DIAGNOSTIC TOOLS FOR FORECAST ENSEMBLES

by

NANA AMA APPIAA BAFFOE

Submitted in partial fulfillment of the requirements

For the degree of Masters of Science

Thesis Adviser: Dr. Jenný Brynjarsdóttir

Department of Mathematics, Applied Mathematics and Statistics

CASE WESTERN RESERVE UNIVERSITY

May, 2018

DIAGNOSTIC TOOLS FOR FORECAST ENSEMBLES

Case Western Reserve University
Case School of Graduate Studies

We hereby approve the thesis[1] of

NANA AMA APPIAA BAFFOE

for the degree of

Masters of Science

Dr. Jenný Brynjarsdóttir

Committee Chair, Adviser — Date
Department of Mathematics, Applied Mathematics and Statistics

Dr. Anirban Mondal

Committee Member — Date
Department of Mathematics, Applied Mathematics and Statistics

Dr. Minwoo Chae

Committee Member — Date
Department of Mathematics, Applied Mathematics and Statistics

March 28, 2018

[1]We certify that written approval has been obtained for any proprietary material contained therein.

Dedicated to my best friend AT
Thanks for being my support for all these years.

Table of Contents

List of Tables

List of Figures

Acknowledgements

First and foremost I would like to express my sincere gratitude to my adviser Dr. Jenny Brynjardottir under whose guidance and mentoring I have learned so much. I deeply appreciate your insights and advice both professionally and personally. You have been an integral contributor to this dissertation.

I would also to express my profound gratitude to Prof Daniela Calvetti, who virtually became my adopted mother in the Mathematics department. You have been a great influence on my life and great support to me through some very difficult times,and I am eternally grateful.

I would like to thank my committee members Dr. Anirban Mondal and Dr. Minwoo Chae who took time from their busy schedules to be a part of the committee.

I am also grateful for the time, efforts and support of Prof. Erkki Somersalo, Dr. Alethea Barbaro, Dr. Patricia Williamson and the entire faulty at Department, who supported and and advised me both professionally and personally. Thank you

I am also grateful for my adopted family here in Cleveland: Amma Anim, Kwaku Anim, Emily and Ruby. You guys made Cleveland feel like home to me and I am so blessed to have had you guys. God bless you.

To my friends : Dr. Thomas Atta-Fosu, Isaac Opoku-Agyemang, Kwaku Addae-Ankrah, Thelma Asare and Beverly Boamah, you have been a great community and I am blessed to have had you along the journey.

The presbytery and the entire congregation of the Church of Pentecost, Cleveland especially Dr. Ackah Toffey and Elder Oscar Johnson. God bless you for all the support and love you shown me in diverse ways.

The graduate students in the Mathematics Department have also been very supportive. Dr. Jamie Prezioso,Paromita Banerjee , Carrie Winterer and Richard Lartey I feel so blessed to have met you guys; My office mates Sararose Nassani and Rhiannon Griffiths; And the entire graduate student body. These past four years have been more successful with your varied support.

And above all, I am thankful to my family: Mr. Kwadwo Baffoe-Abrebese, Ivy Diana Woode Baffoe and Afia Adasi Baffoe-Adusei. Your prayers and love have always been my source of encouragement these past years. You are the best family anyone could have and I am thankful for having you as my mine. Love you.

Diagnostic Tools for Forecast Ensembles

Abstract

by

NANA AMA APPIAA BAFFOE

Forecasting is an important area in statistics and as a result it is important that our forecasts reflect our uncertainties. But most importantly, our forecasts should be as accurate as possible. And how can forecasters tell whether their probabilistic forecast distribution are the same or close to the true distribution, which is unknown most of time (if not all time) to the forecaster. We need to come up with a diagnostic tool that helps us to know how close our probabilistic forecasts distributions are to the true distribution. Verification rank histograms and probability integral transforms (PIT) histograms are the most common diagnostic tools to determine if probabilistic forecast distributions and observations are well calibrated in the univariate settings. Calibration in a nutshell means how statistically compatible the probabilistic forecasts and observations are. The purpose of this study is to compare the sensitivity of the following calibration metrics/multivariate ranking methods - Multivariate ranking method, Minimum Spanning Tree, Band Depth and Average ranking method to mispecifications. A simulation study and a case study of the Orbiting Carbon Observatory 2 (OCO-2) are presented.

The general findings from our study is that, when comparing the four diagnostic tools for forecast ensembles, the minimum spanning tree and the band depth methods are better

at detecting misspecifications than the multivariate rank method. Also the average rank method with the band depth method and/or minimum spanning tree method gives us more information than band depth or minimum spanning tree alone.

1 Introduction

1.1 Probabilistic Forecasting

Have you ever wondered how your local meteorologist or your weather app predicts the weather for the next hour or day or even week? Or even how bankers give out their portfolio values? These and many other predictions or forecasts are made possible with uttermost precision because of probabilistic forecasting. Arguably, the most mature and successful implementation of probabilistic forecasting methods is in weather predicition.[1]

Probability forecasting refers to the process of assigning numerical probabilities to an uncertain event. One of the major purposes of statistical analysis is to make forecasts for the future and therefore it would be important if these forecasts had our uncertainties associated with them (Dawid 1983)[2]. There are two (2) types of probabilistic forecasting commonly used. These are ensemble forecasts and density forecasts.

A forecast ensemble with m ensemble members has a **discrete** predicitive distribution which usually assigns probability mass $\frac{1}{m}$ to each ensemble. Ensemble forecasts provide

an ensemble or set of values or range of future possibilities, hence its name. Forecast ensembles are popularly used in weather and climate predictions. The ensembles are usually obtained by running weather forecasting models using a range of initial conditions and parameter values. These weather models are deterministic.

Ensemble forecasting which is a method commonly used in numerical weather predictions is providing an ensemble (set) of values or range of future possibilities. Ensemble forecasting allows us to incorporate our uncertainty which is why it is preferred to the point (single) forecasts which has no form of uncertainty associated with it. Weather can be very unpredictable that's why it's necessary for its forecast to have some level of uncertainty. The less knowledgeable we are about a condition, the higher our uncertainty. In weather forecasting, the shorter the range of the forecast is ,the better the forecast.

The density forecast on the other hand has a **continuous** predictive distribution. The density forecasts are popular in economic and financial applications. In this paper, we concentrate more on the ensemble forecasting.

1.2 Diagnostic Tools For Univariate Forecasts

Calibration is concerned with the statistical compatibility between the probabilistic forecasts and the realizations and is a joint property of predicive distributions and the vector-valued events that materalize[1]. Essentially, the observations should be indistinguishable from random draws from the predictive distributions for a probabilistic forecast to be well-calibrated.

Suppose at an instance or time t, nature chooses/follows a distribution which is unknown to a forecaster (*which is always the case in practice*). Let's call this distribution $\mathbf{G}_t$. We let $\mathbf{x}_t$ be the observation drawn from natures true distribution $\mathbf{G}_t$. At the same time t, a forecaster makes a probabilistic forecast which is in the form of a predicitive cummulative distribution function (cdf), call it $\mathbf{F}_t$. The probabilistic forecasts chosen/given by the forecaster, usually depends on the expertise and experience of the forecaster and it may or may not be derived from a statistical algorithm. The forecasts made by the forecaster are said to be *ideal* or *perfect* if[3]

$$\mathbf{G}_t = \mathbf{F}_t \quad \text{for} \quad \text{all} \quad \text{t} \tag{1.1}$$

where $t = 1, 2, ...$

Suppose $\mathbf{F}_t$ and $\mathbf{G}_t$ are continuous and therefore strictly increasing. Because the true distribution ($\mathbf{G}_t$), is not observed in practice, we perform any calculations that need to be done on forecasts ($\mathbf{F}_t$) and the observation ($\mathbf{x}_t$) only where $\mathbf{x}_t$ is an observation from $\mathbf{G}_t$.

In the univariate case, calibration can be assessed using the probability integral transform (PIT). That is in order to evaluate the statistical compactibility between the probabilistic forecast and the observation, Dawid(1984) and Diebold et al (1998)[4] proposed the use of the probability integral transform (PIT) value. The probability integral transform (PIT) value is the value that a predictive cumulative distribution function (CDF) attains at the observation. Mathematically is given as:

$$p_t = F_t(\mathbf{x}_t) \tag{1.2}$$

If the forecasts are ideal and the F_t is continuous, then the p_t has a uniform distribution. Hence the uniformity of the PIT is a necessary but not sufficient condition for a forecast to be ideal in the univariate case. So there could be instances where the PIT histograms generated are flat or uniform but the probabilistic forecast and the observation do not follow the same distribution.

Probability integral transform (PIT) is suitable for situations in which forecast is presented as a continuous probability distribution functions. Therefore we describe them with an equation or formula and this equation is known as probability density function (pdf). The probability integral transform says:

If x is a continuous random variable with a cdf $F_X(\mathbf{x})$ and if $Y = F_X(X)$, then Y is a uniform random variable on the interval [0,1]. . So simply put, the idea behind PIT is that if one plugs a random variable into its own cdf, one gets a uniform distribution. And that is why when an observation (x_t) has the same distribution the forecast distribution F_t,the PIT values tend to follow a uniform distribution.

Calibration is empirically tested by plotting histograms of PIT values and checking for uniformity (Dawid 1984; Diebold et al. 1998; Gneiting et al. 2007). These histograms generated by the PIT values are known as PIT histograms. These PIT histograms makes it visually easier for the forecaster to determine if the probabilistic forecasts are ideal. For an ideal forecast, the PIT histograms are flat or uniform. When we do not have the analytical

form of F_t but have samples from it, calibration can be assessed with a verification rank histogram (Hamill 2000).

These rank histograms are generated by tallying the ranks of the verification (i.e. the observations) relative to values from an ensemble sorted from lowest to highest (Hamill 2000)[5]. The rank histogram are equivalent to the PIT histograms in the sense that if the observations come from the same distribution as the m ensembles, the ranks are uniformly distributed over the set $1, 2, ...m$. The rank histograms are used for ensemble forecasts whiles the PIT histograms are for density forecasts.

The PIT histogram or the verification rank histograms can provide information as whether a probabilistic forecast ensemble is well-calibrated or not. For example, a well-calibrated probabilistic forecast ensemble has a flat/uniform histogram, a $\cup$-shaped PIT histogram indicates that predicitive distributions are too narrow or underdispersed and $\cap$-shaped PIT histogram indicates overdispersion or predicitive distributions are too wide on an average[3]. Furthermore, a skewed PIT histogram would suggest that the forecast distribution is biased, i.e. has a different mean than the true distribution.

1.3 Diagnostic Tools for Multivariate Forecasts

The univariate tools do not generalize directly to the multivariate setting because there is no universal way of ordering multivariate vectors, and this is true for both multivariate density forecasts and ensemble forecasts. In the multivariate case, which is the case of interest in this paper, in order to rank an observation, we first apply a pre-rank function

(which differs from each diagnostic tool/calibration metric) to the observation and the forecast ensembles in order to get quantities called pre-ranks. To obtain a verification rank histogram, we find the rank of the pre-rank of the observation when pooled within the ordered pre-ranks ensemble and plot of the histogram of those ranks.

In other words, multivariate properties are mapped to a single dimension through a pre-rank function and the calibration is subsequently assessed through a histogram of the ranks of the observation's preranks[6]. We will explain in more details in chapter 2.

We discuss four ranking methods that have been proposed for assessing ensemble forecast calibration in the multivariate setting in this paper. These ranking methods or calibration metrics are - Multivariate rank, Minimum Spanning tree, Average rank and Band depth ranking method.

In this Thesis, we investigate how well these calibration metrics detect misspecifications in these probabilistic forecast ensembles and their strengths and the weaknesses. For example, which of these methods detects misspecified mean or correlation structure of the probabilistic forecast ensembles well. Which of these calibration metrics works better with a bigger sample among others. These calibration metrics general follow the same procedure, the only difference among them, are how their preranks are assigned.

The remainder of this Thesis is organized as follows. In Chapter 2, we look at four (4) multivariate ranking methods - Multivariate rank, Minimum Spanning tree,average rank and Band depth ranking method. In Chapter 3, we present a simulation study using these

all ranking methods. We look at application of these calibration metrics to a real data set in Chapter 4 and we conclude in Chapter 5.

2 Ranking Methods

2.1 What are Ranking Methods/ Diagnostic Tools?

Ranking methods/diagnostic tools are methods used to order multivariate vectors and check the calibration of multivariate ensemble forecasts. In order to plot these histograms, we need to apply a ranking method on the forecast ensemble and the observation. The Multivariate ranks, Minimum spanning tree, Band depth and Average ranks are the four ranking methods we consider in this Thesis. The general set up for using the four ranking methods for calibration metrics discussed here is the same. The main difference among these are their pre-rank functions i.e. how multivariate vectors are ordered. In other words how pre-ranks are assigned in a ranking method determines the method.

All these ranking methods assume the probabilistic forecasts are given by ensemble. We define $\mathbf{B} = \{\mathbf{x}_1, ..., \mathbf{x}_m\}$ to denote a set of points in $\mathbb{R}^d$ or a d-dimensional subset with $\mathbf{x}_j = (x_{j1}, ..., x_{jd})$. That is to say we assume $\mathbf{B}$ contains m elments of which, $m-1$ are the probabilistic forecast ensembles and the corresponding observation $y = \mathbf{x}_m$. This is the general set-up for all four 4 ranking methods. In order to generate a verification rank histogram(*also called Talagram diagram*), we first need to rank the observation $\mathbf{x}_m$ when

pooled within the ordered ensemble values and plot the histograms of the ranks. To rank the observation ($\mathbf{x}_m$) in $\mathbf{B}$, we follow these two main steps.[7].

(1) Apply a prerank function $\rho_B : \mathbb{R}^d \rightarrow R_+$ to calculate the prerank $\rho_B(\mathbf{x})$, of every $\mathbf{x}_j \in B$.

(2) Set the rank of the observation $\mathbf{x}_m$ equal to the rank of $\rho_B(\mathbf{x}_m)$ in $\rho_B(\mathbf{x}_1)$,...,$\rho_B(\mathbf{x}_m)$ with ties resolved at random.

A visualization of the verification rank histograms generated by these ranking methods provide information about whether our ensemble forecasts are well-calibrated or not. The pre-rank functions have different properties and the rank histograms generated cannot be interpreted in the same way. In this Chapter we go into details of each ranking method/calibration metric. In the subsequent Section 2.2 we explain the concept of the multivariate rank method and give an example. We similarly do that for the average rank method in Section 2.3, the minimum spanning tree rank method in Section 2.4 and the band depth rank method in Section 2.5.

2.2 Multivariate rank

The multivariate rank histogram was proposed as a diagnostic tool by Gneiting et. al (2008)[7] and is defined as follows: We consider ensemble forecasts for a vector-valued quantity that takes values in $\mathbb{R}^d$. Given vectors $\mathbf{x} = (x_1, x_2, ..., x_d)' \in \mathbb{R}^d$ and $\mathbf{y} = (y_1, y_2, ..., y_d)' \in \mathbb{R}^d$, we write

$$\mathbf{x} \leq \mathbf{y} \quad \text{if and only if} \quad x_j \leq y_j \quad \text{for} \quad j = (1, 2, ..., d) \tag{2.1}$$

For example, given the vectors $\mathbf{x}_1 = \{2,4,5\}$, $\mathbf{x}_2 = \{1,5,5\}$ and $\mathbf{x}_3 = \{8,10,9\}$. Here we have $\mathbf{x}_1 < \mathbf{x}_3$ because each element in $\mathbf{x}_1$ is lower than the corresponding element in $\mathbf{x}_3$. Similarly, we have $\mathbf{x}_2 < \mathbf{x}_3$. The case of $\mathbf{x}_2$, although two elements in $\mathbf{x}_1$, (4,5), are less than or equal to the corresponding elements in $\mathbf{x}_2$, it still does not make $\mathbf{x}_1 < \mathbf{x}_2$ so $\mathbf{x}_1 \not< \mathbf{x}_2$. On the other hand, $\mathbf{x}_2 < \mathbf{x}_1$ does not hold either, so we have $\mathbf{x}_2 \not< \mathbf{x}_1$. In order to construct a multivariate rank histograms, we would need to compute the multivariate rank by repeating over individual forecast cases. Simply put, the multivariate rank histogram is a plot of the empirical frequency of the multivariate ranks[7]. For a given ensemble forecast $\mathbf{x}_j \in \mathbb{R}^d : j = 0,1,.....,m$ and a verifying observation $\mathbf{x}_0 \in \mathbb{R}^d$, where m is number of ensemble members, we compute the multivariate ranks as follows. [1]

(1) We assign the pre-ranks. The prerank function is defined as :

$$\rho_j = \sum_{k=0}^{m} \mathbb{1}(\mathbf{x}_k \leq \mathbf{x}_j) \tag{2.2}$$

where $\mathbb{1}$ is the indicator function

$$\mathbb{1}(\mathbf{x}_k \leq \mathbf{x}_j) = \begin{cases} 1 & \text{if } x_k \leq x_j \\ 0 & \text{otherwise} \end{cases}$$

The pre-ranks are integers and they range between 1 and $m+1$. The pre-rank is calculated for all $j = 0,1,....m$

[1] Sometimes we standardize the ensemble member forecast and the verifying observation using a principal component transform as suggested by[7] but we do not do that in the paper.

