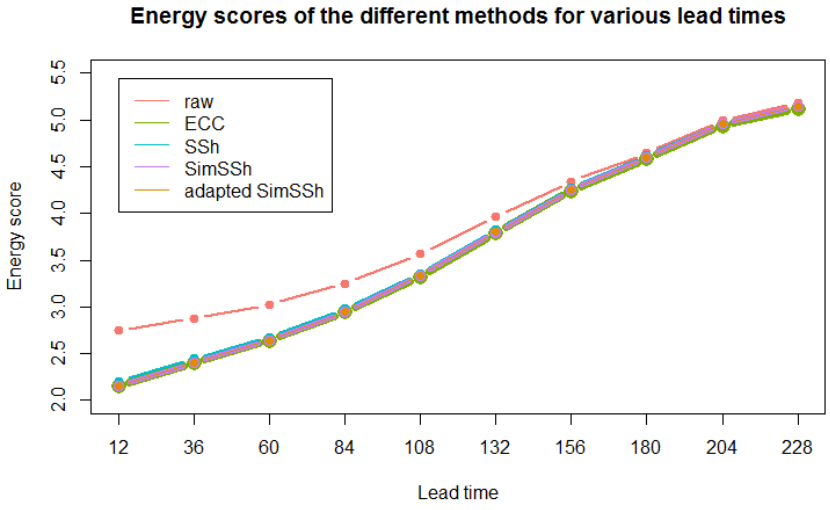


Figure 3.4: *Energy scores of the raw, ECC, Schaake Shuffle, SimSchaake and adapted SimSchaake forecasts*

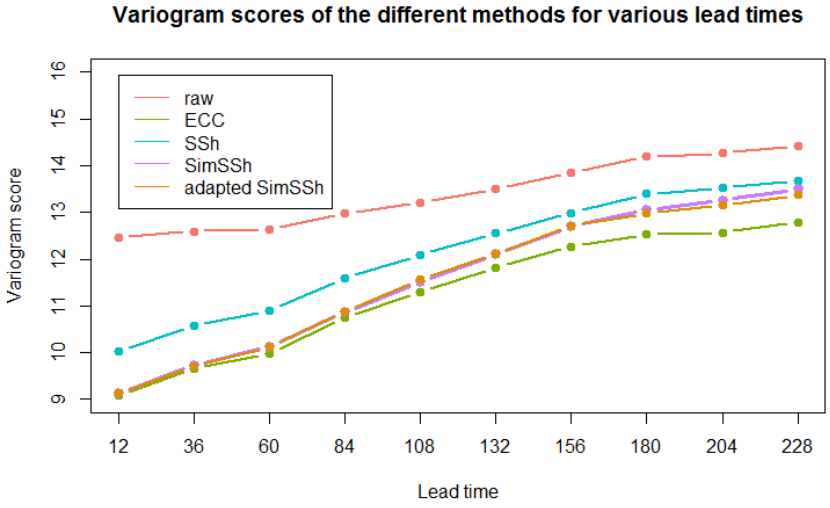


Figure 3.5: *Variogram scores of the raw, ECC, Schaake Shuffle, SimSchaake and adapted SimSchaake forecasts*

We see that the difference between the Schaake Shuffle and the SimSchaake (or adapted SimSchaake) decreases over time. For short lead times the SimSchaake performs better than the Schaake Shuffle. For longer lead times we see that the SimSchaake still performs a bit better, but the difference in variogram score has become quite small.
As explained in Section 2.1.2 the SimSchaake selects the historical observations based on similarity in the raw forecasts, while the Schaake Shuffle random selects historical observations with a same climatology. As discussed before, the raw forecasts for longer lead times are not that skillful and can differ a lot from the real observation. Because these forecasts for longer lead time are not that skillful, the SimSchaake performs equally well to the Schaake Shuffle.

Overall, ECC performs best.

Next, we look at the calibration of the forecasts. We will apply every multivariate verification method discussed in Section 2.2.2 and discuss the results.

### 3.2.1 Multivariate Ranking

First, we start comparing the rank histograms of the different methods when using the multivariate ranking explained in Section 2.2.2. We do this for lead times 12 and 228.

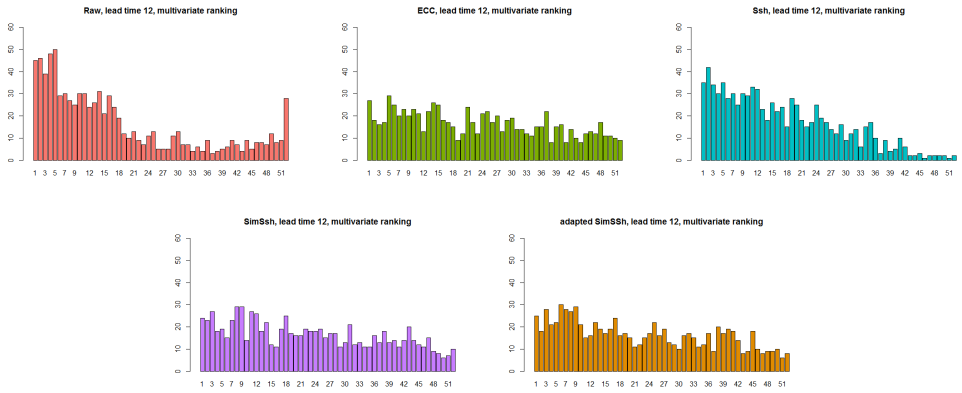


Figure 3.6: *Rank histograms for different methods at lead time 12 using multivariate ranking*

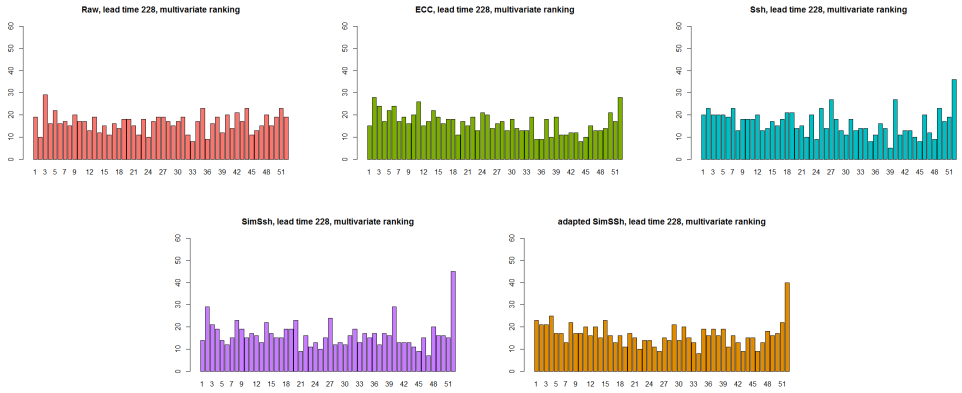


Figure 3.7: *Rank histograms for different methods at lead time 228 using multivariate ranking*

In Figure 3.6 we see that for short lead times not only the raw forecast but also the Schaake Shuffle forecast does not look well calibrated. ECC, the SimSchaake or the adapted Sim-Schaake look more like a flat histogram, which means the forecasts from these three methods are better calibrated than the raw and SimSchaake forecasts.
In the plots in Section 3.1 we saw that the histogram for the raw forecast at long lead times looked already quite flat, and not under- or overdispersed. Now in Figure 3.7 we also see that the histogram for the raw forecasts are flat.

### 3.2.2 Average Ranking

Next, we compare the rank histograms when using the average ranking explained in Section 2.2.2.

