# Probabilistic forecasts using analogs in the idealized Lorenz96 setting

A diploma thesis submitted to the

DEPARTMENT OF METEOROLOGY AND GEOPHYSICS,

UNIVERSITY OF INNSBRUCK

for the degree of

MASTER OF NATURAL SCIENCE

presented by

JAKOB MESSNER

JUNE 2009

# Abstract

In this thesis, methods of probabilistic weather forecasting using 'analogs' are tested and compared. Using 'analogs' means that states of the atmosphere (in forecast or observation space), similar to the current one, are sought in an archive of past forecasts and observations and utilized for forecasting.

The idealized Lorenz96 model, rather than a far higher-dimensional numerical weather prediction (NWP) model is used for testing these methods.

Three basic methods are presented: The first one, termed *no-forecast-model* method, involves scanning the archive for past observations similar to the current one and using their temporal progress as a forecast. The idea of the second approach (*analogs of a deterministic forecast*) is that similar states in model space correspond to similar states in observation space. The archive is scanned for NWP forecasts analog to the current one and the corresponding observations are extracted. These observations then are assumed to form an improved forecast. For the third method, called *analog dressing*, the same technique as for *analogs of a deterministic forecast* is applied separately to each member of an ensemble forecast.

By taking several 'analogs' respectively, these approaches all provide probabilistic forecasts.

In addition, a raw ensemble forecast and methods that statistically postprocess ensemble forecasts are also tested and compared. These are *ensemble dressing*, *logistic regression*, *nonhomogenous Gaussian regression* and *Bayesian model averaging.*

The approaches using analogs show very promising results in this simple model. For longer lead times, the *analogs of a deterministic forecast* and *analog dressing* approaches even perform best among all tested methods.

# Zusammenfassung

In dieser Arbeit werden Methoden zur probabilistischen Wettervorhersage, die 'Analoge' verwenden, getestet und verglichen. 'Analoge' verwenden, bedeutet, dass zum aktuellen Zustand der Atmosphähre (im Modell- oder Beobachtungsraum), ähnliche in einem Archiv aus alten Vorhersagen und Beobachtungen gesucht und zur Vorhersage genutzt werden.

Zum Testen dieser Verfahren, wird anstatt eines viel höher-dimensionalen numerischen Wettervorhersage (NWP) Modells, das idealisierte Lorenz96 Modell verwendet.

Dabei werden drei grundsätzliche Methoden vorgestellt: Für die erste Methode (*no-forecast-model*) werden analoge Beobachtungen zur aktuellen gesucht, und deren weiterer zeitlicher Verlauf als Vorhersage verwendet. Die Idee des zweiten Verfahrens (*analogs of a deterministic forecast*) ist, dass ähnliche Zustände im Modellraum mit ähnlichen Zuständen im Beobachtungsraum zusammenhängen. Das Archiv wird nach analogen NWP Vorhersagen zur aktuellen durchsucht und die dazugehörigen Beobachtungen dem Archiv entnommen. Diese sollen dann eine verbesserte Vorhersage bilden. Die 3. Methode (*analog dressing*) enspricht dem zweiten Verfahren (*analogs of a deterministic forecast*), angewandt jeweils auf die einzelnen Mitglieder einer Ensemble Vorhersage.

Indem jeweils mehrere 'Analoge' verwendet werden, wird durch alle 3 Verfahren eine probabilistische Vorhersage geliefert.

Zum Vergleich werden weiters eine rohe Ensemble Vorhersage und Methoden, die Ensemble Vorhersagen statistisch nachbearbeiten (Ensemble MOS Methoden), getestet. Dazu gehören *Ensemble Dressing*, *Logistic Regression*, *Non-homogenous Gaussian Regression* und *Bayesian Model Averaging.*

Die Methoden, die Analoge verwenden, liefern dabei in diesem einfachen Modell sehr vielversprechende Ergebnisse. Für längere Vorhersagezeiträume funktionieren das zweite und dritte Verfahren (*analogs of a deterministic forecast* und *analog dressing*) sogar besser als alle anderen getesteten Methoden.