(2) We compute the multivariate rank r next. The multivariate rank, is the rank of the observation pre-ranks with ties resolved at random so the multivariate rank, r is within $s^{<} + 1$ and $s^{<} + s^{=}$, where

$$s^{<} = \sum_{j=0}^{m} \mathbb{1}(\rho_j < \rho_0) \tag{2.3}$$

$$s^{=} = \sum_{j=0}^{m} \mathbb{1}(\rho_j = \rho_0) \tag{2.4}$$

Let's apply the multivariate rank method to a hypothetical example. Suppose we have 6 ensemble member forecast ($m = 6$), $\mathbf{x}_j : j = 1,2,3,4,5,6$ and an observation $\mathbf{x}_0$ with $\mathbf{x}_0 = \{4,2,5\}$, $\mathbf{x}_1 = \{3,2,3\}$, $\mathbf{x}_2 = \{5,3,7\}$, $\mathbf{x}_3 = \{2,1,3\}$, $\mathbf{x}_4 = \{9,8,9\}$, $\mathbf{x}_5 = \{2,2,1\}$ and $\mathbf{x}_6 = \{7,4,3\}$. We begin by assigning the pre-ranks, each pre-rank ranges from 1 to 7. We have

$$\begin{aligned}
\rho_0 &= \sum_{k=0}^{6} \mathbb{1}(\mathbf{x}_j \leq \mathbf{x}_0) \\
&= \mathbb{1}(\mathbf{x}_0 \leq \mathbf{x}_0) + \mathbb{1}(\mathbf{x}_1 \leq \mathbf{x}_0) + \mathbb{1}(\mathbf{x}_2 \leq \mathbf{x}_0) + \mathbb{1}(\mathbf{x}_3 \leq \mathbf{x}_0) + \mathbb{1}(\mathbf{x}_4 \leq \mathbf{x}_0) + \mathbb{1}(\mathbf{x}_5 \leq \mathbf{x}_0) + \mathbb{1}(\mathbf{x}_6 \leq \mathbf{x}_0) \\
&= 1 \quad + \quad 1 + \quad 0 + \quad 1 + \quad 0 + \quad 1 + \quad 0 \\
&= 4
\end{aligned}$$

To compute the pre-rank for ensemble member 1 (ρ_1), we go through similar procedure.

$$\rho_1 = \sum_{j=0}^{6} \mathbb{1}(\mathbf{x}_j \leq \mathbf{x}_1)$$

$$= \mathbb{1}(\mathbf{x}_0 \leq \mathbf{x}_1) + \mathbb{1}(\mathbf{x}_1 \leq \mathbf{x}_1) + \mathbb{1}(\mathbf{x}_2 \leq \mathbf{x}_1) + \mathbb{1}(\mathbf{x}_3 \leq \mathbf{x}_1) + \mathbb{1}(\mathbf{x}_4 \leq \mathbf{x}_1) + \mathbb{1}(\mathbf{x}_5 \leq \mathbf{x}_1) + \mathbb{1}(\mathbf{x}_6 \leq \mathbf{x}_1)$$

$$= 0 \quad + \quad 1 + \quad 0 + \quad 1 + \quad 0 + \quad 1 + \quad 0$$

$$= 3$$

Note $\mathbb{1}(x_k \leq x_k) = 1$ is always going to be true. That's why the smallest pre-rank for an observation and or ensemble member forecast is always going to be at least 1.

From the above, we know the pre-ranks for the observation and ensemble member 1 are 4 and 3 respectively. The pre-ranks for the remaining 5 ensemble members were computed in similar manner and are given in the table below. The pre-ranks for ensemble members 2 to 6 are given as $\rho_2, \rho_3, \rho_4, \rho_5, \rho_6$ respectively.

Vector	ρ_0	ρ_1	ρ_2	ρ_3	ρ_4	ρ_5	ρ_6
Pre-rank	4	3	4	1	7	1	4

Table 2.1. Pre-ranks using multivariate ranking method for a hypothetical example

We go ahead and find the multivariate rank r for the observation. We first find $s^<$ and $s^=$.

$s^< = \sum_{j=0}^{6} \mathbb{1}(\rho_j < \rho_0)$

$$= \mathbb{1}(\rho_0 < \rho_0) + \mathbb{1}(\rho_1 < \rho_0) + \mathbb{1}(\rho_2 < \rho_0) + \mathbb{1}(\rho_3 < \rho_0) + \mathbb{1}(\rho_4 < \rho_0) + \mathbb{1}(\rho_5 < \rho_0) + \mathbb{1}(\rho_6 < \rho_0)$$

$$= \mathbb{1}(4 < 4) + \mathbb{1}(3 < 4) + \mathbb{1}(4 < 4) + \mathbb{1}(1 < 4) + \mathbb{1}(7 < 4) + \mathbb{1}(1 < 4) + \mathbb{1}(4 < 4)$$

$$= 0 \quad + 1 \quad + 0 \quad + 1 \quad + 0 \quad + 1 \quad + 0$$

$$= 3$$

$$s^{=} = \sum_{j=0}^{m} \mathbb{1}(\rho_j = \rho_0)$$

$$s^{=} = \mathbb{1}(\rho_0 = \rho_0) + \mathbb{1}(\rho_1 = \rho_0) + \mathbb{1}(\rho_2 = \rho_0) + \mathbb{1}(\rho_3 = \rho_0) + \mathbb{1}(\rho_4 = \rho_0) + \mathbb{1}(\rho_5 = \rho_0) + \mathbb{1}(\rho_6 = \rho_0)$$

$$= \mathbb{1}(4 = 4) + \mathbb{1}(3 = 4) + \mathbb{1}(4 = 4) + \mathbb{1}(1 = 4) + \mathbb{1}(7 = 4) + \mathbb{1}(1 = 4) + \mathbb{1}(4 = 4)$$

$$= 1 \quad + 0 \quad + 1 \quad + 0 \quad + 0 \quad + 0 \quad + 1$$

$$= 2$$

Therefore

$$r \in \{s^{<} + 1, ..., s^{=} + s^{<}\} \tag{2.5}$$

$$r \in \{3 + 1, ..., 2 + 3\} \tag{2.6}$$

$$r \in \{4, 5\} \tag{2.7}$$

Hence in this example,the multivariate rank r could either be 4 or 5. We choose r at random say $r = 5$. We could avoid this random selection of r by using the construction proposed by Czado et. el (2007)[8] but we will not do that here. The multivariate rank reduces to the univariate verification rank when the dimension $d = 1$.

2.3 Average rank

The average rank method was introduced by Thorarinsdóttir et al (2016)[6]. The average rank is simply average over the univariate ranks of each element of the observation vector[6]. Let $rank_B(x_{jk})$ denote the rank of the k^{th} coordinate of $\mathbf{x}_j$ in B, i.e.

$$rank_B(x_{jk}) = \sum_{i=0}^{m} \mathbb{1}(x_{ik} \leq x_{jk}) \tag{2.8}$$

The prerank function is given by

$$\rho_B^a(\mathbf{x}_j) = \frac{1}{d} \sum_{k=1}^{d} rank_B(x_{jk}) \tag{2.9}$$

The average rank is reduced to the classical univariate rank when $d = 1$. The interpretation of the average rank histogram is similar to univariate rank histogram. That is, if the forecast are underdispersive the average rank for the observation is U-shaped, an overdispersive ensemble results in $\cap$ - shaped histogram while a constant bias results in a triangular shaped histogram. Both multivariate rank and average rank have their prerank functions analogously to the univariate rank histogram. They both provide measures of "ascending rank" of the observation vector $\mathbf{x}_0$ relative to the ensembles.

Let $\mathbf{B}$ be a set of $m-1$ ensemble and an observation vector. In order to compute the average rank of an observation $\mathbf{x}_m$, we first rank each component of $\mathbf{x}_m$ relative to its corresponding components of the ensemble members. The ranks are assigned in an ascending order. Then we compute the average rank by computing the average of all the ranks for each

vector. Finally to determine the average rank of the observation $\mathbf{x}_m$, we rank(in an ascending order) the average ranks of all the vectors then determine the rank of the observation vector.

For example,suppose we have a dimension $d = 3$ and $m = 4$,that is we have 3 ensemble vectors and an observation vector. Note, ties are resolved at random. Let $\mathbf{x}_4$ be the observation.

$$\mathbf{B} = [\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4] \tag{2.10}$$

Suppose in this example our $\mathbf{B}$ is given as

$$\mathbf{B} = \begin{bmatrix} 2 & 10 & 5 & 7 \\ 15 & 12 & 13 & 9 \\ 8 & 6 & 12 & 28 \end{bmatrix}$$

To rank $\mathbf{x}_4$, which is the observation vector,we first rank the first component of $\mathbf{x}_4$ (first row) with respect to the ensemble members which is 7 in this case. So we rank 7 with respect to the ensemble members which are 2 ,10 and 5. The rank of 7 is 3 as we can see. Then we come to the second row, the rank of x_{42} i.e. (9) is 1, then last row, the rank of the observation (28) is 1. This procedure is known as pre-ranking. We rank also the ensemble vectors in similar manner. Below is the table of the pre-ranks of observation and ensemble members in B.

$$\text{Pre-ranks of B} = \begin{bmatrix} 1 & 4 & 2 & 3 \\ 4 & 2 & 3 & 1 \\ 2 & 1 & 3 & 4 \end{bmatrix}$$

From the table above, we see $\mathbf{x}_4$ has following ranks 3 ,1 and 1 likewise $\mathbf{x}_1$ has ranks 1 ,4 and 2. The ranks for $\mathbf{x}_2$ and $\mathbf{x}_3$ are given in the second and third columns respectively. We then go ahead and compute averages of ranks for each vector and then rank these averages in an ascending order. Ties are resolved at random. Below is the table for the average of the ranks and their respective ranks. For example, the average of the ranks in $\mathbf{x}_4$ is given by (3 + 1 + 4)/3 which is 2.667.

Vector	Average of ranks	Rank
$\mathbf{x}_{(1)}$	2.333	1
$\mathbf{x}_{(2)}$	2.333	2
$\mathbf{x}_{(3)}$	2.667	3
$\mathbf{x}_{(4)}$	2.667	4

Table 2.2. Pre-ranks using Average ranking method for a hypothetical example

From the table 2.2 above, we see that $\mathbf{x}_4$ has one of two smallest pre-ranks and we conclude (*after a random draw*) that the average rank for the observation $\mathbf{x}_4$ is 4.

2.4 Minimum Spanning Tree

What is a spanning tree? Let $G = (V, E)$ undirected and connected graph *where V are the set of vertices and E are the set of edges.*

A spanning tree of the graph G is a tree that spans G in the sense that it includes every vertex of G and is a subgraph of G (every edge in the tree belongs to G). The cost of the spanning tree is the sum of the weights of all the edges in the tree. Minimum spanning tree is therefore the spanning tree where the cost is minimum among all the spanning trees. A graph can have many minimum spanning trees. Two famous algorithms for finding a Minimum Spanning Tree- Kruskal(1956)[9] and Prim(1957)[10] are given below. Any of these algorithms can be used is solely the choice of the experimenter. MSTs has many applications including the network design, cluster analysis and among others.

Kruskal algorithm

(1) Label each vertex (v)

(2) List the edges in non decreasing order of weight,starting the shortest edge.

(3) Start with the smallest weighted and begin growing the minimum weighted spanning tree from this edge.

(4) Add the next available edge that does not form a cycle to the construction of the minimum weighted spanning tree.

(5) Continue with step 4 until you have a spanning tree(.i.e until we have $v-1$ edges).

Prim algorithm

(1) Choose a starting vertex for your tree at random.

(2) Find the edge with the smallest weight that connects the tree to the vertex that is not in the tree, and add it to the tree.

(3) Continue this until all of the vertices are in the tree.

The basic difference between the Prim algorithm and the Kruskal algorithm is which edge to choose to add to the next spanning tree in each step.In Prim's, you always keep a connected component (.i.e the graph must be a connected graph),starting with a single vertex and then look for all the edges from the current component to other vertices and find the smallest among them.In the Kruskal algorithm, you start by sorting the edges by length and adding them to the tree in order, the shortest edge.Kruskal begins with a forest and then merge into a tree whiles the Prim always remains as a tree.

Minimum Spanning Tree Rank Method

The minimum spanning tree(MST) rank was introduced as a multivariate calibration assessment tool by Smith (2001)[11]. This rank method is used to assess the reliability of ensemble forecast where the ensemble forecast have high dimensions and therefore the regular rank histogram(used for scalar ensemble forecast) cannot be used. In a nutshell, the MST approach yields rank histogram that can evaluate ensembles in an m-dimensional space.

This is how the MST rank method works. For a given n ensemble members and an observation(verification), the ranks range from 1 to $n+1$. First compute the MST length for the ensemble members **only** call it say m_{ens}. Then take turns replacing one ensemble member with the observation and compute the MST length for each.

For example, we replace ensemble member A with the observation O,we then compute the MST length, let's call it m_a. Then we replace ensemble member B with the observation and calculate the MST length and call it m_b. We repeat this process for all n ensemble members, the last MST length would be m_n. In order to rank the ensemble member m_{ens}, we rank the calculated MST lengths in an ascending order. In this is case, we rank $m_f, m_1, m_2, ..., m_n$ in an ascending order. Ties are resolved randomly.

Let's take an hypothetical example, if we have 9 ensemble members and an observation,which implies rank of ensemble member ranges from 1 to 10). Below are the plots of minimum spanning tree (MST) using Prim algorithm.

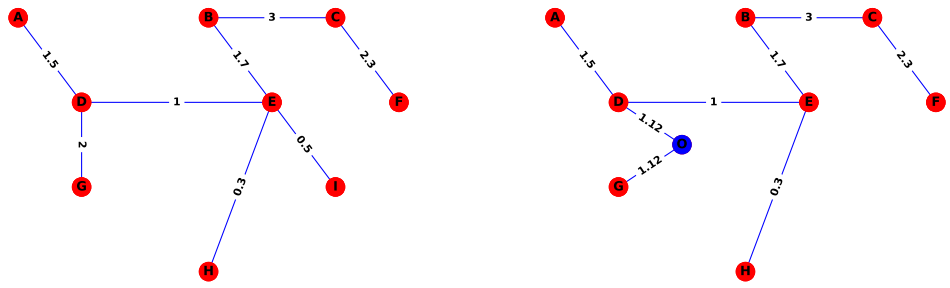

Figure 2.1. Hypothetical example of Minimum spanning tree (MST) in d = 2 dimensions. **Left**: The n_{ens} = 9 ensemble members labeled from A-I. **Right**: We replace the ensemble member I with the observation O (*which is in blue*). The edge weights do not represent the Euclidean distances between nodes. This MST was computed using the Jarnik-Prim algorithm.