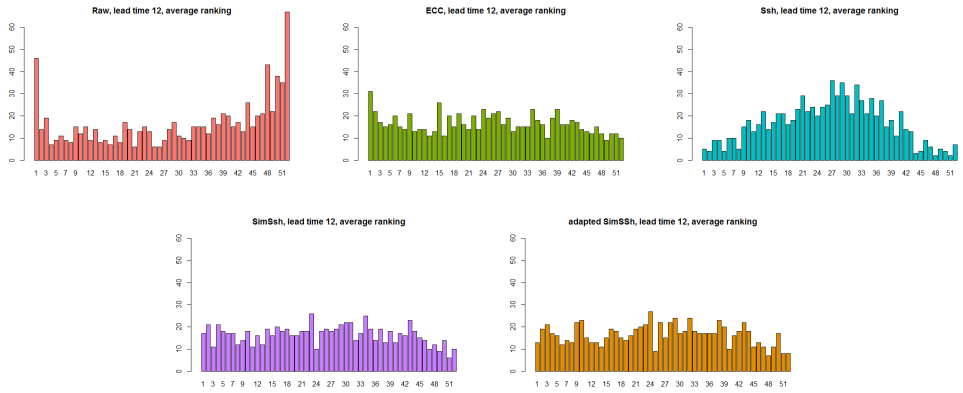


Figure 3.8: *Rank histograms for different methods at lead time 12 using average ranking*

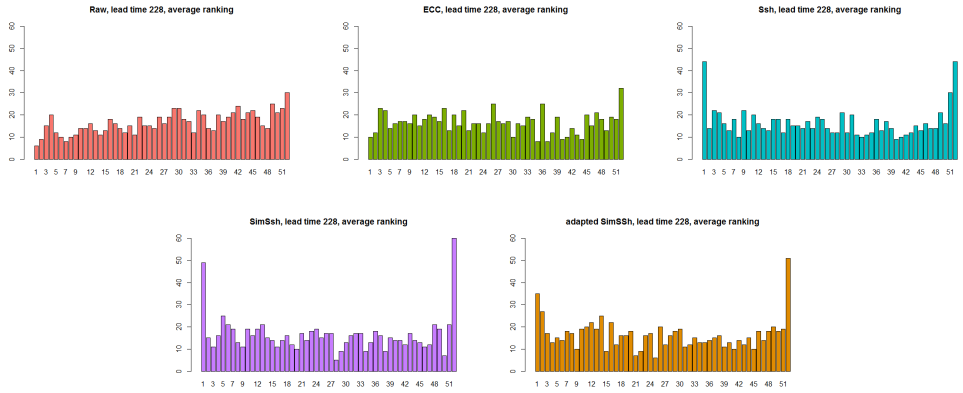


Figure 3.9: *Rank histograms for different methods at lead time 228 using average ranking*

Because of the ∩-shape we see in the histogram for the Schaake Shuffle in Figure 3.8, we know that the Schaake Shuffle uses historical observations that have a high correlation. This can be explained by the fact that we have days from May 21 2011-2014 to select 51 observations from, and because we select within a 15 day window from the day we are forecasting for, we have a high probability of selecting days that are very close and therefore have a high correlation.

Again, for short lead times we see in Figure 3.8 a ∪-shape in the histogram of the raw forecast while for longer lead times in Figure 3.9 the histogram of the raw forecast looks quite flat.

### 3.2.3 Band-depth Ranking

Last, we compare the rank histograms when using the band-depth ranking explained in Section 2.2.2. In these histograms we could clearly see under- and overdispersion by a triangular shape, and we can see the correlation from the ∩- or ∪-shape.

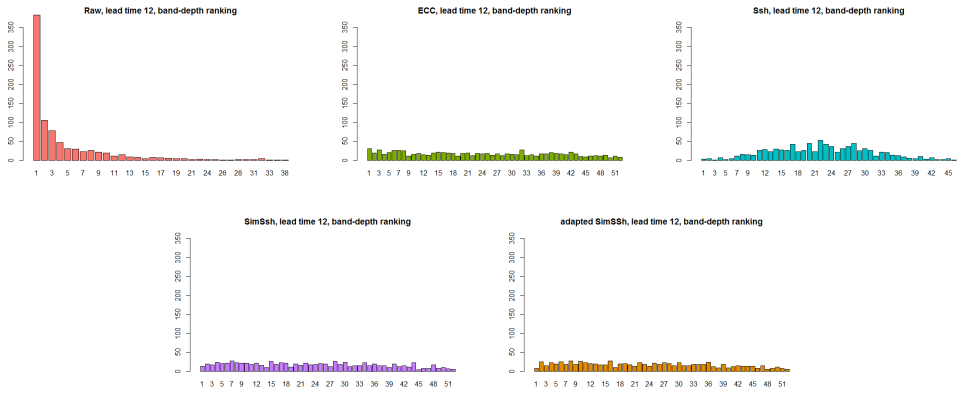


Figure 3.10: *Rank histograms for different methods at lead time 12 using band-depth ranking*

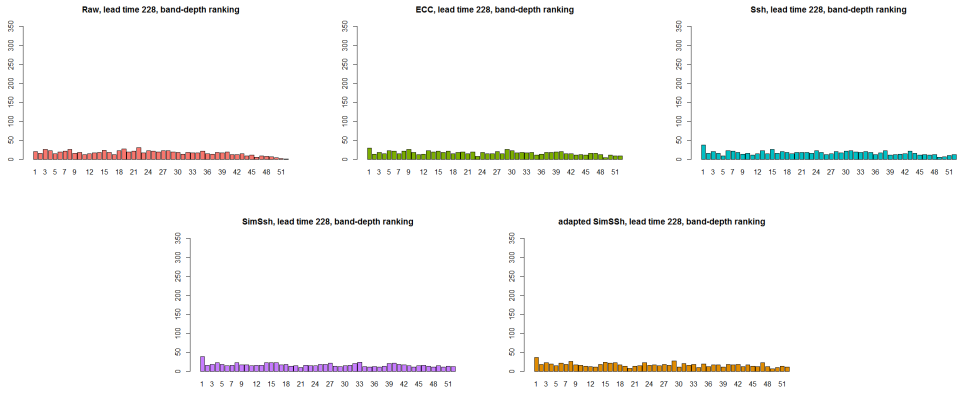


Figure 3.11: *Rank histograms for different methods at lead time 228 using multivariate ranking*

In Figure 3.10 we see a triangular shape in the histogram of the raw forecast. This shape shows that we have a great number of high and low values in the raw forecast, and barely any values in the center. This suggests that we have an underdispersed raw forecast for those lead times.
Again, the forecast for the Schaake shuffle shows a ∩-shape in Figure 3.10 as well as in Figure 3.11 which indicates the high correlation between the random selected historical observations.

In Table 3.1 we give an overview of the multivariate methods and their skill and calibration.

| Method | Skill | Calibration |
|---|---|---|
| **Raw** | The energy and variogram score show that the raw forecast does not perform well | For short lead times the raw forecast is extremely underdispersed. For longer lead times the forecast looks calibrated. |
| **ECC** | The energy and variogram score show that ECC performs best | The forecasts look calibrated for shorter and longer lead times. |
| **Schaake Shuffle** | The energy score shows that the method performs slightly worse than ECC, SimSchaake and the adapted SimSchaake. The variogram score shows larger differences between the Schaake Shuffle and ECC, SimSchaake and the adapted SimSchaake. For longer lead times the Schaake Shuffle performs almost equally well as the SimSchaake and adapted SimSchaake. | Multivariate ranking shows that the forecast is not calibrated for short lead times. Average and Band-depth ranking show that the selected historical observations are highly correlated. For larger lead times the forecasts look calibrated. |
| **SimSchaake** | The energy score shows that the SimSchaake performs very well. The variogram score shows that ECC slightly outperforms the SimSchaake. | The forecasts look calibrated for shorter and longer lead times. |
| **Adapted SimSchaake** | The energy score shows that the adapted SimSchaake performs very well. The variogram score shows that the adapted SimSchaake and SimSchaake perform equally well for short lead times. For larger lead times the adapted SimSchaake performs slightly better than the SimSchaake. | The forecasts look calibrated for shorter and longer lead times. |

Table 3.1: *An overview on the performance of the different methods for skill and calibration*

## 3.3 Summary

- We use the CRPSS to quantify the skill of the corrected forecast compared to the raw forecast. We conclude that EMOS improves the forecasts for every lead time and station, but performs best for short lead times.
- We see an underdispersed raw forecast for short lead times. We notice that this changes for the corrected forecast and the histogram looks more flat. Both the histogram for the raw forecast and the corrected forecast look calibrated for longer lead times.
- The Energy scores show that ECC and the SimSchaake methods perform best. The variogram scores show that ECC is the best method. The SimSchaake and the adapted SimSchaake perform equally well for short lead times. For longer lead times the adapted SimSchaake performs slightly better.
- The rank histograms give that ECC, SimSchaake and adapted SimSchaake have calibrated forecasts for short lead times. We notice that the selected days in the Schaake Shuffle have a high correlations and therefore the histograms are not flat.
  For longer lead times all methods seem to perform equally well in terms of calibration.