# Contents

**Abstract** i

**Zusammenfassung** iii

**Contents** v

**1 Introduction** 1

1.1 Motivation . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 1

1.2 Outline . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 3

**2 Lorenz96 model and performance measures** 5

2.1 Lorenz96 model . . . . . . . . . . . . . . . . . . . . . . . . . . 5

2.1.1 Model equations . . . . . . . . . . . . . . . . . . . . . . 5

2.1.2 Initialization of ensembles . . . . . . . . . . . . . . . . . 8

2.2 Performance measures . . . . . . . . . . . . . . . . . . . . . . . 11

2.2.1 Brier (skill) score . . . . . . . . . . . . . . . . . . . . . . 11

2.2.2 Ranked probability (skill) score . . . . . . . . . . . . . . 11

2.2.3 Reliability diagram . . . . . . . . . . . . . . . . . . . . . 12

**3 Forecast methods** 17

3.1 Ensemble-MOS approaches . . . . . . . . . . . . . . . . . . . . 17

3.1.1 Direct model output (DMO) . . . . . . . . . . . . . . . . 17

3.1.2 Ensemble dressing . . . . . . . . . . . . . . . . . . . . . 18

3.1.3 Logistic regression . . . . . . . . . . . . . . . . . . . . . 20

3.1.4 Non-homogeneous Gaussian regression (NGR) . . . . . . . . . 21

3.1.5 Bayesian model averaging (BMA) . . . . . . . . . . . . . . 22

3.2 Approaches using analogs . . . . . . . . . . . . . . . . . . . . . . 24
3.2.1 No-forecast-model (NFM) . . . . . . . . . . . . . . . . . . 24
3.2.2 Analogs of a deterministic forecast (ADF) . . . . . . . . . . . 24
3.2.3 analog dressing . . . . . . . . . . . . . . . . . . . . . . . . 25
3.2.4 Analogy criterion . . . . . . . . . . . . . . . . . . . . . . . 27

**4 Results 29**

4.1 'Traditional' ensemble MOS methods . . . . . . . . . . . . . . . . . 29
4.2 Analog methods . . . . . . . . . . . . . . . . . . . . . . . . . . . . 36

**5 Summary and conclusion 47**

**Bibliography 51**

**List of Figures 54**

**Acknowledgments 57**

**Curriculum Vitae 59**

# Chapter 1

# Introduction

## 1.1 Motivation

The common method to predict future weather is to run a numerical weather model. If the present state of the atmosphere is known, future states can be computed with the largely known physical laws of the weather system.

However, due to imperfection of observations, the state of the atmosphere can at most be approximated. In addition, because of limited computational resources, not all processes in the atmosphere can be resolved (e.g. convection), and therefore have to be parametrized. Because of both uncertainties in the initial conditions and imperfection of the model equations, forecasts always exhibit uncertainties. Therefore, for many users, a probabilistic forecast (i.e. predicting a probability distribution), quantifying these uncertainties, is superior to a deterministic one (i.e. forecasting a single value).

The most common way to obtain an outline of these uncertainties is to run an ensemble system. One or several deterministic numerical forecast models are integrated several times with slightly perturbed initial conditions. If the perturbed initial conditions represent the uncertainty of the analysis (i.e. approximation of the state of the atmosphere), it is assumed that after integrating the model forward in time, the different model states represent the uncertainty of the forecast. Large ensemble dispersion (i.e. large differences between the ensemble members) indicates large uncertainty, while a forecast is assumed to be more certain if the ensemble members are similar to each other.

Furthermore, the use of different models (multi-model ensemble) or model equations can take into account model errors.

To obtain a probabilistic forecast from an ensemble, the simplest approach is to take the ensemble relative frequencies as probability density function. However, it is a well known problem of many ensemble systems that they are subject to underdispersion (e.g. Hamill 2001; Wang and Bishop 2005; Bishop 2008). In this case, probabilistic forecasts made with ensemble relative frequencies are overconfident (i.e. the forecasted probability density function is sharper than the density function of truth given the forecast).

As well as improving the ensemble forecasts themselves, statistical postprocessing (model output statistics - MOS) is a good way to achieve better probabilistic forecasts. Therefore, an archive of past ensemble forecasts with a frozen model (reforecast archive) and observations is needed. By utilizing past forecast errors, current forecasts can be corrected. In other words, forecasts made in model space (phase space of the simplified atmosphere, described by the numerical model) are projected into observation space (phase space of the true atmosphere).