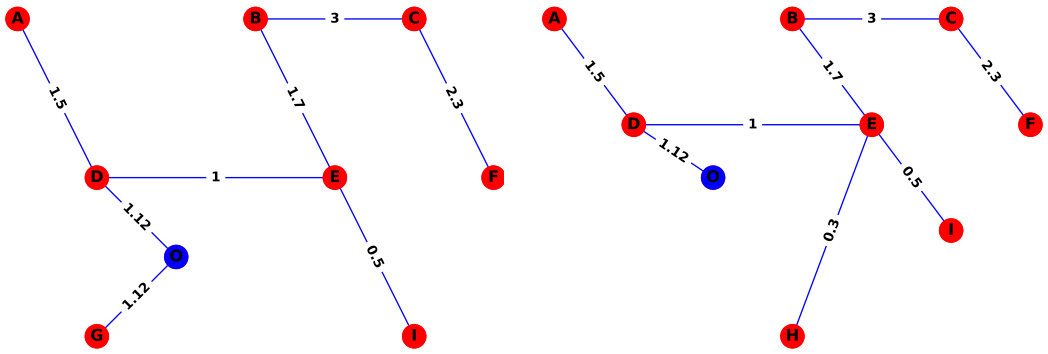

Figure 2.2. **Left**: We replace the ensemble member H with the corresponding observation O (*which is in blue*).**Right**: We replace the ensemble member G with the corresponding observation O (*which is in blue*). The length/weights of the edges are indicated on the lines. This MST was computed using the Jarnik-Prim algorithm.

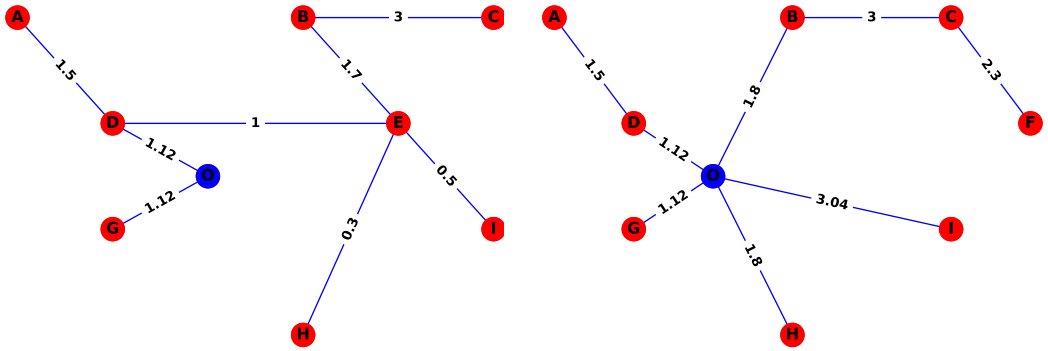

Figure 2.3. **Left**: We replace the ensemble member F with the corresponding observation O (*which is in blue*). **Right**: We replace the ensemble member E with the corresponding observation O (*which is in blue*). The length/weights of the edges are indicated on the lines. This MST was computed using the Jarnik-Prim algorithm.

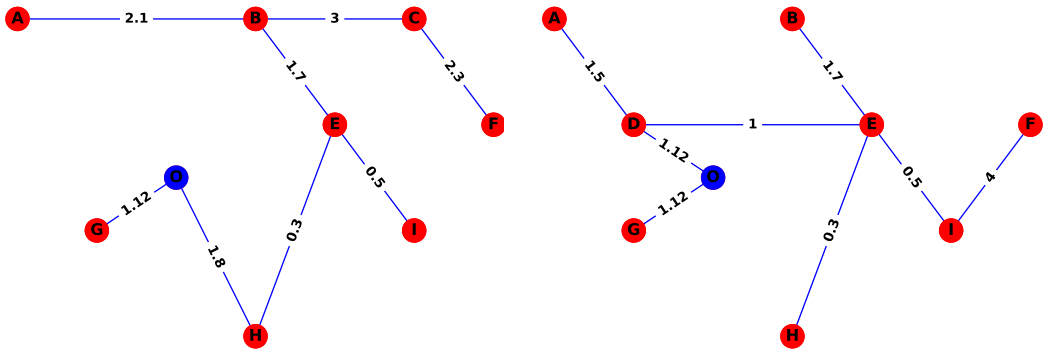

Figure 2.4. **Left**: We replace the ensemble member D with the corresponding observation O (*which is in blue*). **Right**: We replace the ensemble member C with the corresponding observation O (*which is in blue*). The length/weights of the edges are indicated on the lines. This MST was computed using the Jarnik-Prim algorithm.

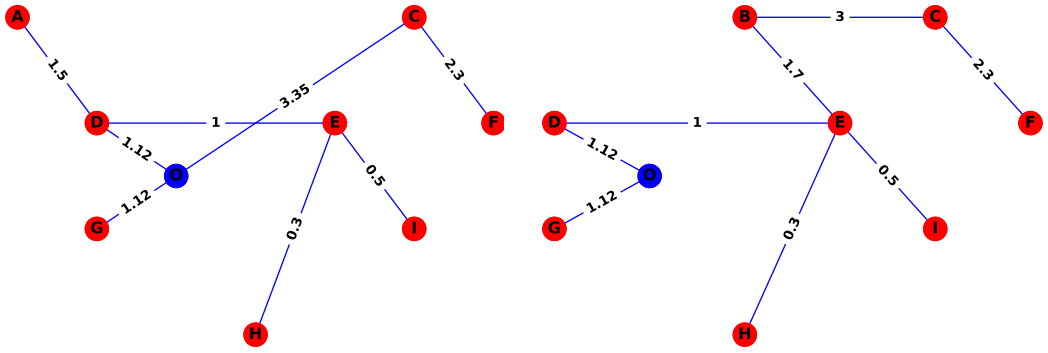

Figure 2.5. **Left**: We replace the ensemble member B with the observation O (*which is in blue*). **Right**: We replace the ensemble member A with the observation O (*which is in blue*). The length/weights of the edges are indicated on the lines. This MST was computed using the Jarnik-Prim algorithm.

In order to find the minimum spanning tree (MST) lengths, we sum the weights of each minimum spanning tree. For example, the MST length for m_i (*see figure 2.1*) is given as

$m_i = 1.5 + 1.12 + ... + 3 + 2.3$ which gives 12.04. The rest of MST lengths were computed in similar manner and they are shown below:

$$m_a = 11.04$$

$$m_b = 11.19$$

$$m_c = 11.24$$

$$m_d = 12.82$$

$$m_e = 15.68$$

$$m_f = 10.24$$

$$m_g = 11.42$$

$$m_h = 12.24$$

$$m_i = 12.04$$

$$m_{ens} = 12.3$$

We then rank these MST lengths in an ascending order shown in table 2.3, and find the the rank for the ensemble member is 8.

When the ensemble member forecast and observation follow the same distribution, the MST rank histogram is going to be flat or uniform. Likewise if we have smaller ranks the MST histogram is going to be skewed to the right and it would be skewed to the left when we have bigger ranks. We note that MST ranks give an indication of how "central" the

Names	MST length	Rank
m_f	10.24	1
m_a	11.04	2
m_b	11.19	3
m_c	11.24	4
m_g	11.42	5
m_i	12.04	6
m_h	12.24	7
m_{ens}	12.30	8
m_d	12.82	9
m_e	15.68	10

Table 2.3. Ordered MST lengths for a hypothetical example

observation is with respect to ensemble members , m_{ens} will be smaller than most MST's that include the observation. Likewise, if the observation is centrally located the MST's that include it will tend to be smaller than m_{ens}.

2.5 Band depth and Modified Band depth

The band depth or the modified band depth rank method is another diagnostic tool introduced by Thorarinsdóttir et al (2016) [6] and it is based on the concept of band depth for functional data Lopez-Pintado and Romo (2009) [12], introduced a center-outward ordering of curves, which they called band depth. Band depth of a multivariate vector $\mathbf{x}_j$ in a set of vectors B is defined as the proportion of elements of $\mathbf{x}_j = (\mathbf{x}_{j1}, ..., \mathbf{x}_{jd})$ that are inside bands defined by subsets of n vectors. Usually n is set as equal to 2 and we follow that convention here. Therefore, the band depth of $\mathbf{x}_1$ is the proportion of elements of $\mathbf{x}_j$ that are between the corresponding element of pairs of vectors in B. The proportion is taken over all pairs $\binom{m}{2}$ and all elements d. The difference between the band depth and the modified

band depth, is that the band depth rank method has an indicator function which makes it harder to work with in practice but the modified band depth rank method doesn't and thus makes it easier to work with. The modified band depth takes the proportion of times that a curve is in a band of two other curves into account and hence it avoids having too many depth ties. The modified band depth is more convenient to obtain the most representative curves in terms of magnitude, while the band depth rank on the other hand is more dependent on the shape of the curves often yielding ties and thus it can be used to obtain the most representative curves in terms of shape. If a curve $r(t)$ always lies inside the band, the modified band depth degenerates to the band depth. A curve is contained in a band even if this curve is on the border. The more centered the curve, the higher the rank.

The pre-rank function for the band depth is given as:

$$\rho_B^{bd} = \quad \frac{1}{d}\sum_{k=1}^{d}[m - rank_B(x_k)][rank_B(x_k) - 1] + (m+1) \tag{2.11}$$

where $rank_B(x_k) = \sum_{i=1}^{m} \mathbb{1}\{x_{ik} \leq x_k\}$ is the rank of the k^{th} element of k among the k^{th} element of the forecast ensembles.

$$0 \leq \rho_B^{bd} \leq 1 \quad \forall \mathbf{x}$$

The closer the band depth is to 1, the deeper or the more centered the observation curve is. Likewise the closer the band depth is to 0 the more the observation curve is among the outlying curves. In the band depth ranking method, we consider the probabilistic forecast ensembles and observation as curves. In a nutshell, the more centered the observation

curve (*in red*) is in the forecast ensembles, the higher the ranks likewise the more outlying the observation curve (*in red*) is, the lower its rank. For example, observation curve in figure 2.6, a would have a higher rank than observation curve in b.

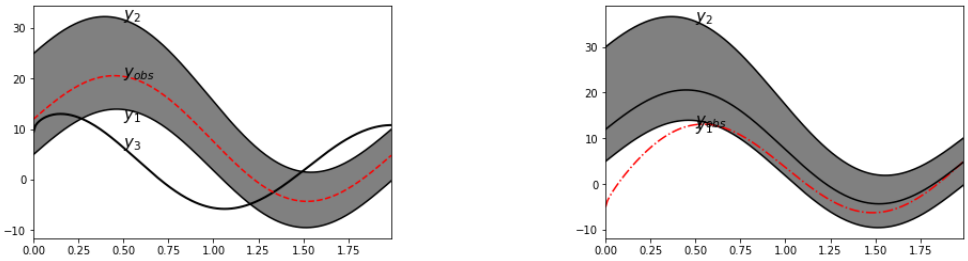

Figure 2.6. An example of band depth for forecast ensembles and observation(*in red*). The grey area is the band delimited by y_2 and y_1. **Left**: Figure (a) **right**: Figure (b)

Although Lopez-Pinatado and Romo (2009) proposed the computation of a bandepth and modifoed rank, the method proposed by Sun et al (2012) [13] is faster and also easier to implement.

For a curve of interest say j, in order to implement the exact method by Sun et al (2012), we must first determine the number of curves that that are completely above the curve of interest, call it n_a, and the number of curves that are completely below the curve of interest, call it also n_b. Note, a curve is also defined as "contained in a band", if it is on the border of the band [13]. Then to compute the band depth of j, the formula by Sun et al (2012) is given by:

$$(n_a \times n_b + n - 1) / \binom{n}{j}$$

We apply the Sun et al method to a hypothetical example. Suppose we have 3 forecast ensemble curves (y_1, y_2 and y_3) and an observation curve (y_{obs}),which implies that $n = 4$. This is shown in figure 2.6 below. From figure 10, the band is delimited by y_1 and y_2 and therefore, $j = 2$. The possible number of bands is 6, that is $\binom{4}{2}$.

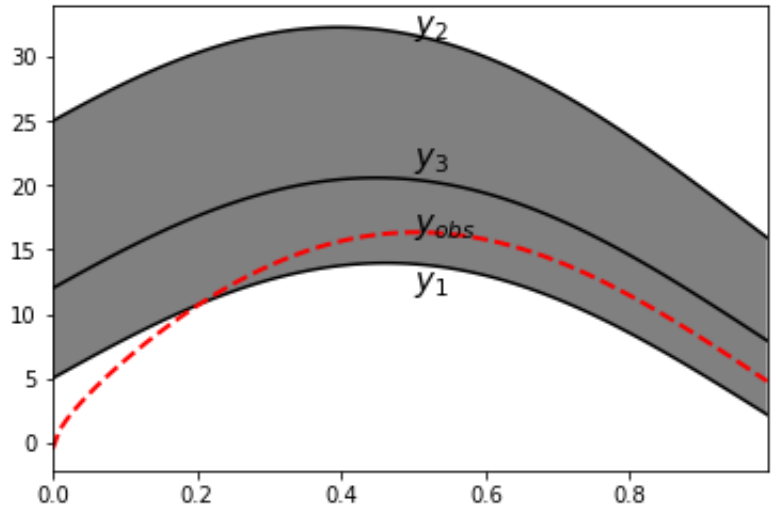

Figure 2.7. An example of the band depth and modified band computation: the grey area is the band delimited by y_2 and y_1. The curve y_3 completely belongs to the band, but y_{obs} only partly does.

To compute the band depth (BD) for y_3, the number of curves that are completely above it (n_a) is 1 (as shown in figure 2.7) and the curves completely above (n_b) is 2. Using the Sun et al method,

$$BD(y_3) = n_a \times n_b + n - 1$$

$$= 1 \times 2 + 4 - 1$$

$$= 5$$

Therefore $BD(y_3)$ is $\frac{5}{6}$ (0.83). Because y_3 is completely in the band, the modified band depth of y_3 is same as the band depth. The band depth for y_{obs}, is given as

$$BD(y_{obs}) = n_a \times n_b + n - 1$$

$$= 3 \times 0 + 4 - 1$$

$$= 3$$

Hence $BD(y_{obs})$ is $\frac{3}{6}$ (0.5).

The computation for modified band depth (MBD) for y_{obs} is little different. For y_{obs}, the curve is partly in the band. Approximately 80% of y_{obs} belongs the band thus MBD(y_{obs}) is $(3 + 0.8 + 0.8)/6 = \frac{4.6}{6} = 0.77$. The band depth and modified band depths for y_1 and y_2 were computed inn similar manner. BD(y_1) = MBD(y_1) = $\frac{3}{6}$ and BD(y_2) = MBD(y_2) = $\frac{3}{6}$.

The closer the value of the band depth/modified band depth is to 1, the more centered the curve is in the band and this has been verified by the example above.

3 Simulation Study

3.1 General Simulation setup

The general setup for this simulation study is to, generate a multivariate observation from a distribution say **G** and a multivariate ensemble from a forecasting distribution, say **F**. The forecasting distribution (**F**) is usually different from that of the distribution of the observation **G** in some way, such as misspecified means and/or covariances. We calculate the rank of the observation in the forecasting ensembles, using one of the four ranking methods. This procedure is repeated many times and then histograms are generated using the ranks.

If **G** = **F**, the histograms will be approximately uniform for all four ranking methods. If **F** is misspecified in some way the histograms of ranks may deviate from uniformity in some way. The goal of this simulation study is to investigate if and how the four ranking methods are able to identify misspecification of the forecasting distribution.

In this simulation study the forecast ensembles $\mathbf{Y} = (Y_1, ..., Y_d)$ follows a Gaussian autoregressive (AR) process with a correlation function given as:

$$cov(Y_i, Y_j) = \exp(-|i-j|/\tau),\ \tau > 0 \quad i, j = 1, .., d \tag{3.1}$$

or a standard multivariate Gaussian $N(0, I)$. The parameter τ controls how fast correlations decay with time lag. τ is set to 3 for the forecast throughout the simulations. We assume that the forecast ensemble consists of 19 members and has a dimension, $d = 16$. The mean of the forecast will either be a vector of zero or shifted in some way to mimic a biased forecast.

We consider five true distributions (**G**s) for the observations in this simulation study. We can think of these observations in terms of weather forecasting as five weather observations from five different locations. All these five observations were simulated from a multivariate normal distribution with mean 0 and variance of 1. The main difference in the observations are their correlation structures which in effect gives them different variance-covariance matrices. We now go into details on how the variance-covariance matrices were formed for each observation. We will examine several cases of forecast misspecification (**F**) by varying the mean vector and variances. Furthermore, the forecasting distribution will have the same correlation structure as one of the true distributions (observation 4) although we also consider a diagonal covariance matrix.

The variance-covariance matrices (Σ) for the forecast and the observations are formed by correlation functions and therefore all the five observations have different variance-covariance matrices (Σs).