# 4 The effect of ties on rank-based multivariate postprocessing methods

When performing ECC, the Schaake Shuffle or the SimSchaake, ties in the raw forecast or in historical observations are traditionally solved randomly. This might not be an issue if there are hardly any ties in the raw forecasts. However, in our data ties do occur.
To give an idea on the number of ties in the raw forecasts we calculate per day, per station and per lead time the maximum number of ties. This is done by calculating how often every value occurs in the raw forecast, and then we take the maximum of these numbers. We do this for the days between April 1 2015 and October 31 2018, because all of these days are used as a dependence template for ECC. We represent the findings in Figure 4.1.

From Figure 4.1, we notice that there are more ties for short lead times. This is expected because the dispersion of the raw forecasts is smaller for short lead times. For lead time 12 the average of the maximum number of ties is between 7 and 8, with an outlier for station De Kooy. For the longest lead time we see that the maximum number of ties is 5 and there are even forecasts with 10 ties.

Since a large amount of ties occurs in the raw forecasts, it is interesting to look at various ways for solving them. First, we will introduce two other methods to solve ties. Second, we perform ECC with the random method for solving ties multiple times on the forecasts and we investigate the differences in the variogram scores.

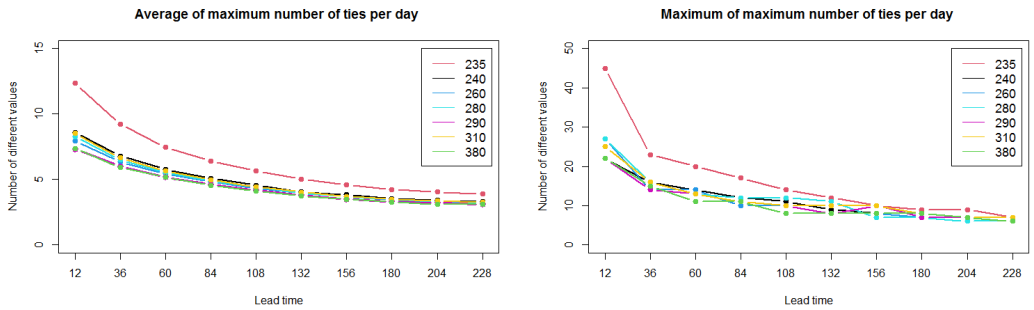


Figure 4.1: *Average and maximum number of maximum ties per day, per stations and per lead time*

## 4.1 Alternative methods for solving ties than random

Besides the random approach, there are multiple other methods that can be applied when assigning ranks to ties in the raw forecast or in historical observations. We discuss two methods that could be interested when solving ties in ECC, the Schaake Shuffle, or the SimSchaake.

### 'First' method

In this method called *first*, the ranks are assigned to ties in order from the first value to the last.

**Example 4.1**
If we have the raw ensemble

$$S = \begin{pmatrix} 4.2 & 3.6 & 4.2 & 8.8 & 5.0 & 4.2 \end{pmatrix}$$

$$R = \begin{pmatrix} * & 1 & * & 6 & 5 & * \end{pmatrix}$$

By assigning ranks using the method 'first' for ties, we get the ranks

$$R = \begin{pmatrix} 2 & 1 & 3 & 6 & 5 & 4 \end{pmatrix}$$

### 'Last' method

In this method called *last*, the ranks are assigned to ties in order from the last value to the first.

**Example 4.2**
If we have the raw ensemble

$$S = \begin{pmatrix} 4.2 & 3.6 & 4.2 & 8.8 & 5.0 & 4.2 \end{pmatrix}$$

$$R = \begin{pmatrix} * & 1 & * & 6 & 5 & * \end{pmatrix}$$

By assigning ranks using the method 'last' for ties, we get the ranks

$$R = \begin{pmatrix} 4 & 1 & 3 & 6 & 5 & 2 \end{pmatrix}$$

## 4.2 Impact on the variogram scores of the random method

To select the best method to solve ties, we could reshuffle forecasts with ECC using the 'first' method, random method and 'last' method and then compute the variogram score per day. At that point we could compare these scores and conclude that the method that gives the lowest score is the best for that specific day.
The problem with this approach is that the random method will give a different score at every run. Therefore, it would be interesting to study the variability in the random score and to compare different methods to the random method. This aspect has not been analyzed in the literature.

To analyze this point, we focus on two groups of two stations each:

- We look at the difference in random methods for station De Kooy and Maastricht. As can be seen in Table 2.2, these stations have the lowest correlation in the raw forecast.
- We look at the difference in random methods for station Schiphol and De Bilt. As can be seen in Table 2.2, these station have the highest correlation in the raw forecast.

We run ECC 100 times for all forecasts between 2015 and 2018 for lead time 12 so we have 100 variogram scores per day. In Figure 4.2 the mean of the variogram score per day is given for all days between 2015-2018 for stations De Kooy and Maastricht and for stations Schiphol and De Bilt. Almost all variogram scores are between 0 and 1. The difference between the variogram scores is mostly between 0 and 0.1.

To investigate the source of the differences we distinguish days with a small difference in variogram score and days with a large difference in variogram score. We set a threshold at 0.025 and 0.1. Days where the difference between the minimum and maximum value of the variogram score is larger than 0.1 are considered days with a large difference, while days where the difference between the minimum and maximum value of the variogram score is smaller than 0.025 are considered days with a small difference. This gives the following groups:

- Group 1 consists of all days with a difference in variogram score below 0.025
- Group 2 consists of all days with a difference in variogram score between 0.025 and 0.1
- Group 3 consists of all days with a difference in variogram score above 0.1

For stations De Kooy and Maastricht we have 546 days in group 1 and 44 days in group 3. For stations Schiphol and De Bilt we have 411 days in group 1 and 61 days in group 3.
If we compare these number of days in the groups, we see that the variation in variogram scores is larger for stations with a higher correlation because stations Schiphol and De Bilt have more days in group 3 and less days in group 1.

We now compare the days in group 1 to the days in group 3 and see what the cause of a large or low difference is. We also investigate the correlation between the raw forecasts, the number of ties in the raw forecasts and the location of the ties.

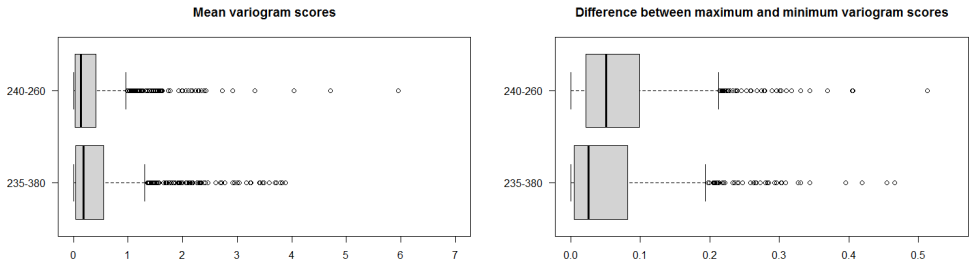


Figure 4.2: *Mean variogram scores per day for stations De Kooy (235) and Maastricht (380) on the left and for stations Schiphol (240) and De Bilt (260) on the right*

### 4.2.1 Correlation between the raw forecasts

We now look at the Spearman's correlation between the raw forecasts. If the Spearman's correlation is high for the raw forecasts, we know that the order of the assigned ranks to every value is very similar for both the forecasts. If there are ties in the raw forecasts and the raw forecasts also have a high correlation, we could expect a method as 'first' or 'last' would perform better than a random method. For both pairs of stations we see the results in Figure 4.3.
We see in both subfigures that the correlation between the raw forecasts for all days in group 1 is equal to the correlation between the raw forecasts for all the days in group 3.