Several ensemble MOS approaches have been developed and tested in the past few years. In particular a lot of research has been carried out on ensemble dressing (e.g. Roulston and Smith 2003; Wang and Bishop 2005; Fortin et al. 2006), regression (e.g. Hamill et al. 2004; Gneiting et al. 2005), and Bayesian methods (e.g. Stephens 2005; Raftery et al. 2005; Bishop 2008). Furthermore, some of these methods have been compared in Wilks (2006b, 2007), both, theoretically in a simple model framework (Lorenz96 model - Lorenz 1996) and for a real atmospheric variable (temperature).

In contrast, very simple but quite promising ensemble MOS approaches that use analogs (Hamill et al. 2006; Hamill and Whitaker 2006) are treated rarely in literature. The basic idea of the analog methods of Hamill and Whitaker (2006) is that similar states in model space correspond to similar states in observation space. So if a forecast similar to the current one can be found in the archive, the corresponding observed state of the atmosphere is assumed to be similar to the state to be forecasted.

The aim of this study is to compare these and further methods using analogs with some of the best-working ensemble MOS methods - *ensemble dressing*, *logistic regression*, *nonhomogeneous Gaussian regression* and *Bayesian model averaging* - in the same simple model setting (Lorenz96) as used in Wilks (2006b).

One of the biggest difficulties of analog methods is to find an appropriate criterion for the analogy (i.e. variables, region, height levels, ...). In a simple system like the Lorenz96 model (one resolved variable, 1 dimensional grid), the criterion is clearly much more simple than in the real atmosphere. Thus, the use of such a model to test these methods can reveal their theoretical performance without testing a large number of different criteria.

## 1.2 Outline

In the second chapter, first the Lorenz96 system is described. Subsequently, several scores used to test the different methods are introduced.

In chapter three, both the ensemble MOS methods and the approaches using analogs are presented. Their comparison can be found in chapter four. Chapter five provides a summary and conclusion.

# Chapter 2

# Lorenz96 model and performance measures

In the first part of this chapter, the Lorenz96 model is described. In the second part, several measures to estimate the performance of probabilistic forecast methods are presented

## 2.1 Lorenz96 model

To test the various ensemble MOS methods and approaches using analogs, a forecast and verification dataset is needed. For simplicity, the Lorenz96 model, which shall simulate weather and weather forecasts, is used instead of real weather data. After a basic description of the model, the initialization of ensembles is explained in this section.

### 2.1.1 Model equations

The Lorenz96 model describes a system that consists of two types of variables. A set of a large scale quantity, which is connected to a set of a faster, smaller scale parameter. The 'true' state of the systems is described by the two equations:

$$\frac{dX_k}{dt} = -X_{k-1}(X_{k-2} - X_{k+1}) - X_k + F - \frac{hc}{b} \sum_{j=J(k-1)+1}^{kj} Y_j; \quad k \in \{1, ..., K\} \tag{2.1a}$$

$$\frac{dY_j}{dt} = -cbY_{j+1}(Y_{j+2} - Y_{j-1}) - cY_j + \frac{hc}{b} X_{int_{[(j-1)/J]}+1}; \quad j \in \{1, ..., JK\} \tag{2.1b}$$

The variables $X_{k\in\{1,\dots,K\}}$ and $Y_{j\in\{1,\dots,JK\}}$ are assumed to have cyclic boundary conditions ($X_K = X_0$, $Y_{JK} = Y_0$) and can thus be interpreted as gridpoints (in a one dimensional grid), arranged in a latitudinal circle (Lorenz 2004). Each $X_{k\in\{1,\dots,K\}}$ is connected to J variables with smaller amplitude and frequency ($Y_{j\in\{J(k-1)+1,\dots,kJ\}}$). The system is illustrated schematically in Figure 2.1.