The correlation functions for the observations are given in equations 3.2 – 3.6 below. The plots of the correlation functions of the observations (1 to 4) and the forecast against the lags are shown in figures 3.1 and 3.2.

$$obs1 : cov(Y_i, Y_j) = \exp(-|i-j|/4.5)(0.75 + 0.25\cos(\pi|i-j|/2)) \tag{3.2}$$

$$obs2 : cov(Y_i, Y_j) = (1 + |i-j|/2.5)^{-1} \tag{3.3}$$

$$obs3 : cov(Y_i, Y_j) = \mathbb{1}_{|||i-j|\leq 5|/}(1 - |i-j|/5) \tag{3.4}$$

$$obs4 : cov(Y_i, Y_j) = \exp(-|i-j|/\tau),\ \tau > 0 \tag{3.5}$$

$$obs5 : cov(Y_i, Y_j) = 0 \quad \forall i \neq j \tag{3.6}$$

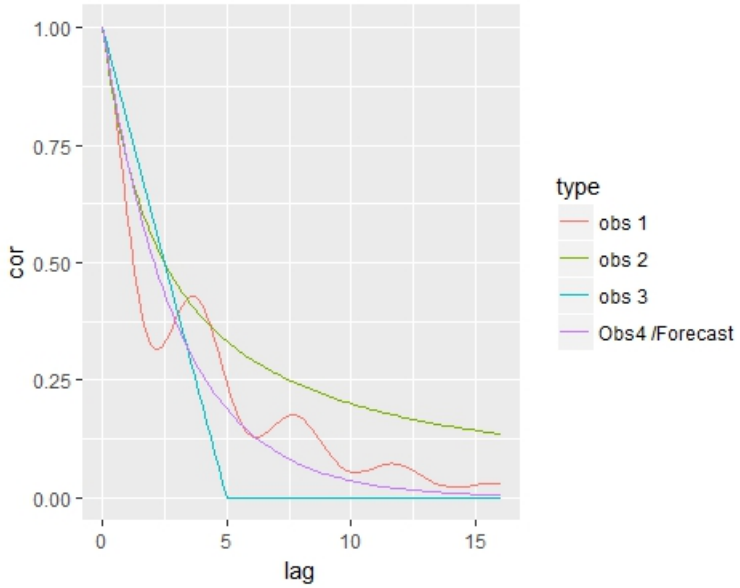

Figure 3.1. Plots of four observation (4) correlation structure/function. Observation 5 has no correlation structure and therefore it is not represented here.

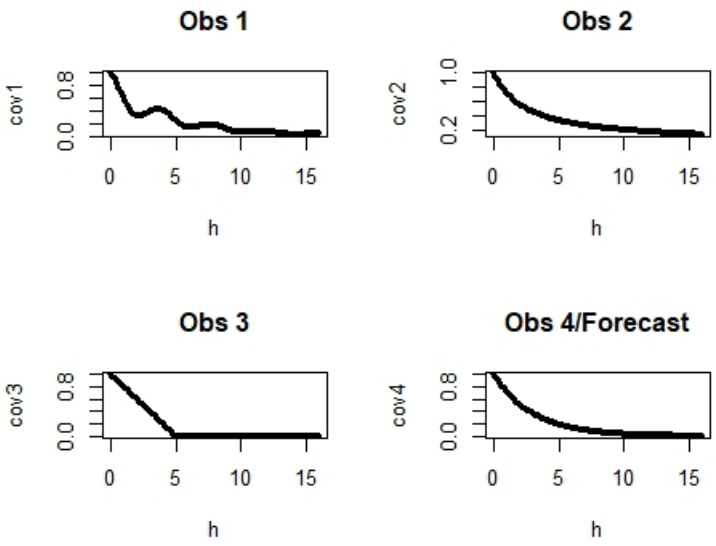

Figure 3.2. Plots of the correlation structure/function against lags. Individual plots of the correlation structure. Observation 4 has the same correlation structure as the forecast. Observation 5 has no correlation structure and therefore it is not represented here.

The elements of the 16 $\times$ 16 dimensional covariance matrices are calculated using these the equations (above) with $i = 1, 2, ...16$ and $j = 1, 2, ..., 16$.

The correlation structure for observation 1 is a damped cosine that oscillates around the exponential model. The correlation function for observation 2 is an exponential of lags (a distance matrix). This exponential model has a much stronger correlations at larger time lags. The correlation function for observation 3 is also an exponential. The difference between that of observation 2 and 3 is that, the correlation structure for observation 3 is a piecewise function. The exponential model for observation 3 has zero correlations for larger time lags[6]. The correlation function/structure for observation 4 is an exponential

of lags (a distance matrix) divided by τ and the Observation 5, on the hand has no correlation structure/function and therefore its variance-covariance matrix is the identity of size 16 (I_{16}). The variances, i.e. the diagonal of the variance-covariance matrices are ones (1) in all the five (5) distributions(observations). The distributions of the five observations remain unchanged throughout these simulations. The forecast distribution (**F**) considered in this simulation study, either has the covariance structure same as observation 4 or are independent correlation function as observation 5. There are times, when we change the mean and/or the variance of the forecast, while the observations remain unchanged. Intuitively we think about the difference in correlation functions between forecast and observation this way:

(1) Forecast has a weaker correlation (shorter correlation length) than observation 2.

(2) Forecast has a slightly stronger correlation (longer correlation length) than observation 3.

(3) Forecast has a similar correlation length as observation 1 but has a monotone correlation function while observation 1 has an oscillating correlation function.

(4) Forecast has much stronger correlation structure than observation 5.

For each observation (vector), 19 ensemble member forecasts (vectors) are generated from **F**. That is we apply the four diagnostic tools discussed in Chapter 2 to a 19-member ensemble forecast generated from forecast distribution such as $N \sim (\mu, \Sigma_4)$, paired with each of the 5 observation vectors generated from the distribution of the observation respectively. This implies the pre-ranks for each observation is between 1 and 20.

3.2 Cases of forecast misspecifications

We now apply all these four ranking methods/diagnostic tool to several hypothetical cases and see how each one behaves. We also wanted to know the sensitivity of these ranking methods to misspecifications better. These cases are misspecifications of various forms to the ensemble member forecast distribution, the ensemble size and the dimension size. In other words, the misspecifications were only made for the ensemble member forecast, leaving the mean, variance-covariance matrices and the correlation functions of all five(5) observations unchanged. The misspecifications of the ensemble member forecasts included but were not limited to the mean, variance-covariance matrix and the correlation structure.

Six (6) cases were considered in this chapter, and these cases are as follows :

In case 1, the correlation structure/function for the ensemble member forecast was misspecified, the forecast distribution had a correlated correlation function or an independent correlation function. That is the correlation function for observation 4 and observation 5 was used for the ensemble member forecast respectively. For Case 2, the variance-covariance matrix of the forecast distribution was specified. The mean (positive) of the ensemble member forecast distribution was misspecified in Case 3. In Case 4, the mean of the forecast was also misspecified. This time the mean of forecast distribution has a negative mean. We considered smaller ensemble size and smaller dimension for Cases 5 and 6 respectively.

3.2.1 Case 1: Misspecified correlation structure/function

Case 1.1: Forecast has correlated correlation structure/function

In case 1.1, the forecast distribution has a correlation structure, that is the forecast distribution has the same correlation structure as observation 4. The distributions of forecast and observations are shown in the table 3.1.

Observation	Distribution	Mean	Variance
Forecast	$N \sim (\mu, \Sigma)$	[0,0,0,...,0]	[1,1,1,...1]
Obs1	$N \sim (\mu_1, \Sigma_1)$	[0,0,0,...,0]	[1,1,1,...1]
Obs2	$N \sim (\mu_2, \Sigma_2)$	[0,0,0,...,0]	[1,1,1,...1]
Obs3	$N \sim (\mu_3, \Sigma_3)$	[0,0,0,...,0]	[1,1,1,...1]
Obs4	$N \sim (\mu_4, \Sigma_4)$	[0,0,0,...,0]	[1,1,1,...1]
Obs5	$N \sim (\mu_5, \Sigma_5)$	[0,0,0,...,0]	[1,1,1,...1]

Table 3.1. Setup for Case 1.1. Forecast distribution has the same mean vector and variances as the five observations and the same covariance function as observation 4.

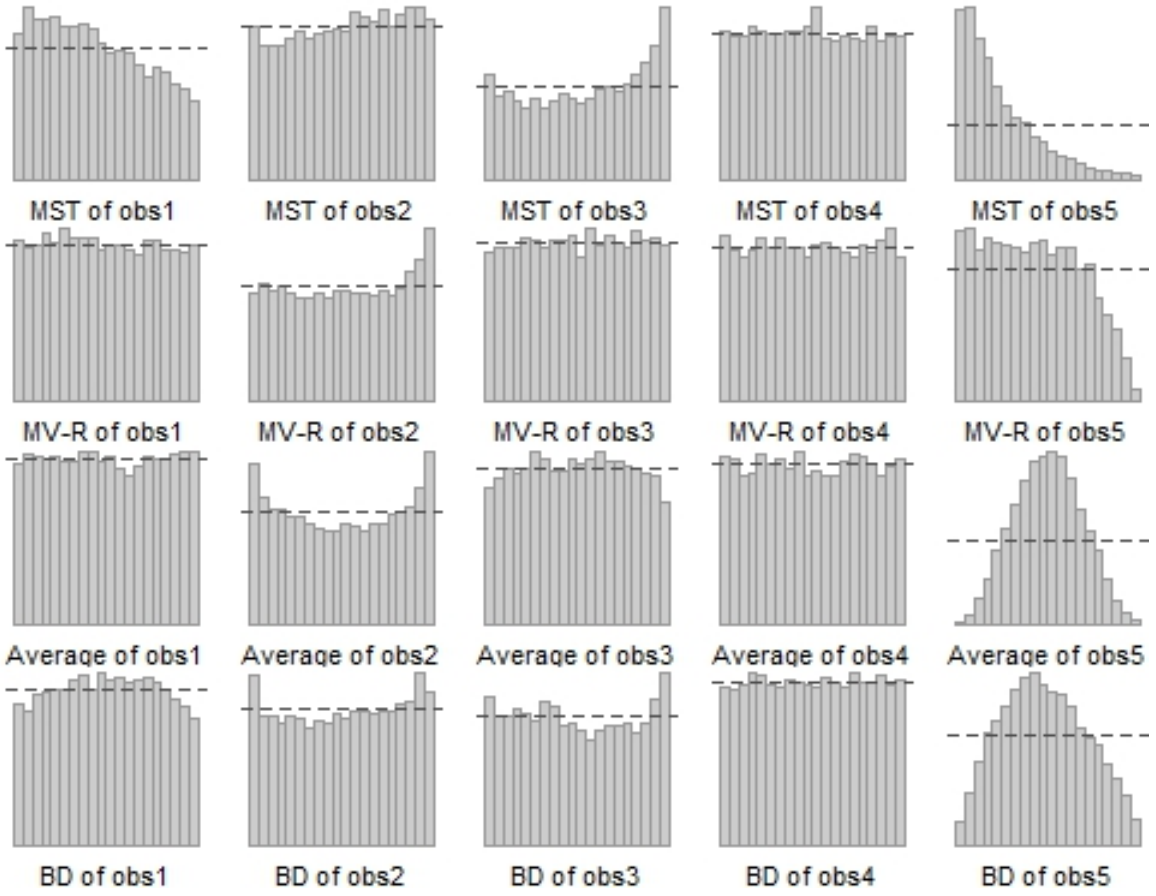

Figure 3.3. Rank histogram for case 1. The means for both forecast and observations are 0*s* and all have a variance of 1. The first row shows the rank histogram for minimum spanning tree,the second shows the multivariate rank, the third average rank and last row show the band depth rank. The five columns show the 5 observations. Note that here the forecast has the same distribution as observation 4. (a) Observation 1 has similar but oscillation correlation. (b)Observation 2 has a stronger correlation than forecast. (c) Observation 3 has a weaker correlation than forecast. (d)Observation 5 has a weaker correlation than forecast. The results are based on 10,000 repetitions.

As expected, in figure 3.3, the rank histograms of observation 4 is uniform for all of the four (4) calibration metrics and this is because both the ensemble member forecast and

observation 4 follow the same distribution. Since the means of the forecast distribution are all zeros 0 and variances are all 1s we are not seeing the effects of bias or under/over dispersion in the rank histogram, only the effect of misspecified correlation function. Although observation 1 doesn't follow the same distribution as the forecast, the multivariate rank and average rank of observation 1 have their verification rank histograms to be uniform/flat. This is an example of the fact that a flat/uniform verification rank histograms do not necessarily mean the observation and the forecast distribution are well calibrated. Likewise the flatness of rank histogram for observation 3 by the multivariate rank may be illusory, this is because observation 3 doesn't follow the same distribution as the ensemble member forecast.

It is interesting how all the four calibration metrics rank histograms interpret observation 5. Observation 5's variance-covariance matrix has no correlation structure. The rank histogram of the minimum spanning tree (MST) rank method for observation 5 is skewed to the right, indicating more outlying observations than inlying as so a lack of correlation in the forecast gives similar results biased or underdispered in the forecast. That of the multivariate rank appears to be somewhat skewed to the right which is also what we would expect from biased forecast ensembles. The rank histograms of average rank method on the other hand shows a pattern consistent with overdispersed, i.e. $\cap$-shape. The verification rank histograms of observation 5 by the average rank and band depth rank method may appear similar but they have different interpretations. The band depth pre-ranks assess the centrality of the observation vector. The more centered the observation vector is in the ensembles, the higher the rank. The $\cap$-shaped histogram by the band depth rank

method indicates that there are less inlying and less outlying observations than would be expected from an forecast.

Overall the mismatch of correlation functions between the forecast and observations 1, 2 and 3 does not result in rank histograms that are strikingly different from uniform. However there are a few noteworthy patterns that emerge. First, for the MST rank method, the slightly triangular shaped histogram for observation 2 means more low ranks was generated by this correlation function than high ranks.

We now take a look at the rank histograms generated by the multivariate rank method. There is a slight increase in high ranks for observation 2, the histogram skewed slightly to the left, this is as a result of the forecast having a weak correlation function. In the case of observation 5, there appears to be more low ranks (*histogram skewed to the right*) than high ranks and therefore we can conclude that the forecast has a very strong correlation. As for observation 1 and 3, they both did not detect more subtle mismatch. Next we take a look at the rank histogram by the average rank method. From figure 3.3, the slightly $\cup$-shaped histogram for observation 2 is by virtue that the forecast has a weak correlation as compared to that of the correlation of observation 2. On the other hand, the slightly $\cap$-shaped histogram for observation 3 is an indication the forecast has a somewhat strong correlation structure. In observation 5, the forecast has a very strong correlation and this is shown by a $\cap$-shaped histogram. Lastly, the slightly $\cup$-shaped histogram by the band depth method for observation 3 and the $\cap$-shape histogram for observation 5, have a similar kind of response as the average rank method.

Case 1.2: Forecast has independent correlation structure/function

Case 1.2 is similar to case 1.1 expect that the forecast distribution in this case, has an independent correlation structure, that is, the correlation structure of the forecast is the identity matrix. The correlation structure of the forecast ensemble is the same as observation 5, that is the variance-covariance matrix is an identity matrix of size 16. The distribution of the forecast is shown in the table 3.2 below:

Observation	Distribution	Mean	Variance
Forecast	$N \sim (\mu, I_{16})$	[0,0,0,...,0]	[1,1,1,...1]

Table 3.2. Setup for Case 1.2. Forecast distribution has the same mean vector and variances as the five observations and the same covariance function as observation 5.