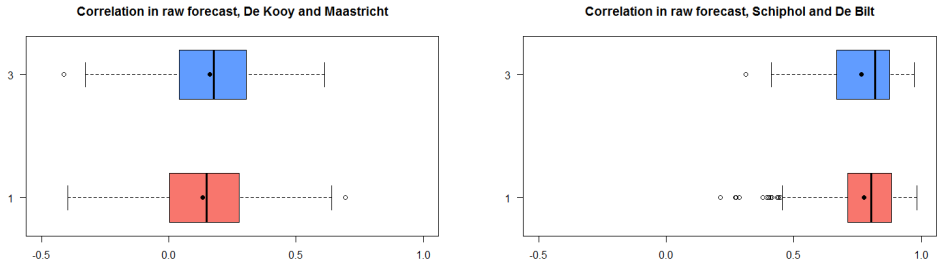


Figure 4.3: *Correlation between the raw forecasts for the days with in group 1 and 3 for stations De Kooy (235) and Maastricht (380) on the left and for stations Schiphol (240) and De Bilt (260) on the right*

### 4.2.2 Unique values in the raw forecast

When multiple ties occur in the raw forecast, there will be more possibilities to shuffle the values from the univariate corrected forecast. Therefore, we could expect a high variation in the variogram scores. The number of unique values in the raw forecast for the days in group 1 and group 3 can be seen in Figure 4.4.
We see that the number of unique values at station De Kooy is very low and almost the same for the days in group 1 and 3. At station Maastricht we see that the days in group 3 have more unique values in the raw forecast which means there are less ties. This is surprising, because we expected the days in group 3 to have more ties, as there are more ties in the raw forecast lead to more possibilities to reorder the corrected forecast. We will investigate this further in the simulation study in Section 5.

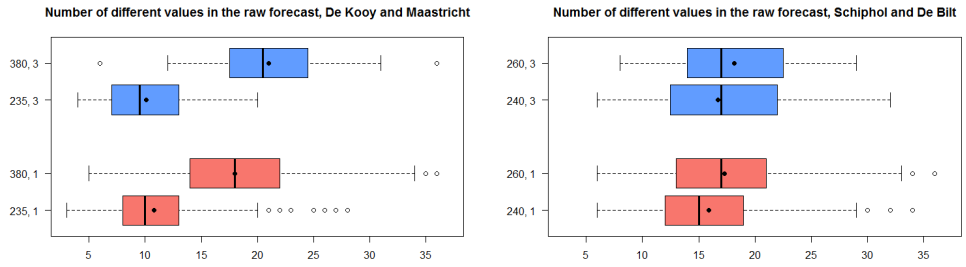


Figure 4.4: *Number of unique values in the raw forecasts for the days with in group 1 and 3 for stations De Kooy (235) and Maastricht (380) on the left and for stations Schiphol (240) and De Bilt (260) on the right*

### 4.2.3 Location of the Ties

Besides the number of ties in the raw forecast, we investigate if the location of these ties is of great influence. When the ties occur for the same ensemble member, we can expect larger differences in the random method.
For every day, we count the maximum value of how many ensemble members have the same forecast value at station 235 and 380. For example, the raw forecast

$$S_{raw} = \begin{pmatrix} 9.0 & 8.0 & 9.0 & 8.0 & 8.0 & 9.0 \\ 4.0 & 3.0 & 4.0 & 5.0 & 5.0 & 4.0 \end{pmatrix}$$

has a value of three, because there are three ensemble members that have the same value at both the stations.

We plot the number of matches in the raw forecasts for the days in group 1 and group 3. Results can be seen in Figure 4.5.
We see that all histograms show the same shape, and we see that 2 or 3 maximum pairs are most common. In Section 5 we will consider raw forecasts with more pairs and see if this has an influence on the performance of the random method.

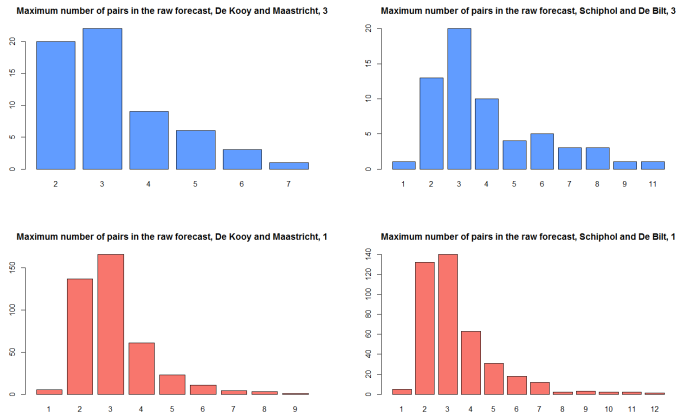


Figure 4.5: *Maximum number of matches in the raw forecasts for the days in group 1 at stations De Kooy (235) and Maastricht (380), the days in group 3 at stations De Kooy (235) and Maastricht (380), the days in group 1 at stations Schiphol (240) and De Bilt (260) and the days in group 3 at stations Schiphol (240) and De Bilt (260)*

## 4.3 Summary

- We here consider two other possible approaches to solve ties, namely:
    1. ties solving by assigning tied ranks in order from the first value to the last
    2. ties solving by assigning tied ranks in order from the last value to the first
- We investigate the influence of ties for two pairs of two stations each: De Kooy and Maastricht, and Schiphol and De Bilt.
- We investigate the importance of three different related aspects: the correlation between the raw forecasts, the unique values in the raw forecast and the location of the ties.
- We do not observe a difference in any of the aspects between the days with a small difference in variogram score and the days with a large difference in variogram score.

# 5 Simulation Study

We perform a simulation study to further investigate the differences in the variogram scores. We investigate the same conditions as in the case study with ties in Section 4, namely the number of unique values in the raw forecasts, the locations of the ties and the correlation between the raw forecasts.

In this simulation we consider 51-member forecasts at two stations. We simulate data similar to the data described in Section 2.3 and used in Section 3. This means we have 8 years of data from April 1 until October 31, and we make forecasts for the last 856 days. We simulate the observations and raw forecasts based on the observations and raw forecasts for a lead time of 12 hours.

We consider the following settings:

- In setting 1 (random ties), we do not consider a specific amount of ties.
- In setting 2 (matched ties), we ensure that there are 25 members with ties and we restrict them to be at the first 25 ensemble members for both stations.
- In setting 3 (unmatched ties), we ensure that there are 25 members with ties and we restrict them to be at the first 25 ensemble members for the first station, and at the last 25 ensemble members for the second station. Therefore, there will be no ties at the same ensemble member for both stations.

We investigate if the number of ties influences the variability of the of the variogram scores, i.e., the difference between the maximum and minimum variogram score, for 100 runs of ECC with ties resolved at random. By comparing setting 2 to setting 3 we investigate the affect of the locations of the ties ion the variability of the variogram score.
In every setting we consider the correlations between the raw forecasts around 0.1, 0.5 and 0.9.

## 5.1 Simulating the different settings

In this Section we explain how we generate the observations and the raw forecasts for the three different settings.

We start with simulating the observations. There are multiple ways to do this. We could have used the observations from the data used in the case study. We decided to not take the exact same observations but simulate them as follows.

Every year in the data consists of the days between April 1 and October 31, which gives a total of 214 days. The temperature shows an increasing trend from 13.3 to 21 degrees between April and July and a decreasing trend from 21 to 13.3 degrees between July and October. To simulate this, we take 107 quantiles of the normal distributions with mean 16 and standard deviation 4 to have the observations of the first half of one year. Then we take 107 quantiles in decreasing order of the normal distributions with mean 16 and standard deviation 4 to have the observations of the second half of one year.

We do not want to have the observations as perfect quantiles, and because we do not always have the same observation at both of the stations, we add random noise to the quantiles. We do this by generating a random number from the normal distribution with the mean equal to the quantile value, and the standard deviation equal to 0.5.
The pseudocode can be found in Algorithm 1.

We compare the observations from the data in the case study to the simulated observations. The results are reported in Figure 5.1. The monthly climatologies of the simulated and observed temperatures match reasonably well. Only the months May and August give slightly higher temperatures in the real data.

### Setting 1

In setting 1 ties can occur, but we do not make restriction on the number and location of ties. First we generate the mean of the raw forecast per station. We do this in the exact same way as simulating the observations in Algorithm 1.

**Algorithm 1:** Simulating observations

```
for all 8 years do
    for first 107 days in a year do
        take quantile of the normal distribution with mean 16 and standard deviation 4
    end
    for last 107 days in a year do
        take quantile of the normal distribution with mean 16 and standard deviation 4 in
         decreasing order
    end
end
for all days do
    simulate the observation as a random number with mean equal to the quantile value
     and standard deviation 4
end
```

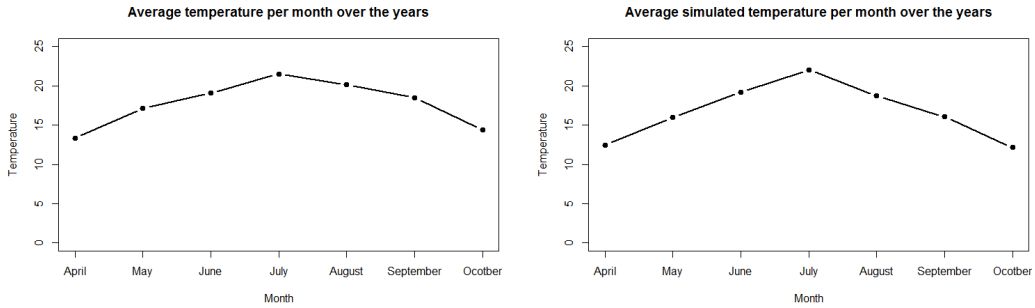


Figure 5.1: *Average temperature per month over the years on the left and average simulated temperature per month over the years on the right*

In Figure 5.2 all standard deviations of the raw forecast with lead time 12 are given. We find a mean standard deviation of 0.456 and from the boxplots we see that except for some outliers, almost all values are between 0.05 and 1.0. The standard deviation of all standard deviations is equal to 0.223.

We simulate the standard deviation of the raw forecasts by taking a random positive number from the normal distribution with mean 0.45 and standard deviation 0.2. If we have the mean and standard deviation per station we can take 51 values from the normal distribution which represents the raw forecast. By using the function '*mvrnorm*' in R we restrict the Pearson correlation between the two vectors to be around 0.1, 0.5 and 0.9. We repeat this for every day.

The function '*mvrnorm*' in R ensures the Pearson's correlation between vectors and not the Spearman's correlation. The Spearman's correlation is the correlation that should be used when applying rank-based methods. However, the Pearson's and Spearman's correlation are very close, and because we consider correlations close to 0.1, 0.5 and 0.9 we can use the Pearson's correlation.

In Algorithm 2 we described the simulation when the correlation between the raw forecasts is around 0.1.

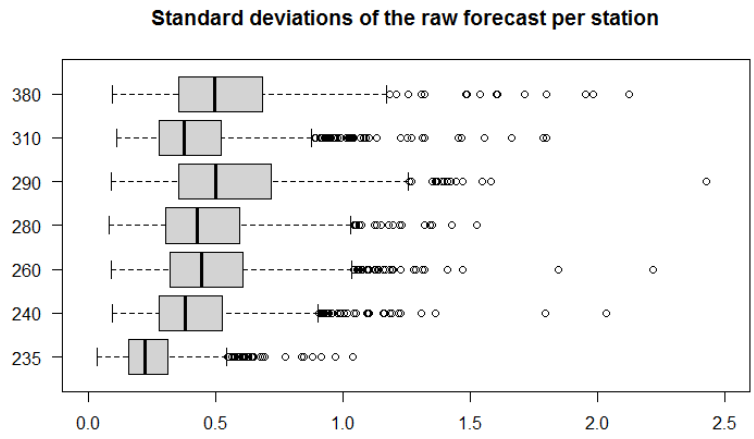


Figure 5.2: *Standard deviations of the raw forecast per station*

## Setting 2

In setting 2 we consider 25 ties that are located at the first 25 members for both station.
In both simulated vectors we change the second until 25th value to the first value, so we have 25 ties at the first 25 places for both vectors. We again repeat this for every day.
In Algorithm 3 we described the simulation when the correlation between the raw forecasts is around 0.1.

## Setting 3

In setting 3 we consider 25 ties that are located at the first 25 places for the first station and at the last 25 places for the second station.
In the first simulated vector we change the second until 25th value to the first value. In the second vector we change the 27th until 50th value to the 51th value.
In Algorithm 4 we described the simulation when the correlation between the raw forecasts is around 0.1.

**Algorithm 2:** Simulating setting 1: no ties

```
for all days do
    Generate the mean per station
    Generate the standard deviation per station
    Generate two vectors of length 51 with the generated mean and standard deviation
     with correlation 0.1
end
```

**Algorithm 3:** Simulating setting 1: matched ties

```
for all days do
    while correlation is between 0.05 and 0.15 do
        Generate the mean per station
        Generate the standard deviation per station
        Generate two vectors of length 51 with the generated mean and standard
         deviation with correlation 0.1
        for index 2 until 25 do
            change value equal to first ensemble member in both vectors
        end
        calculate correlation between both vectors
    end
end
```

**Algorithm 4:** Simulating setting 1: unmatched ties

```
for all days do
    while correlation is between 0.05 and 0.15 do
        Generate the mean per station
        Generate the standard deviation per station
        Generate two vectors of length 51 with the generated mean and standard
         deviation with correlation 0.1
        for index 2 until 25 do
            change value equal to first ensemble member in first vectors
        end
        for index 27 until 50 do
            change value equal to last ensemble member in second vectors
        end
        calculate correlation between both vectors
    end
end
```

## 5.2 Results

For every method we first perform EMOS and then apply ECC on the days in the years 2015-2018, where we solve the ties in the raw forecast randomly. We do this 100 times.

In Figure 5.3 we have the mean variogram scores for every setting and correlation. Most days have a mean variogram score close to zero.

For all the days in all settings we compare the difference between the maximum and the minimum variogram score for the 100 runs. Results are in Figure 5.4. We notice that the difference in variogram scores is small for setting 1. The differences increase for setting 2, but the differences in variogram scores are the largest in setting 3. This shows that more ties in the raw forecast result in a larger difference in variogram score and that ties located at the same ensemble member do not result in larger difference in the random score.

Next, we investigate the performance of the 'first' method for resolving ties compared to the random method. We start with looking at the difference between the minimum of all 100 variogram scores for the random method and the variogram score when we use the 'first' method. We subtract the variogram score for the 'first' method from the minimum variogram score of the random method for every day. If this difference is greater that zero, we have that the 'first' method performs better than the best case scenario of the random method, because we were considering the minimum value of the 100 runs with the random method.
In Figure 5.5 the results for this difference is given.

While the differences are very small for the days with a correlation around 0.1 or 0.5 in the raw forecast, we can observe that the differences between the methods are larger when the correlation in the raw forecast is higher for all three settings. If we compare setting 2 to setting 3 in Figure 5.5, we notice a large positive difference for quite a number of days in the second setting. This means that the 'first' method performs better when ties are matched for the same ensemble members.