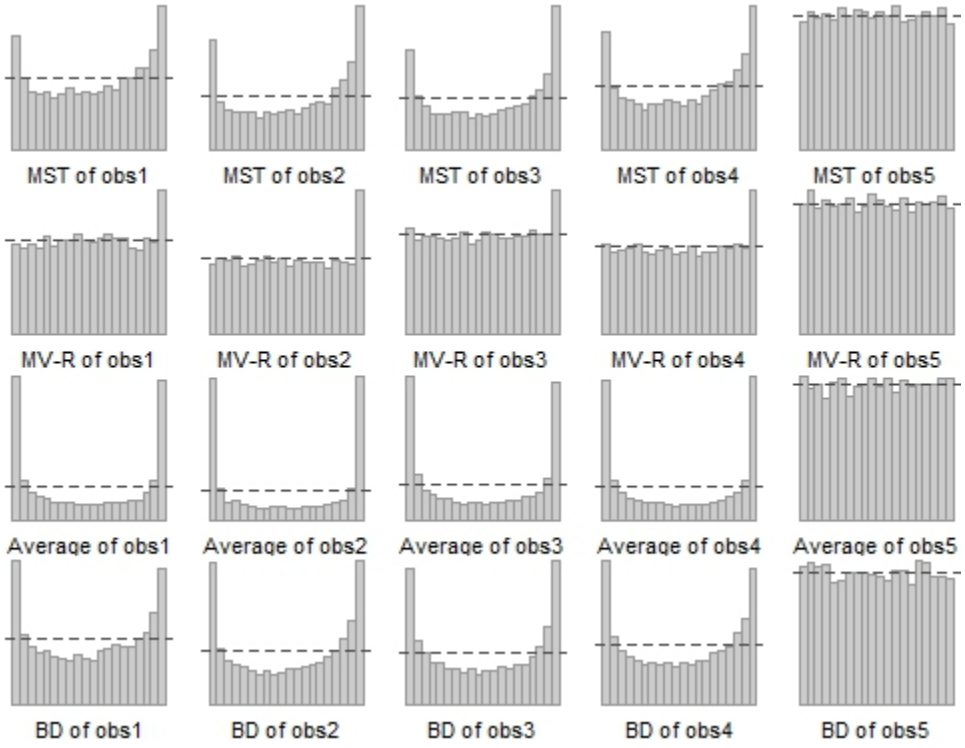

Figure 3.4. Rank histogram for case 1.2. The forecast has a misspecified correlation structure. The first row shows the rank histogram for minimum spanning tree,the second show the multivariate rank, the third average rank and last row show the band depth rank. The five columns show the 5 observations. Note that here the forecast has the same distribution as observation 5. So observation $1-4$ all have stronger correlation structure than the forecast. The results are based on 10,000 repetitions

Observation 5 in figure 3.4 has the forecast ensemble at each sample point to be reliable, in other words, observation 5 has the same distribution as the forecast ensemble that's why the resulting rank histogram is uniform and this was captured by all four ranking methods. The rank histograms for observation $1-4$ in figure 3.4 are extreme versions of what we saw for observation 2 in figure 3.3, which was also a case of forecast having too weak correlation structure. That is, average rank and band depth exhibit a strong $\cup$-shape

and it turns out that MST does as well (*not so clear in figure 3.3*. Furthermore, multivariate rank shows this peculiar pattern of basically even number of ranks, except it has a spike in th enumber of highest rank.

In conclusion, it can be difficult to distinguish between a misspecified correlation function and bias and over/under dispersion of forecasts. However, it seems that a $\cup$ shaped MST and band depth histogram may be an indicator that forecast lacks correlation structure and perhaps a side-ways L shape of multivariate rank can as well.

Case 2: Forecast has a misspecified variance-covariance matrix

In this case, we assume the variance-covariance matrix of the forecast ensemble member was misspecified. The misspecified (wrong) variance-covariance matrix was computed this way. Let's assume the vector S has the same length as the variance-covaraince matrix of the forecast ensemble member which 16 in this cases and contains the variance we want to impose on the 16 elements of a forecast member. However, we also want to maintain the same correlation function. Let's assume the current variance-covariance matrix of size 16×16 is D. In order to change the variance-covariance of the forecast ensemble without changing the correlation structure, we form a new variance-covariance matrix (NV). We represent mathematically below how the new variance-covariance matrix was formed:

Let $S = (s_1, ..., s_{16})$ be a vector of 16 variance values and D be the current variance-covariance of the forecast distribution, then the new variance-covariance matrix NV is given by:

$$NV = (diag(\sqrt{S}))^T * D * diag(\sqrt{S})$$

where $diag(\sqrt{S})$ is a diagonal matrix with the values $\sqrt{s_1}, ..., \sqrt{s_{16}}$ on the diagonal. NV has a dimension of 16×16.

We consider 3 subcases of misspecified variances: For case 2.1, forecast has (mostly) larger variance than observations. In case 2.2, forecast has (mostly) smaller variance than the observations and lastly in case 2.3, half of the forecast vector has larger variance than the observation and the other half has smaller variances.

Case 2.1: Forecast has misspecified variance (big)

In this case our forest is overdispersed. In the univariate setting this would lead to a verification rank histogram with a $\cap$-shape. Since multivariate and average ranks are extension a of univariate orderings we would expect to see a $\cap$-shape in those histograms. Minimum spanning tree and band depth histograms show the effect of overdispersion much clearer (comparing to figure 3.5). They show a left skewed histogram for observations $1-4$ as we would expect. But observation 5 is different, both MST and Band depth show a (*skewed*) $\cap$-shape. Comparing these to the corresponding histograms in figure 3.5 we see that the overdispersion is represented, it just doesn't completely overwrite the effect the effect of misspecified correlation structure(*forecast is correlated, observation is not*). In the cases

of big variances, because the minimum spanning tree and the band depth methods measure the centrality or the outlyingness of the observation curve, we expect the observation curve in this case to be among the inlying curves most of the time. This suggest high ranks are going to be generated and therefore the rank histograms have to skewed to the left. The distribution of the forecast ensemble is given in the table 3.3 below and the verification rank histograms are given in figure 3.5.

We generated the members of vector S randomly from an exponential distribution with a rate of 0.5 (*mean 2*). Most of the variances are bigger than that of the observation (1) but a few were smaller (5 $out \quad of$ 16) than that of the observation curve.

The minimum spanning tree and the band depth rank histograms are more skewed to the left because the observation curves was most of the time more centered, which generated high ranks most of the time and very few ranks.

As a result of most of the variances in this case are bigger as compared to the observation variances, we expect the ensemble members of the forecast to be overdispersed .i.e. the rank histograms should have $\cap$-shaped rank histograms in the case of the average rank and multivariate rank methods.

From figure 3.5, we observe that the rank histograms produced by the average rank method for all the observations aside from observation 2 have a $\cap$-shaped histograms which signifies overdispersion among the forecast ensembles. The slightly $\cup$-shaped histogram for observation 2 might be taken as a sign of underdipersion but recall this histogram in figure 3.3. Observation 2 has stronger correlations than the forecast and that tends to result in

a ∪ shape. Here we added overdispersion to the forecast distribution but its effect on the rank histogram is overshadowed by the wrong correlation function. The multivariate rank histograms are very similar to those in 3.5.

Taking these into account, the minimum spanning tree and the band depth methods are the best methods to detect the misspecification of variance (big) in forecast ensembles.

Observation	Distribution	Mean	Variance
Forecast	$N \sim (\mu, \Sigma)$	[0,0,.....,0]	[1.520, 2.204, 1.608, 0.352, 3.574, 0.349, 2.561, 0.119, 2.768, 4.492, 2.229, 1.165, 0.744, 1.830, 3.171, 0.286]

Table 3.3. Setup for Case 2.1 is a subset of Case 2. Forecast distribution has the same mean vector but bigger variances. The mean vector and variances of the five observations remains unaltered.

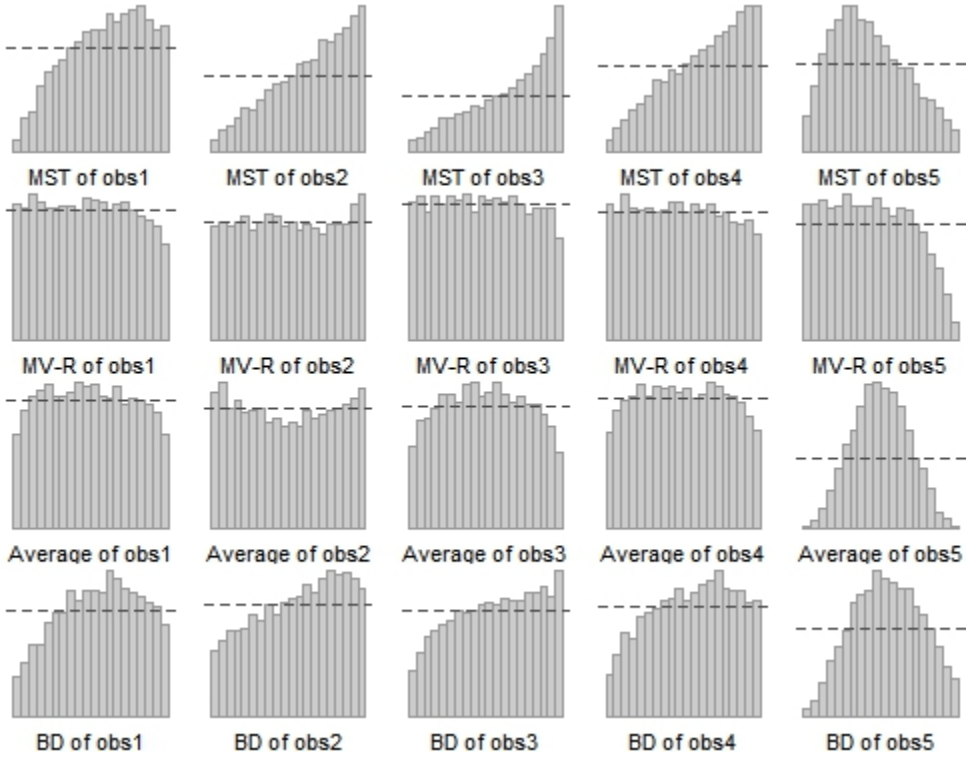

Figure 3.5. Rank histogram for case 2.1. The forecast distribution has misspecified variance (big). The first row shows the rank histogram for minimum spanning tree,the second show the multivariate rank, the third average rank and last row show the band depth rank. The five columns show the 5 observations. The results are based on 10,000 repetitions

Case 2.2: Forecast has misspecified variance(small)

The small variances in case 2.2 were generated randomly from an exponential distribution with a rate of 1.5 (mean 0.667). All the variances smaller than 1, variance of the observation, except for 2 variances which are bigger see table 3.4.

Because the variances are small, we expect the rank histograms generated by the average rank and the multivariate rank method to have ∪-shaped histograms and the ones by the

band depth rank and minimum spanning tree rank method to be right-skewed since we would get more outlying observation (low rank) than inlying (high ranks). All these rank histograms are indication of lack of variability or spread of the ensemble members of the forecast that is the ensemble members of the forecast are underdispered.

Let's take a look at figure 3.6 (*seen below*), we see that this misspecification was perfectly detected by the minimum spanning tree rank and the band depth rank methods for all 5 observations including observation 5 (*has no correlation structure*). The average rank method was able to determine this misspecification for just observations 1 to 4. For observation 5, the average rank method interprets this misspecification as an overdispersion but this also happened in case 1, that is, using correlated forecast for uncorrelated observations gives a ∩-shaped histogram in this case and the amount of underdispersion in case 2.2 is not enough to counter act that. The multivariate rank method again, cannot detect this misspecification. We can conclude from here that, when our forecast ensembles have misspecified variances (*small variances*) the band depth rank and the minimum spanning tree rank methods are the best calibration metrics/ranking methods to detect that.

Observation	Distribution	Mean	Variance
Forecast	$N \sim (\mu, \Sigma)$	[0,0,.....,0]	[0.507, 0.735, 0.536, 0.117, 1.191, 0.116, 0.854, 0.040, 0.923, 1.497, 0.743, 0.388, 0.248, 0.610, 1.057, 0.095]

Table 3.4. Setup for Case 2.1 is a subset of Case 2. Forecast distribution has the same mean vector but smaller variances. The mean vector and variances of the five observations remains unaltered.

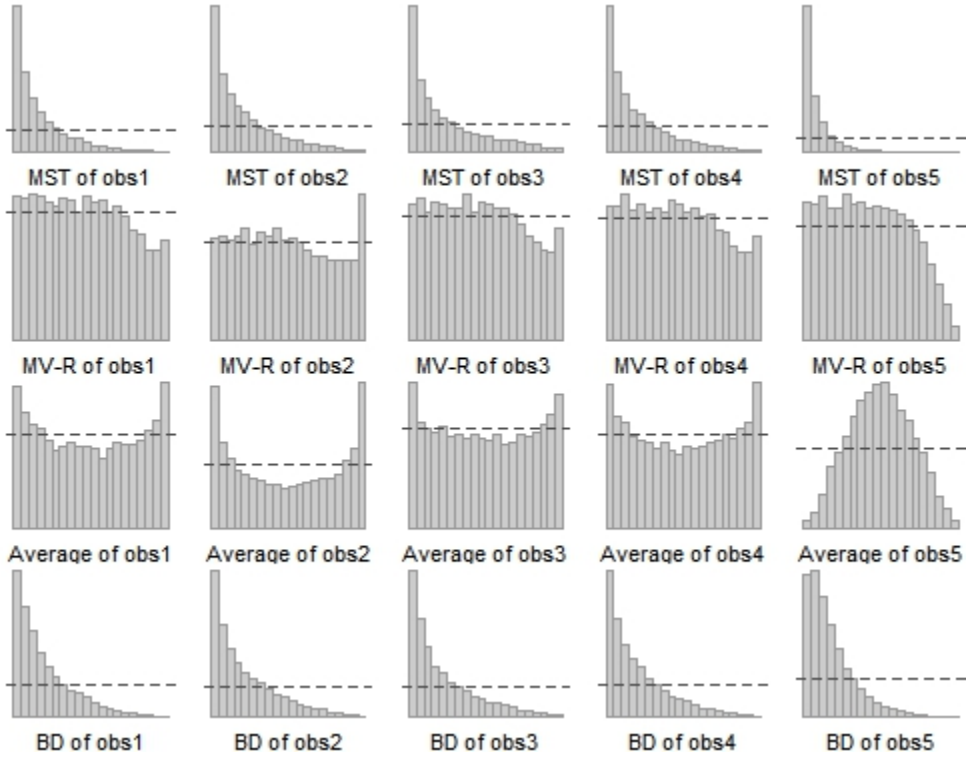

Figure 3.6. Rank histogram for case 2.2. The forecast distribution has misspecified variance (small). The first row shows the rank histogram for minimum spanning tree,the second show the multivariate rank, the third average rank and last row show the band depth rank. The five columns show the 5 observations. Note that here the forecast has the same distribution as observation 4. The results are based on 10,000 repetitions

Case 2.3: Forecast has misspecified variance(half and half)

We randomly generated from an exponential distribution with a rate of 0.5 for the first half of S and the remaining eight 8 values were randomly generated from an exponential distribution with a rate of 1.5. The forecast ensemble with a misspecified variance-covaraince matrix is given in table 3.5 below. The new variance-covariance matrix has equal number of big variances as well as small variances.

Observation	Distribution	Mean	Variance
Forecast	$N \sim (\mu, \Sigma)$	[0,0,.....,0]	[1.520, 2.204, 1.606, 0.352, 3.574, 0.349, 2.561, 0.119, 0.923,1.497, 0.743, 0.388, 0.248, 0.610, 1.057, 0.095]

Table 3.5. Setup for Case 2. Forecast distribution has the same mean vector but different variances. The mean vector and variances of the five observations remains unaltered

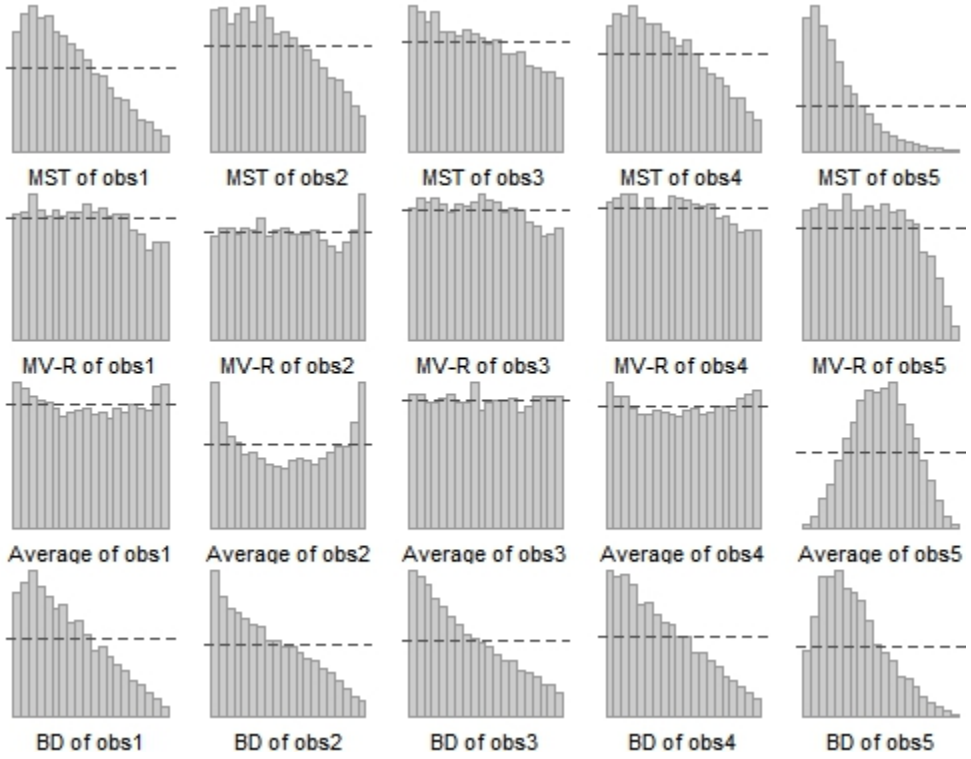

Figure 3.7. Rank histogram for case 2.3. The forecast distribution has misspecified variance (half are bigger and half are smaller). The first row shows the rank histogram for minimum spanning tree,the second show the multivariate rank, the third average rank and last row show the band depth rank. The five columns show the 5 observations. The results are based on 10,000 repetitions

The rank histograms generated by the minimum spanning tree and band depth rank methods for all five (5) observations, as shown figure 3.7, are skewed to the right which indicates the forecast ensembles have too low ranks which means the forecast ensembles are underdispersed. Even though the rank histogram is predominantly skewed to the right we can still tell that there was quite a number of high ranks here too. The multivariate rank histograms especially for observations 1 to 4 appears to be somewhat flat or almost uniform. The average rank histograms for observation 2 indicates underdispersion ($\cup$-shaped rank histogram) in the forecast ensemble and for observation 5, the forecast ensemble members are overdispersed ($\cap$-shaped histogram). The rank histograms for observations 1, 3, and 4 appears to be almost flat or uniform.