Still a large number of days has a negative value for the difference, which means the 'first' method does not perform better than the best option for the random methods. We look further into this by also comparing the 'first' method to the maximum value of the 100 variogram scores for the random method. We subtract the variogram score for the 'first' method from the maximum variogram score when performing the random method. Results of this difference are in Figure 5.6. The differences might not be that large for the days with a correlation in the raw forecast around 0.1 or 0.5, but over 50$\%$ of the days have a positive difference, which tells us that the 'first' method performs better than the worst case scenario of the random method.

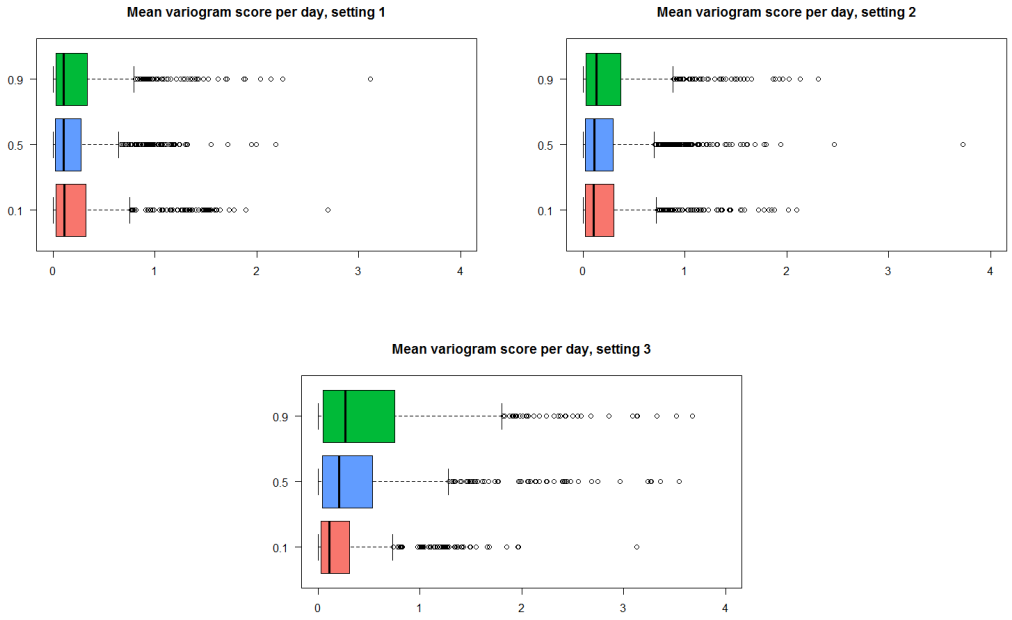


Figure 5.3: *Mean variogram score for the random method for the three different settings considering three different correlations in the raw forecast*

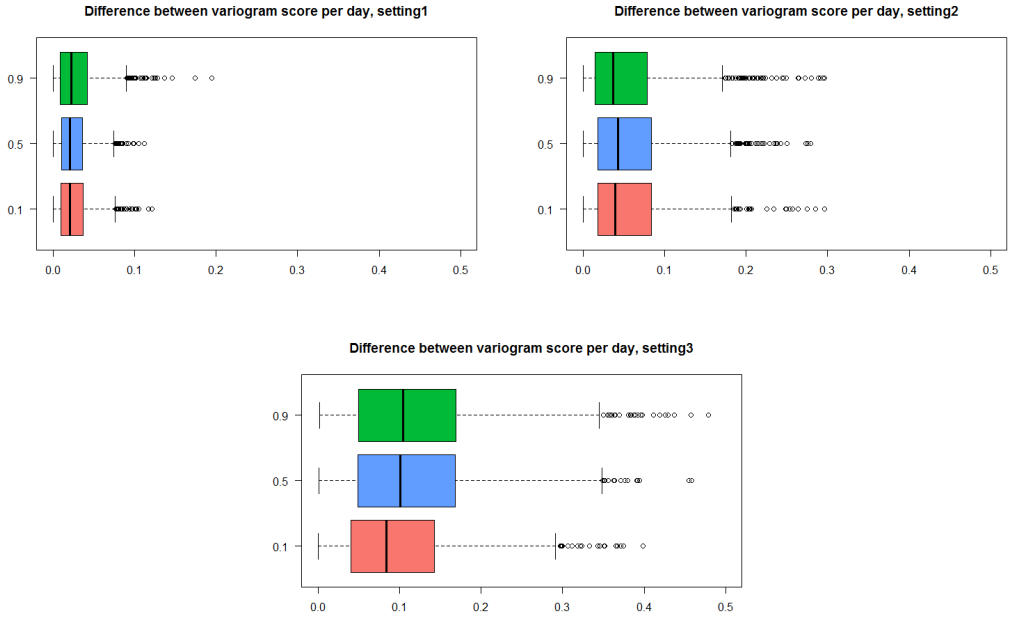


Figure 5.4: *Differences in minimum and maximum variogram score for the random method for the three different settings considering three different correlations in the raw forecast*

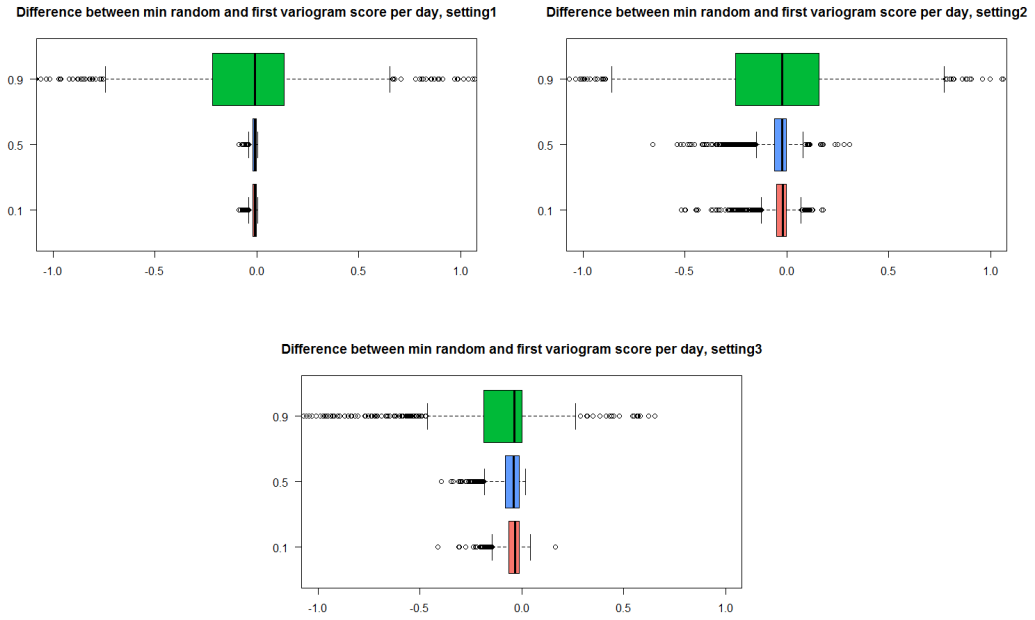


Figure 5.5: *Differences in variogram score for the 'first' method and the minimum variogram score for the random method for the three different settings considering three different correlations in the raw forecast*

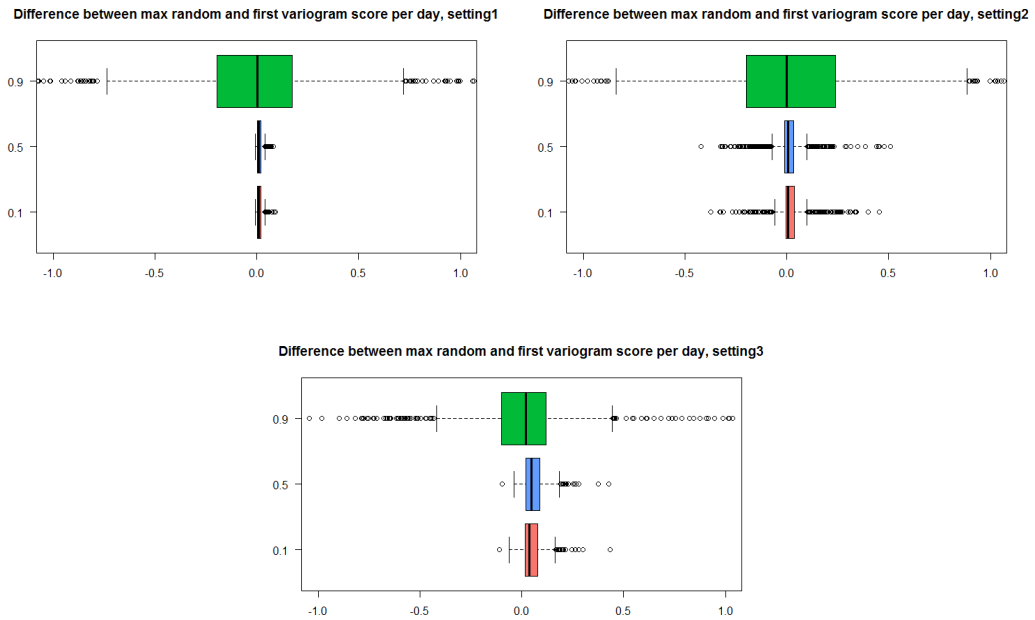


Figure 5.6: *Differences in variogram score for the 'first' method and the maximum variogram score for the random method for the three different settings considering three different correlations in the raw forecast*

## 5.3 Summary

- In this simulation study we further investigate the variability of the variogram scores in case of ties and we look at the performance of other ways of solving ties, compared to the random method. We consider three settings and investigate if the number of ties and their locations have an influence on the variogram scores:

  1. In setting 1 (random ties), we do not consider a specific amount of ties.
  2. In setting 2 (matched ties), we ensure that there are 25 members with ties and we restrict them to be at the first 25 ensemble members for both stations.
  3. In setting 3 (unmatched ties), we ensure that there are 25 members with ties and we restrict them to be at the first 25 ensemble members for the first station, and at the last 25 ensemble members for the second station.

  For every setting we look at raw forecasts with a correlation around 0.1, 0.5 and 0.9.

- We simulate observations and raw forecasts similar to the observations and forecasts of lead time 12 in the ECMWF data discussed in Section 2.3.
- After 100 runs, the difference in variogram scores is small for setting 1, it increases for setting 2, and it is the largest in setting 3. This shows that more ties in the raw forecast result in a larger variability in the variogram scores, while ties located at the same ensemble members do not result in larger difference in the variogram score.
- We compare the variogram scores after performing the 'first' method to the minimum and maximum variogram score of the 100 random runs separately. We see that for days with a correlation around 0.9 in the raw forecast the difference between the minimum variogram score of the random method and the 'first' method is larger than zero, which tells us that the 'first' method performs better than the best random run.
- However, when we look at the results for setting 2, we often see positive differences. This means that very often the 'first' method performs better than the random method when there are many ties and they are located at the same ensemble members.

# 6 Conclusion and Discussion

In this work we discussed several methods for postprocessing multivariate weather forecasts and investigated their limitations. For univariate postprocessing we used EMOS, and for multivariate postprocessing we compared the methods ECC, Schaake Shuffle, SimSchaake and an adapted SimSchaake where we used similarities in the corrected forecast instead of the raw forecasts to select historical observations. In Section 1 we gave five research questions. We will answer these questions by giving the main findings.

*How does EMOS perform for the different stations and lead times?*

The CRPS, CRPSS and the rank histograms show that EMOS improves the skill of the forecasts for all lead times. However, EMOS performs better for short lead times compared to longer lead times.

*How do the multivariate postprocessing methods ECC, the Schaake Shuffle and the SimSchaake perform for the different lead times?*

The energy score, variogram score and the rank histograms from the multivariate, average and band-depth ranking show that ECC gives the best forecast for all different lead times. The SimSchaake performs equally well as ECC for short lead times, but compared to ECC the performance decreases when the lead time increases. This is because the SimSchaake uses the raw forecasts to select historical observations, and the raw forecasts are not very skillful for longer lead times.
The Schaake Shuffle does not perform well. This can be due to the selection of historical observations that are too highly correlated and do not display the dependence structure between the stations well.
Our results were expected, as in Section 1 we already discussed that ECMWF is very capable of simulating the dependencies between stations and variables, resulting in a good performance of ECC.

*How does a multivariate method that selects historical observations based on similarities in the corrected forecasts perform for the different lead times?*

Wee see that the SimSchaake and SimSchaake perform equally well for short lead times, but the adapted SimSchaake outperforms the SimSchaake for longer lead times. Therefore, when the SimSchaake is performed, we highly recommend to also perform the adapted SimSchaake using the corrected forecast to improve the forecasts for longer lead times.

For further research, the relative new method called Total divergence is interesting. In this method the historical observations are selected such that the marginal distributions of the observation trajectories resemble those of the corrected forecast [33]. Therefore, raw forecasts are not necessary anymore for the selection of historical observations. Total divergence was invented for precipitation forecasts, but can be adapted for temperature forecasts. We did initial test, but computational future work can explore this method.

When we apply one of the multivariate methods and we shuffle the corrected forecast, ties

in the dependence template are typically solved randomly. We investigated the influence of this random approach on raw forecasts with ties. The results will especially be interesting for forecasts for short lead time, as these forecasts have a small dispersion compared to forecasts for longer lead time, and therefore are more likely to contain more ties.

We run ECC 100 times for the ECMWF forecasts with lead time 12 and compute in every run the variogram score. This results in 100 variogram scores per day. We looked at the variability within the variogram score per day, and we considered days with a large difference in variogram score, and days with a small difference in variogram score. Comparing these two groups by looking at the correlation between the raw forecasts, the number of ties and the location of ties did not give any clear indication about the source of the variability in variogram score. Therefore, we performed a simulation study to investigate further the three aspects correlation, number of ties and location of ties.

*What is the effect of ties on the performance of the multivariate postprocessing methods?*

In the simulation study we see that raw forecasts with more ties result in larger differences in the random method if we run multiple times. When the ties occur at different ensemble members for both stations we get even larger differences then when the ties occur at the same ensemble members for both station.

*How do other methods for solving ties perform compared to the random method?*

In the simulation study we showed that the 'first' method very often performs better than the random methods. This is often the case when ties are at the same ensemble member for both stations. Also, when the raw forecasts are highly correlated the 'first' method seems to perform better. The latter is quite interesting, as not only the number of ties, but also the correlation between raw forecast can influence the variability of the random scores.

In the research for the ties we always compared variogram scores after performing ECC and we investigated the number of ties in the raw forecast. For the multivariate methods Schaake Shuffle, SimSchaake or adapted SimSchaake ties can occur in the selected historical observations. However, the ties in the Schaake Shuffle could be avoided by randomly selecting other historical observations if any ties start occurring. The ties in the SimSchaake or adapted SimSchaake are harder to avoid, but one could select the days with the most similarity between the raw or corrected forecasts that do not have ties with the other already selected historical observations. This can influence the performance in a negative way, the results are also relevant to the Schaake Shuffle or SimSchaake is performed and the selected historical observations contain ties.

This thesis provides a starting point for the research into the effect of ties in the raw forecasts on the multivariate correction methods. While our results show that the presence of ties can influence the performance of the methods, previous work on the topic does not take this limitation into account.

Further research on the effect of ties on the methods, the simulation study in Section 5 can be extended. In particular, we only considered two stations, which made it easier to investigate the correlations between the raw forecasts. When more stations are considered the correlation between the raw forecasts of pairs of stations might still play an important role.