Case 3: Forecast with misspecified mean (positive)

The forecast ensemble in this case, has a misspecified mean *positive* of 1 which is a positive bias. The variance-covariance matrix remains unchanged, same as observation 4 (*this variance-covariance matrix is assumed to be the original one in this study*). The distribution of the forecast is shown in the table 3.6 below.

Observation	Distribution	Mean	Variance
Forecast	$N \sim (\mu, \Sigma_4)$	[1,1,1,...,1]	[1,1,1,...,1]

Table 3.6. Setup for Case 3. Forecast distribution has negative mean vector and has variance-covariance matrix as observation 4. The mean vector and variances of the five observations remains unaltered.

From figure (*3.9*), we observe that the verification rank histograms of all the calibration metrics are skewed to the right. Let's go into details of how each ranking method dealt with the misspecified mean.

In the cases of the band depth and minimum spanning tree method, we considered the ensemble member forecast and observation as curves, with the mean of the observation being 0 and that of the ensemble member forecast being 1's, then as shown the figure 3.8 below the observation curve is below the the ensemble member curves all the time.

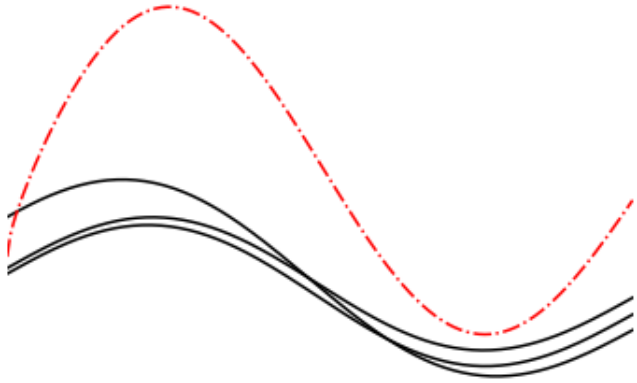

Figure 3.8. The means of the observations are in black and that of the forecast ensemble is in red

This implies the observation curve would be most of time be among the outlying curves and because both the band depth and minimum spanning tree rank methods are about the centrality of the observation curve, in this case, the ranks generated by these ranking methods are going to be low and hence the rank histograms generated by them are expected to be skewed to the right. And this is what can be seen in the figure 3.9. Thus we

conclude that, the band depth and the minimum spanning tree rank methods are able to detect when the forecast distribution has a misspecified mean.

The average rank and the multivariate rank methods were also able to notice this misspecification. Even though the multivariate rank detected this misspecification, its rank histograms did not show it as vividly as the other three (3) ranking methods. This may be because of the high dimensionality ($d = 16$) of the ensemble member forecasts and the observation. The multivariate rank was introduced originally for smaller dimensions (d) and thus it works when the dimension of the ensemble member forecast is small say 2. For the average rank and multivariate rank methods when the forecast distribution has a misspecified mean(*positive*), the rank histograms are skewed to the right which indicates positive bias. In conclusion, all four (4) ranking methods were able to detect that the forecast distribution had a misspecified mean even if the forecast distribution doesn't have a correlation structure (*as evident in observation* 5).

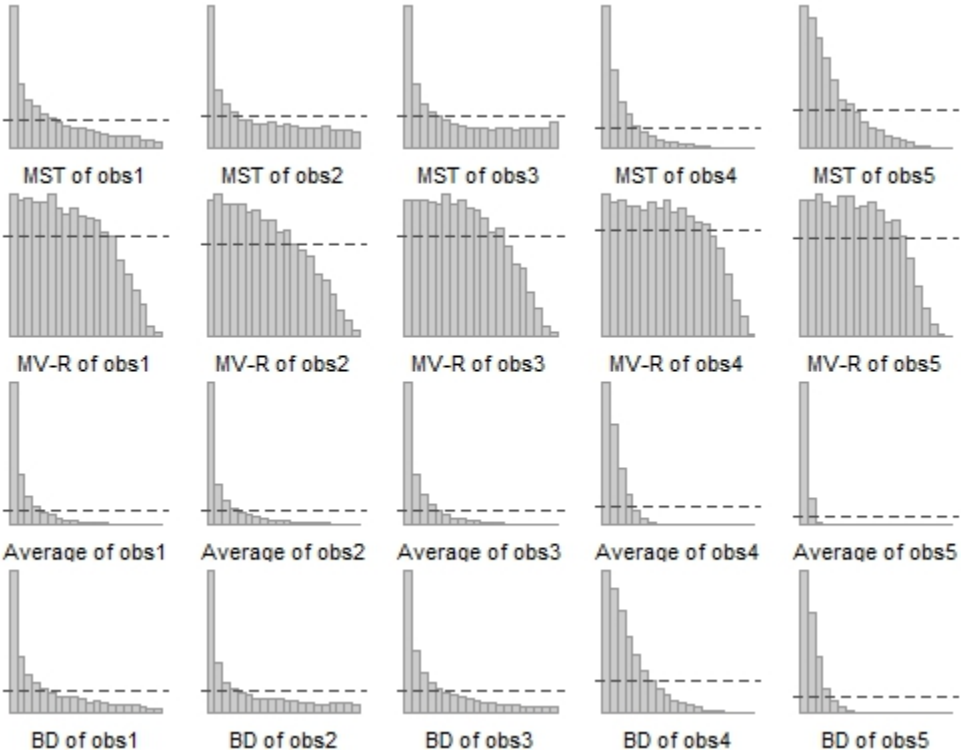

Figure 3.9. Rank histogram for case 3. The forecast distribution has misspecified mean (1). The first row shows the rank histogram for minimum spanning tree,the second show the multivariate rank, the third average rank and last row show the band depth rank. The five columns show the 5 observations. Note that here the forecast has the same distribution as observation 4. The results are based on 10,000 repetitions

Case 4: Forecast with misspecified mean(negative)

The forecast distribution in Case 4 has a negative mean (-0.5), which is a negative bias. The variance-covariance matrix here has a correlated correlation structure/function i.e. it has the same correlation function as observation 4. The distribution of the forecast is given in the table 3.7 below.

Observation	Distribution	Mean	Variance
Forecast	$N \sim (\mu, \Sigma_4)$	[-0.5,-0.5,-0.5,...,-0.5]	[1,1,1,...,1]

Table 3.7. Setup for Case 4. Forecast distribution has positive mean vector and has variance-covariance matrix as observation 4. The mean vector and variances of the five observations remains unaltered.

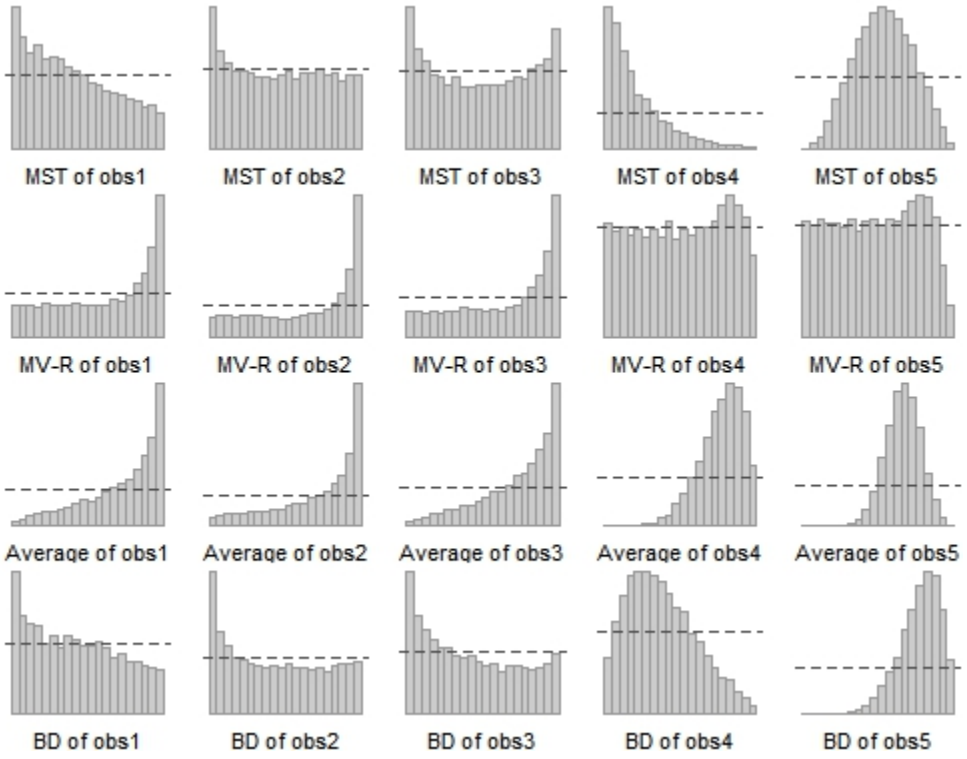

Figure 3.10. Rank histogram for case 4. The forecast distribution has misspecified mean (-0.5). The first row shows the rank histogram for minimum spanning tree,the second show the multivariate rank, the third average rank and last row show the band depth rank. The five columns show the 5 observations. The results are based on 10,000 repetitions

The rank histograms in figure 3.10 are interesting but not surprising. The minimum spanning tree rank and band depth rank have most of their rank histograms similar as in, they

are both skewed to the right. This is not surprising because the band depth rank and minimum spanning tree rank methods provide a center-outward ordering of the curves. When the forecast ensemble member has a negative mean, i.e. a negative bias, we expect the rank histograms generated by the average rank and the multivariate rank methods to be skewed to the left. The average rank method detected this misspecification very well in all five observations. The multivariate rank histogram was able to identify this misspecifications for observations 1 ,2 and 3. As for observations 4 and 5, the rank histograms appear almost flat/uniform which is usually occurs when the forecast ensemble members and observation follows the same distribution. In this situation, we know that is illusory.

The band depth and the minimum spanning tree rank methods on the other hand interprets a misspecified mean (*negative*) of ensemble member forecast differently. We expect the ranks generated by the band depth and minimum spanning tree ranking methods in this case to be low and therefore their histograms should be skewed to the right. The figure 3.11 (below) gives a visual representation of what's going on.

We assume the ensemble member forecast and the observation are curves. From the figure 3.11 above, we see that the observation curve is an outlying curve at all times. We therefore expect the ranks produced in this case to be low and thus the rank histograms should be skewed to the right.

The minimum spanning tree rank method also detected this misspecification for observations 1 and 4 but not so well for other observations. The rank histogram for observation 2 is almost flat which would imply both the observation and ensemble member forecast follow the same distribution which obviously is false. The $\cup$-shaped rank histogram for

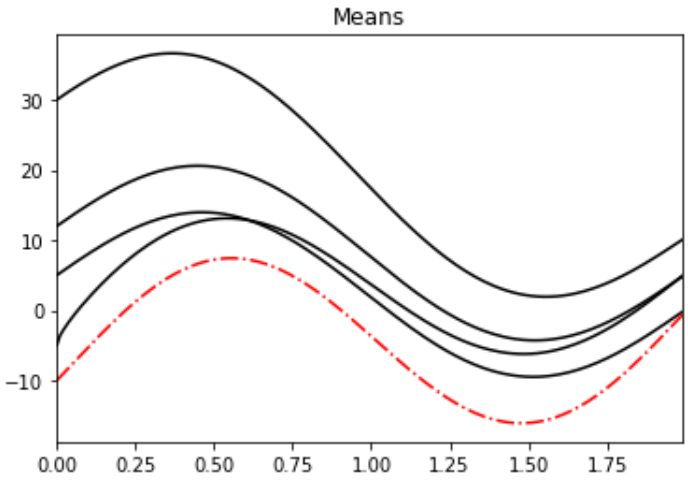

Figure 3.11. The means of the observations are in black and that of the forecast ensemble is in red

observation 3 generated by the minimum spanning tree ranking method implies the forecast ensemble has no correlation structure which is untrue.

The band depth rank method also was able to detect this misspecification for all the observations expect for observation 5. In observation 5, the rank histogram of band depth rank method is skewed to the left which indicates higher ranks, showing an interesting effect of observation 5 having no correlation structure and forecast being biased.

Overall, in conclusion, the average rank was the best ranking method to detect this misspecification even if the forecast ensemble has no correlation structure.

Case 5: Small ensemble size

It is said that miscalibration is generally easier to detect when we have large number of forecast ensemble than fewer forecast ensembles. We put the theory to test in case 6.

We reduce the number of forecast ensemble members from 19 to 5. We test this on the original case. From figure 3.12 we see all the rank histograms generated by all four ranking methods appears to be flat/follows a uniform distribution irrespective of the distributions of the observations. We conclude misspecification by a ranking method is hard to detect when the number of ensembles are small.

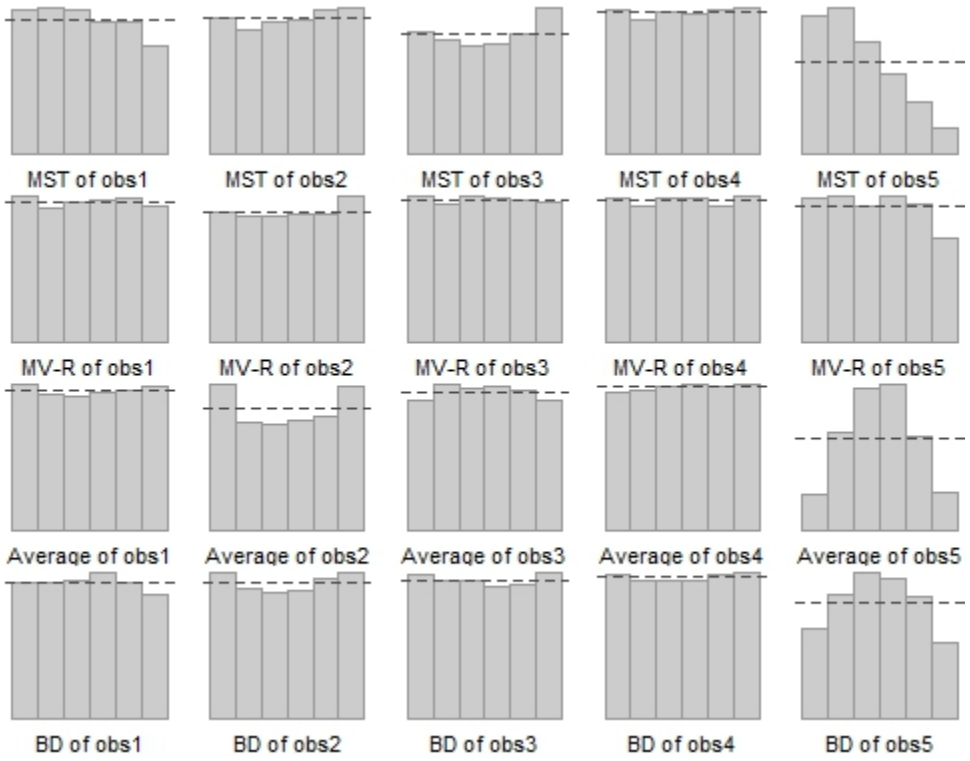

Figure 3.12. Rank histogram for case 6. The number of ensembles is now 5. The forecast has same distribution as case in 1. The first row shows the rank histogram for minimum spanning tree,the second show the multivariate rank, the third average rank and last row show the band depth rank. The five columns show the 5 observations. Note that here the forecast has the same distribution as observation 4. The results are based on 10,000 repetitions

Case 6: Small dimension

In case 6, we would like to know if the dimension has an effect on the calibration metrics. That is we would like to know if miscalibration is easier to detect with bigger or smaller dimensions. We reduce the dimension from 16 to 3 in case 6. These changes were done to the case of misspecified mean (case 4).