We only considered a setting where no restrictions on the ties were made for both stations, and two setting where we had 25 ties for both stations. We did not consider a setting where we varied in the number of ties for both the stations in one setting. This option can be explored in further research. At last, when reshuffling the corrected forecast, we could even decide on a different method for solving ties per station. For example, when for a specific day many ties occur at the first station, we could apply the 'first' method for reshuffling, while for that

same day at the second station barely any ties occur, we could apply the random method for reshuffling.

In conclusion, we have shown the univariate and multivariate postprocessing of weather forecasts and introduced a new promising multivariate method. Besides this, we provided a starting point for the research on ties which can be further explored in future work.

# Bibliography

[1] Adrian E Raftery, Tilmann Gneiting, Fadoua Balabdaoui, and Michael Polakowski. Using bayesian model averaging to calibrate forecast ensembles. *Monthly weather review*, 133(5):1155–1174, 2005.

[2] Annette Möller, Alex Lenkoski, and Thordis L Thorarinsdottir. Multivariate probabilistic forecasting using ensemble bayesian model averaging and copulas. *Quarterly Journal of the Royal Meteorological Society*, 139(673):982–991, 2013.

[3] Tilmann Gneiting, Adrian E Raftery, Anton H Westveld III, and Tom Goldman. Calibrated probabilistic forecasting using ensemble model output statistics and minimum crps estimation. *Monthly Weather Review*, 133(5):1098–1118, 2005.

[4] Maxime Taillardat, Olivier Mestre, Michaël Zamo, and Philippe Naveau. Calibrated ensemble forecasts using quantile regression forests and ensemble model output statistics. *Monthly Weather Review*, 144(6):2375–2393, 2016.

[5] Kirien Whan and Maurice Schmeits. Comparing area probability forecasts of (extreme) local precipitation using parametric and machine learning statistical postprocessing methods. *Monthly Weather Review*, 146(11):3651–3673, 2018.

[6] Michael Scheuerer and Thomas M Hamill. Statistical postprocessing of ensemble precipitation forecasts by fitting censored, shifted gamma distributions. *Monthly Weather Review*, 143(11):4578–4596, 2015.

[7] Thordis L Thorarinsdottir and Tilmann Gneiting. Probabilistic forecasts of wind speed: Ensemble model output statistics by using heteroscedastic censored regression. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 173(2):371–388, 2010.

[8] Sándor Baran and Sebastian Lerch. Mixture emos model for calibrating ensemble forecasts of wind speed. *Environmetrics*, 27(2):116–130, 2016.

[9] John Schaake, Jean Pailleux, Jutta Thielen, Ray Arritt, Tom Hamill, Lifeng Luo, Eric Martin, Doug McCollor, and Florian Pappenberger. Summary of recommendations of the first workshop on postprocessing and downscaling atmospheric forecasts for hydrologic applications held at meteo-france, toulouse, france, 15–18 june 2009, 2010.

[10] Veronica J Berrocal, Adrian E Raftery, Tilmann Gneiting, and Richard C Steed. Probabilistic weather forecasting for winter road maintenance. *Journal of the American Statistical Association*, 105(490):522–537, 2010.

[11] Pierre Pinson and Robin Girard. Evaluating the quality of scenarios of short-term wind power generation. *Applied Energy*, 96:12–20, 2012.

[12] Pierre Pinson. Wind energy: Forecasting challenges for its operational management. *Statistical Science*, 28(4):564–585, 2013.

[13] Pierre Pinson and Jakob W Messner. Application of postprocessing for renewable energy. In *Statistical postprocessing of ensemble forecasts*, pages 241–266. Elsevier, 2018.

[14] Georgios Chaloulos and John Lygeros. Effect of wind correlation on aircraft conflict probability. *Journal of Guidance, Control, and Dynamics*, 30(6):1742–1752, 2007.

[15] Veronica J Berrocal, Adrian E Raftery, and Tilmann Gneiting. Combining spatial statistical and ensemble information in probabilistic weather forecasts. *Monthly Weather Review*, 135(4):1386–1402, 2007.

[16] Kira Feldmann, Michael Scheuerer, and Thordis L Thorarinsdottir. Spatial postprocessing of ensemble forecasts for temperature using nonhomogeneous gaussian regression. *Monthly Weather Review*, 143(3):955–971, 2015.

[17] Roman Schefzik, Thordis L Thorarinsdottir, Tilmann Gneiting, et al. Uncertainty quantification in complex simulation models using ensemble copula coupling. *Statistical science*, 28(4):616–640, 2013.

[18] Martyn Clark, Subhrendu Gangopadhyay, Lauren Hay, Balaji Rajagopalan, and Robert Wilby. The schaake shuffle: A method for reconstructing space–time variability in forecasted precipitation and temperature fields. *Journal of Hydrometeorology*, 5(1):243–262, 2004.

[19] Roman Schefzik. A similarity-based implementation of the schaake shuffle. *Monthly Weather Review*, 144(5):1909–1921, 2016.

[20] Daniel S Wilks. Multivariate ensemble model output statistics using empirical copulas. *Quarterly Journal of the Royal Meteorological Society*, 141(688):945–952, 2015.

[21] Kirien Whan, Jakob Zscheischler, Alexander I Jordan, and Johanna F Ziegel. Novel multivariate quantile mapping methods for ensemble post-processing of medium-range forecasts. *Weather and Climate Extremes*, 32:100310, 2021.

[22] Elisa Perrone, Irene Schicker, and Moritz N Lang. A case study of empirical copula methods for the statistical correction of forecasts of the aladin-laef system. *Meteorologische Zeitschrift*, pages 277–288, 2020.

[23] Tilmann Gneiting, Fadoua Balabdaoui, and Adrian E Raftery. Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2):243–268, 2007.

[24] Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378, 2007.

[25] Sam Allen, Gavin R Evans, Piers Buchanan, and Frank Kwasniok. Accounting for skew when post-processing mogreps-uk temperature forecast fields. *Monthly Weather Review*, 2021.

[26] Manuel Gebetsberger, Jakob W Messner, Georg J Mayr, and Achim Zeileis. Estimation methods for nonhomogeneous regression models: Minimum continuous ranked probability score versus maximum likelihood. *Monthly Weather Review*, 146(12):4323–4338, 2018.

[27] Michael Scheuerer and Thomas M Hamill. Variogram-based proper scoring rules for probabilistic forecasts of multivariate quantities. *Monthly Weather Review*, 143(4):1321–1334, 2015.

[28] Thomas Jung and Martin Leutbecher. Scale-dependent verification of ensemble forecasts. *Quarterly Journal of the Royal Meteorological Society: A journal of the atmospheric sciences, applied meteorology and physical oceanography*, 134(633):973–984, 2008.

[29] Alexander Jordan, Fabian Krüger, and Sebastian Lerch. Evaluating probabilistic forecasts with scoringrules. *arXiv preprint arXiv:1709.04743*, 2017.

[30] Tilmann Gneiting, Larissa I Stanberry, Eric P Grimit, Leonhard Held, and Nicholas A Johnson. Assessing probabilistic forecasts of multivariate quantities, with an application to ensemble predictions of surface winds. *Test*, 17(2):211, 2008.

[31] Thordis L Thorarinsdottir and Nina Schuhen. Verification: assessment of calibration and accuracy. In *Statistical postprocessing of ensemble forecasts*, pages 155–186. Elsevier, 2018.

[32] Thordis L Thorarinsdottir, Michael Scheuerer, and Christopher Heinz. Assessing the calibration of high-dimensional ensemble forecasts using rank histograms. *Journal of computational and graphical statistics*, 25(1):105–122, 2016.

[33] Michael Scheuerer, Thomas M Hamill, Brett Whitin, Minxue He, and Arthur Henkel. A method for preferential selection of dates in the s chaake shuffle approach to constructing spatiotemporal forecast fields of temperature and precipitation. *Water Resources Research*, 53(4):3029–3046, 2017.