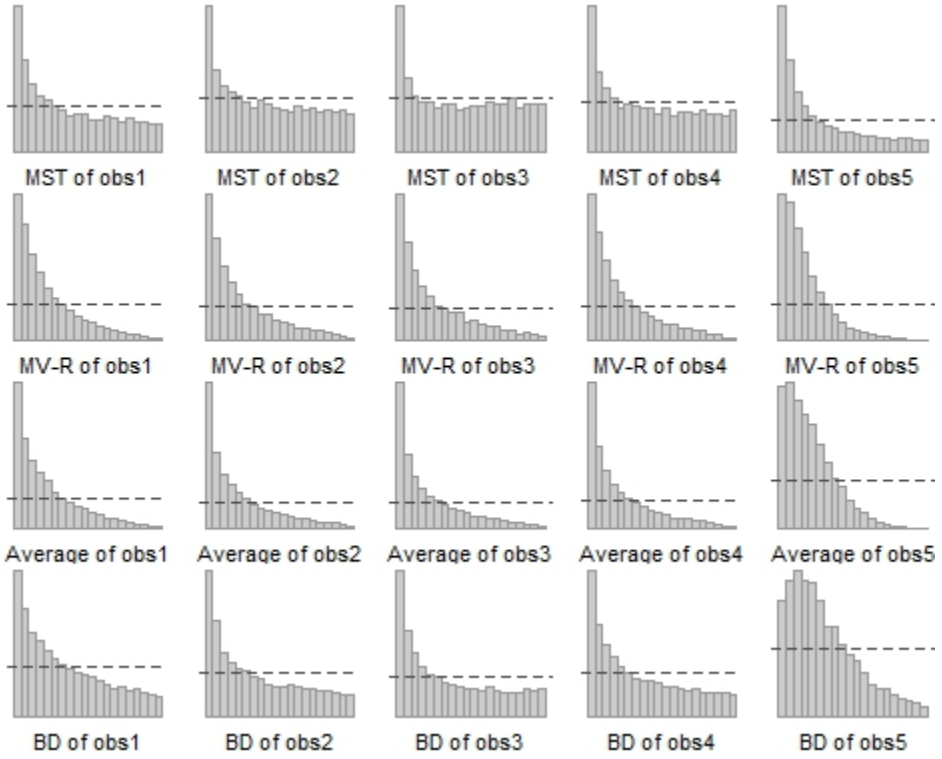

Figure 3.13. Rank histogram for case 7. The forecast has same distribution as in case 4. The first row shows the rank histogram for minimum spanning tree,the second show the multivariate rank, the third average rank and last row show the band depth rank. The five columns show the 5 observations. The dimension is now 5 instead of 16. The results are based on 10,000 repetitions

From figure 3.13, we observe that dimension has no significant effect on the ranking method. The only difference here (*compared to figure 3.11*) is the multivariate rank method was able to detect this misspecification much better when the dimension was smaller.

We have confirmed that the multivariate rank method works better when the dimension is small. Also, reducing the dimension does not have a strong effect on a calibration metrics ability to detect a misspecification.

Ranking Methods Response to Forecast Misspecification

In the following experiments, we summarize the response of the ranking methods to misspecifications in the forecast distribution. The experiments consist of the following setups:

(1) Ensemble forecast has the **same** distribution as observation 4, but different correlation functions from observations $1,2,3$ and 5. (*same as case* 1.1

(2) Ensemble forecast has a **bigger** mean ($\mu = 1.5 * [1,1,\ldots,1]$), but **same** covariance function as observation 4,

(3) Ensemble forecast has **same** mean ($\mu = \mathbf{0}$) as observation 4 but **bigger** (1.5 times) variance-covariance values than observation 4,

(4) Ensemble forecast has **same** distribution as observation 5,

(5) Forecast has **negative mean** ($\mu = [-1,-1,\ldots,-1]$) but has **same** variance-covariance as observation 4.

In the figures below (Figures 3.14 - 3.17), the row 1 corresponds to Experiment 1 above, row 2 to Experiment 2, and so on. The columns correspond to rank histogram for the observations 1 through 5.

We see from rows 2 and 3 that the MVR is able to accurately capture the mispecificaiton in the forecast (from observation 4) when the mean of the forecast is $\mu = 1.5 * [1, 1, \ldots, 1]$, or variance-covariance is increased, and observation 4 has mean $\mu = \mathbf{0}$. Also, when the mean of the forecast is changed to $\mu = [-1, -1, \ldots, -1]$ (row 5), there is a dramatic change in the shape of the histogram, meaning the MVR is able to detect the mispecification with respect to mean.

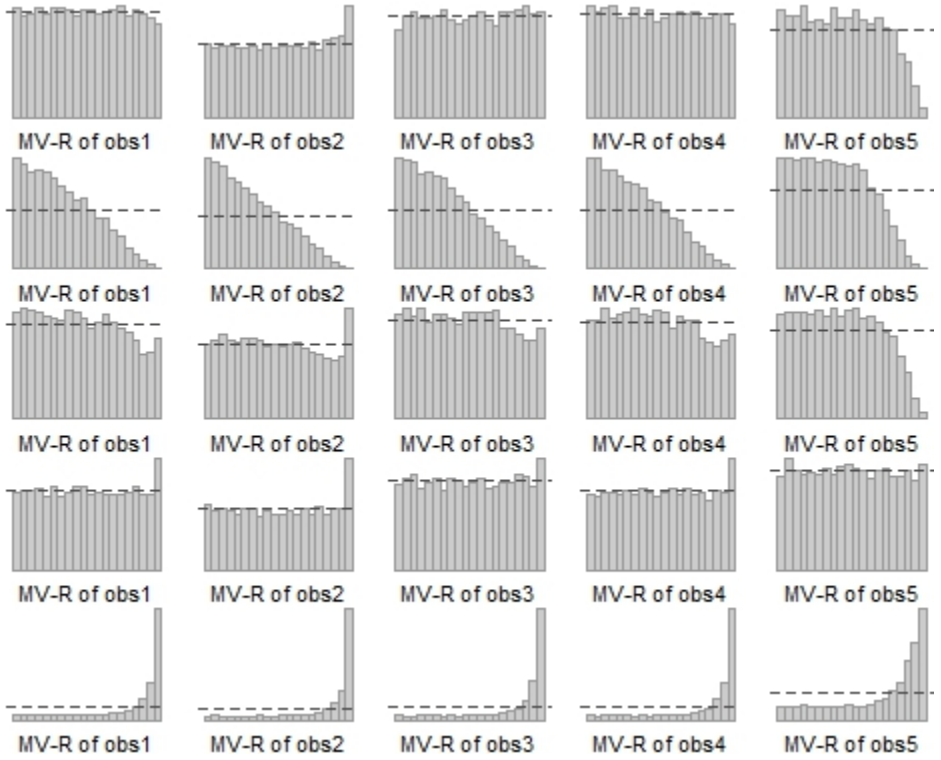

Figure 3.14. This figure shows experiments using the Multivariate Ranking (MVR). Columns $1-5$ show rank histograms for observations $1-5$ as before. First row: Forecast has same distribution as observation 4. Second and last row: Forecast has a positive and negative bias, respectively. Third row: Forecast is overdispersed. Fourth row: Forecast has no correlation.

We see from row 2 that the change in forecast specification not only affects histogram for observation 4, but also the other observations (causing them to have the same shapes.

The BDR detects mispecifications in mean and covariance with the similar rank histograms (rows 2, 3, and 5).

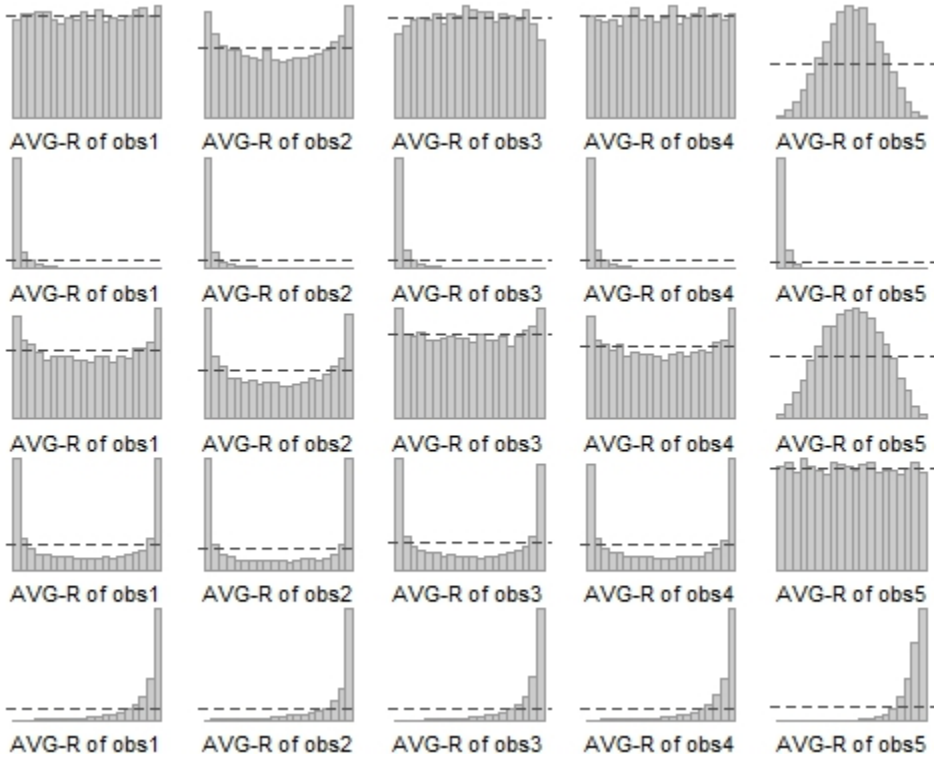

Figure 3.15. This figure shows experiments with Average Ranking (AVR). Columns $1-5$ show rank histograms for observations $1-5$ as before. First row: Forecast has same distribution as observation 4. Second and last row: Forecast has a positive and negative bias, respectively. Third row: Forecast is overdispersed. Fourth row: Forecast has no correlation.

Just like the BDR, the MSTR detects mispecifications in means and covariance with similar shaped rank histograms (rows 2,3, and 5).

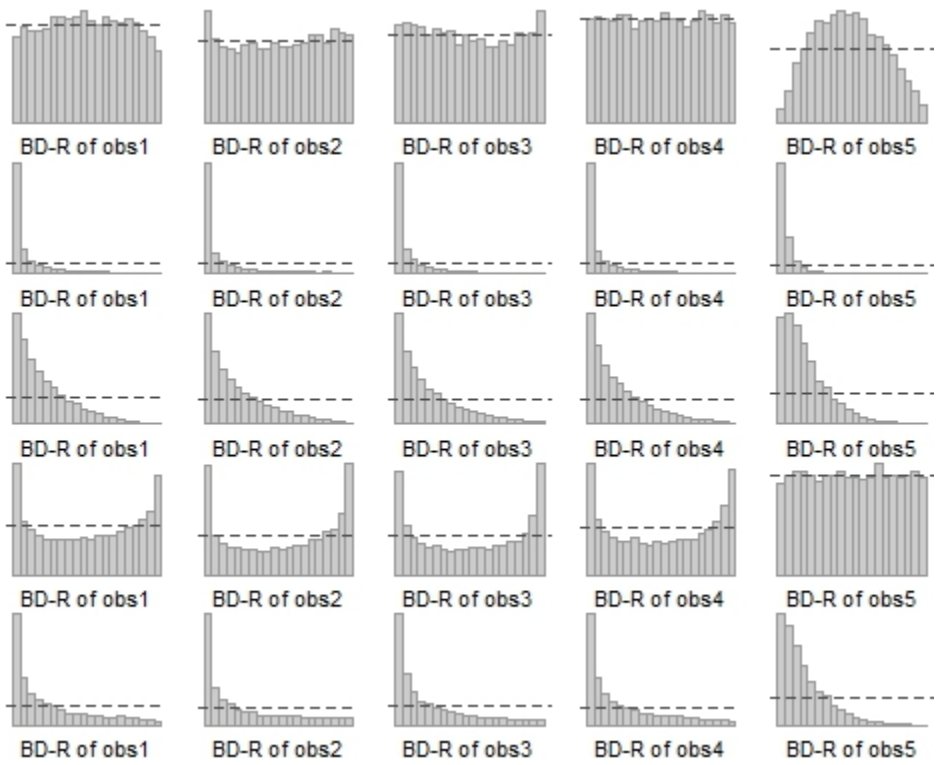

Figure 3.16. This figure shows the Band Depth Rank (BDR) histograms for the five experiments. Columns $1-5$ show rank histograms for observations $1-5$ as before. First row: Forecast has same distribution as observation 4. Second and last row: Forecast has a positive and negative bias, respectively. Third row: Forecast is overdispersed. Fourth row: Forecast has no correlation.

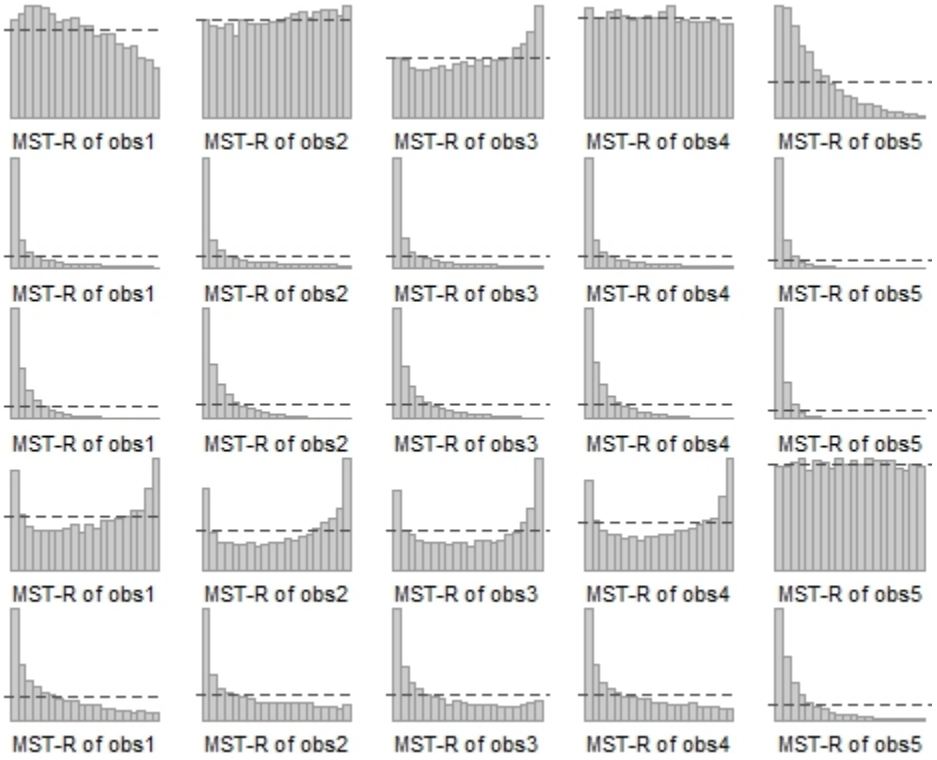

Figure 3.17. This figure shows the five experiments using the Minimum Spanning Tree Rank method. Columns $1-5$ show rank histograms for observations $1-5$ as before. First row: Forecast has same distribution as observation 4. Second and last row: Forecast has a positive and negative bias, respectively. Third row: Forecast is overdispersed. Fourth row: Forecast has no correlation.

4 Case Study: CO_2 Retrievals By OCO-2

4.1 Background on the OCO-2 Instrument and Data used

This case study is a smaller fraction of the work done by Brynjarsdóttir et al, (2018)[14]. The data set and models used in this case study were all from the paper. For the purpose of this study, certain technical details which were not relevant to this thesis such as the background of the OCO-2, how the posterior variances were formed among others have been omitted from this thesis. For more information about these exempted details, look at the paper.

The Orbiting Carbon Observatory-2 (OCO-2) is an instrument on a satellite that collects infrared spectra from which atmospheric properties are retrieved. It measures radiances (i.e. reflected sunlight) in a range of wave-lengths that are known to be affected by CO_2 and O_2 absorption. Let's call the vector of radiances $\mathbf{Y}$. The OCO-2 instrument collects 24 observations every second. Each observation consists of 1016 radiances from three band wavelength and the 3048 dimensional observed vector $\mathbf{Y}$ is called sounding. This

vector of radiances, **Y**, is inverted to an estimate of a state vector **X** that represents the atmospheric conditions at that time and conditions. The concept of inversion is similar to inverse problems where we know the results and we want to deduce the causes. In this case, the vector of radiances **Y** by the OCO-2 is modeled as model:

$$\mathbf{Y} = \mathbf{F}(\mathbf{X}, \mathbf{b}) + \varepsilon \tag{4.1}$$

where **F** is the forward model called the full physics model and **b** is a vector of known constants and ε is the error. We derive the state vector, let's call it **X** from the model above. The state vector **X** contains CO_2 concentrations at 20 pressure levels of the atmospheric column and about 40 other elements such as surface pressure, albedo and aerosol information. For more information about OCO-2, see Eldering et al (2017) [15].

The operational retrievals, i.e. estimation of the state vector **X**, is performed with an algorithm called the Optimal Estimation (OE). The OE algorithm finds the posterior mode of **X** and give an estimate of the whole posterior distribution of **X**. Brynjarsdóttir et al (2018) performed an extensive simulation study and compared results obtained by OE to estimates of the said posterior distributions of **X** obtained via Markov Chain Monte Carlo (MCMC) methods. In this chapter we expand on multivariate diagnostics of MCMC ensembles done in the paper.

Brynjarsdóttir et al (2018) used a surrogate model, $\mathbf{F}^{surr}$ instead of **F** because it is computationally faster than the full physics model. Therefore our model becomes:

$$\mathbf{Y} = \mathbf{F}^{surr}(\mathbf{X}, \mathbf{b}) + \varepsilon \tag{4.2}$$

Furthermore, $\mathbf{X}$ is a 39-dimensional vector which contains the 20 layers of CO_2 concentration, surface pressure, coefficients of 4 different species of aerosals and albedo for three spectral bands. 600 true sate vectors, $\mathbf{X}^{true}$, were simulated in such a way that they represent the variability of physical conditions encountered by OCO-2 between November 2014 and January 2016. Then the radiance vectors were simulated according to equation 4.2 with $\mathbf{X} = \mathbf{X}^{true}$ and $\varepsilon \sim N(0, \Sigma_\varepsilon)$, where Σ_ε is a diagonal matrix.

Samples from the posterior $[\mathbf{X}|\mathbf{Y}]$ were obtained by using the adaptive Metropolis algorithm of Haario et. al (2001)[16]. The adaptive MCMC works like the basic metroplis algorithm, the only difference is that the adaptive metropolis updates the covariance matrix of the proposal distribution along time by employing the information learned. For 600 soundings, four (4) independent chains were ran starting with different initial values of 250000 iterations per chain, making a total of 2400 chains (600×4). Some of the chains, had to terminated due to unsuccessful cholesky factorization and other numerical issues. Out of the 600 MCMC retrievals (soundings), only 457 were regarded to have converged. For each chain, 100,000 iterations out of the 250,000 iterations, were set in as burn-in. Acceptance rates ranged from under 0.5% to 14.5% with a median of 37.2%. The chains were then thinned leaving 6000 MCMC samples (i.e. 1500 for each chain) for inference. For more details about the CO2 retrievals by the OCO-2, see Brynjarsdóttir et al (2018).

In this simulation study of Brynjarsdóttir et al (2018), the true state of the atmosphere $\mathbf{X}^{true}$, is known and we treat them as observations in order to apply the calibration diagnostic tools. We are interested in the estimation of the multivariate state vector $\mathbf{X}$ (39 dimensional) and the CO_2 profile vector $\mathbf{X}_p$ (20 dimensional). The retrievals of the CO_2

by the Markov chain Monte Carlo (MCMC) method are considered as the probabilistic forecast ensembles. In order to determine how well the posteriors by the MCMC method capture the true state vector, we use the minimum spanning tree and the multivariate rank methods as diagnostic tools to determine that. The other two calibration metrics - band depth and average rank methods have already been used in Brynjarsdóttir et al (2018)and are shown in the figure 4.1.

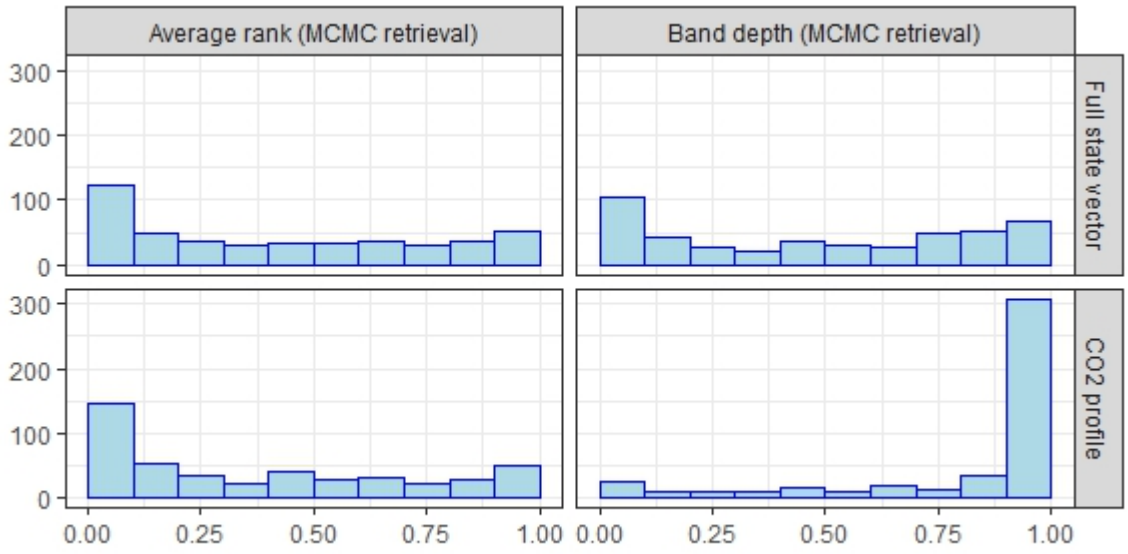

Figure 4.1. Forecasting diagnostics for the full state vector **X** (top row) and the CO_2 profile vector (bottom row). Left: Average rank histograms (MCMC retrieval), Right: Band Depth rank histograms

From figure 4.1, we notice that the rank histograms generated by the band depth method for the full state vector, **X**, (*top right*) is slightly ∪-shaped or right skewed and that of the profile vector, CO_2 element of interest, has its rank histogram (*bottom right*) to be skewed to the left, with much more high ranks than expected. Getting more high ranks than low

is as a results of our observations (true values of parameters) most of the time being in the center of bands defined by the forecast ensembles (MCMC samples from the posterior). This indicates that our forecast ensembles are overdispersed for the profile vector. That of the rank histograms generated by the average rank for both the full state vector, **X**, and CO_2 profile vector, indicates bias in the forecast ensembles.

4.2 Results

Using all 457 converged retrievals, we now consider the retrievals of the 20-dimensional CO_2 profile vector ($\mathbf{X}_p$) and the 39-dimensional full state vector (**X**). The verification rank histograms generated by the multivariate rank and the minimum spanning tree rank method for the MCMC retrieval are shown in the figure 4.2. Calculating the minimum spanning tree ranks for this example is computationally intensive ($\approx$ 10 hours per sample). We utilized high performance computing cluster at both JPL and CWRU to perform these calculations in parallel.

From figure 4.2, we notice that, the rank histograms generated by the minimum spanning tree for both full vector **X** (*top left*) and that of CO_2, element of interest, $\mathbf{X}_p$, (*bottom left*) are somewhat similar. They both show more high ranks, i.e. their rank histograms are skewed to the left. This means our forecast ensembles are overdispersed, too much variance.

The rank histograms by the multivariate method for the full vector **X** (*top right*) and that of CO_2, element of interest, $\mathbf{X}_p$, (*bottom right*)on the other hand, does not tell us much about the forecast ensembles.

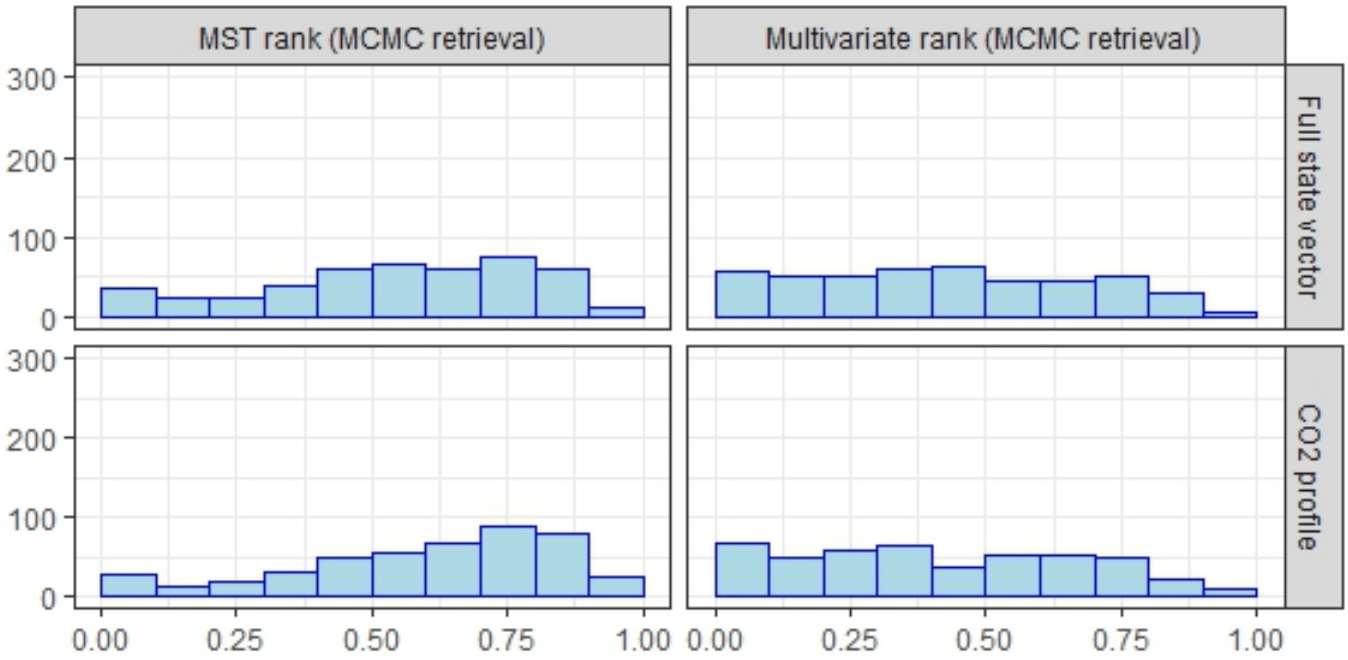

Figure 4.2. Forecasting diagnostics for the full state vector (top row) and the CO2 profile vector. Left: minimum spanning tree rank histograms (MCMC retrieval) and Right: the multivariate rank histograms (MCMC retrieval).

We conclude based on the rank histograms by the average rank method, band depth method and the minimum spanning tree rank method that our forecast ensembles are slightly biased and overdispersed. This means that, the multivariate cases for the marginal posterior distribution of the CO_2 profile vector, $[\mathbf{X}_p|\mathbf{Y}]$ and the posterior distribution of the full vector, $[\mathbf{X}|\mathbf{Y}]$ the true state of the atmospheric ($\mathbf{X}^{true}$) are not perfectly calibrated with.

5 Conclusions

Based on the simulation study and the case study, we can conclude given a verification rank histograms by minimum spanning tree, multivariate, band depth and the average rank methods, we can easily tell the misspecifications. Each method has its own way of indicating misspecification (*through the rank histograms*). For example, a $\cap$-shaped histograms generated by the average rank method would be mean our ensemble forecasts are overdispersive that is to say our probabilistic ensemble forecast are too wide on an average. This same $\cap$-shaped histograms by the band depth rank method, would indicate our probabilistic forecast ensembles have too high correlation. The left-skewed rank histogram by the band depth rank and the minimum spanning tree methods would rather indicate overdipersion of the probabilistic ensemble forecast. A skewed rank histogram generated by an average rank method would indicate a bias in our forecast ensembles. Also $\cup$-shaped rank histogram by the minimum spanning tree and the band depth rank method would suggest a lack of correlation in forecast ensembles whiles the $\cup$-shaped rank histogram by the average rank method would suggest underdispersion among the forecast ensembles.

In conclusion, the multivariate rank method does not seem to detect any of the misspecifications so far. This could be as the result of the high dimensionality of the forecast ensemble members used in this thesis, this is because the multivariate rank method was designed for smaller dimensions. The band depth rank and the minimum spanning tree rank methods are better at detecting misspecification. Also relying on just one calibration metric is not going to be helpful. The best way to assess multivariate forecast ensembles, is to use the average rank method with the band depth rank method and/or minimum spanning tree method simultaneously.

For future work, the forecast ensembles could be generated from other distributions other than a Gaussian.

Complete References

[1] Tilmann Gneiting. Probabilistic forecasting. Journal of the Royal Statistical Society. Series A (Statistics in Society), pages 319–321, 2008.

[2] A Philip Dawid. Probability forecasting. Encyclopedia of statistical sciences, 1986.

[3] 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.

[4] Francis X Diebold, Jinyong Hahn, and Anthony S Tay. Multivariate density forecast evaluation and calibration in financial risk management: high-frequency returns on foreign exchange. Review of Economics and Statistics, 81(4):661–673, 1999.

[5] Thomas M Hamill. Interpretation of rank histograms for verifying ensemble forecasts. Monthly Weather Review, 129(3):550–560, 2001.

[6] 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.

[7] 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.

[8] Claudia Czado, Tilmann Gneiting, and Leonhard Held. Predictive model assessment for count data. Biometrics, 65(4):1254–1261, 2009.

[9] Shiro Ishikawa. Fixed points by a new iteration method. Proceedings of the American Mathematical Society, 44(1):147–150, 1974.

[10] John C Gower and Gavin JS Ross. Minimum spanning trees and single linkage cluster analysis. Applied statistics, pages 54–64, 1969.

[11] Leonard A Smith and James A Hansen. Extending the limits of ensemble forecast verification with the minimum spanning tree. Monthly Weather Review, 132(6):1522–1528, 2004.

[12] Sara López-Pintado and Juan Romo. On the concept of depth for functional data. Journal of the American Statistical Association, 104(486):718–734, 2009.

[13] Ying Sun, Marc G Genton, and Douglas W Nychka. Exact fast computation of band depth for large functional datasets: How quickly can one million curves be ranked? Stat, 1(1):68–74, 2012.

[14] Jenný Brynjarsdóttir, Jonathan Hobbs, Amy Braverman, and Lukas Mandrake. Optimal estimation versus mcmc for co2 retrievals. Journal of Agricultural, Biological, and Environmental Statistics, pages https://doi.org/10.1007/s13253–018–0319–8, 2018.

[15] Annmarie Eldering, Chris W O'Dell, Paul O Wennberg, David Crisp, Michael R Gunson, Camille Viatte, Charles Avis, Amy Braverman, Rebecca Castano, Albert Chang, et al. The orbiting carbon observatory-2: First 18 months of science data products. Atmospheric Measurement Techniques, 10(2):549, 2017.

[16] H Haario, E Saksman, and J Tamminen. An adaptive Metropolis algorithm. Bernoulli, 7(2), 2001.