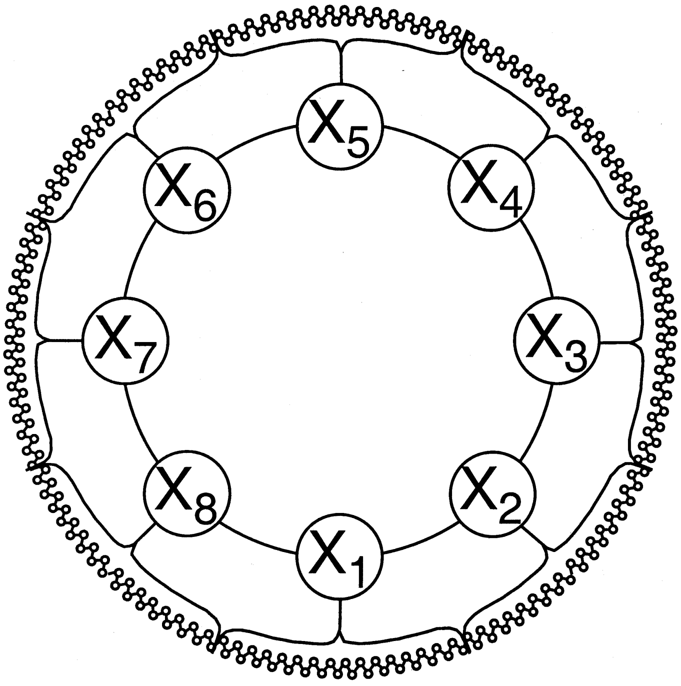


**Figure 2.1:** Schematic illustration of the Lorenz96 system with K=8 large scale variables $X_k$ (large circles) each connected with J=32 smaller scale variables $Y_j$ (small circles) (Wilks 2005)

The smaller scale $Y_j$ can be interpreted as a quantity unresolved in NWP (e.g. convection), while the larger scale $X_k$ represents a parameter that favors the unresolved mechanism (e.g. static instability; Lorenz 1996).

The system simulates advection (quadratic terms), dissipation (linear terms) and external forcing (constant terms) (Lorenz 1996). The last terms on the right sides of Eq. (2.1a) and (2.1b) specify the coupling between the large scale variable $X_k$ and the J small scale variables $Y_j$, which refer to it.

In the literature two different methods to specify 1 'day' can be found: Eq. (2.1a) is scaled such that the quadratic and linear term do not have coefficients. Hence, a time unit has to be regarded as the damping time of the system. This is approximately 5 days for the atmosphere (Lorenz 2004). With this definition, one 'day' equals 0.2 time units.

Wilks (2006a) defined one day as the timespacing with a lag-1 autocorrelation of 0.5 for each of the X variables. This corresponds to 0.15 time units, and a damping time of $6.\dot{6}$ days. In this thesis, the second definition is used.

The coefficients c and b denote that the smaller scale variables have a c times higher frequency, while their amplitude is b times smaller. Furthermore, the parameter h describes the strength of the coupling (Lorenz 1996).

The specific parameter values used in the present study are K=8, J=32, h=1, b=10, c=10 and F=20, as used by Wilks (2005, 2006a).

The equations are advanced using fourth-order Runge-Kutta integration scheme with time step 0.0001. While Eq. (2.1a) and (2.1b) are used to compute the 'true' state of the system, the equations

$$\frac{dX_k^*}{dt} = -X_{k-1}^*(X_{k-2}^* - X_{k+1}^*) - X_k^* + F - \frac{hc}{b}U(X_k^*) \tag{2.2a}$$

$$U(X_k^*) = 0.262 + 1.45X_k^* - 0.0121X_k^{*2} - 0.00713X_k^{*3} + 0.000296X_k^{*4} \tag{2.2b}$$

shall simulate a forecast for the system. The parameterization of the unresolved variable Y (Eq. 2.2b) is obtained through calculating a polynomial regression (eq. 2.2b) of the unresolved tendencies $U(t)$ (Eq. 2.3) as function of $X_k(t)$ (Wilks 2005, 2006a).

$$U(t) = [-X_{k-1}(t)(X_{k-2}(t) - X_{k+1}(t)) - X_k(t) + F] - [\frac{X_k(t+\Delta t) - X_k(t)}{\Delta t}] \tag{2.3}$$

In addition to the fact that the Y variables are parametrized here, Eq. (2.2a) is integrated with Runge-Kutta 2nd order and a timestep of 0.005. Together with a perturbed initialization, the main reasons for forecast errors in operational numerical weather prediction (NWP) are simulated. Consequently, it is quite probable that ensemble forecasts in the Lorenz96 model behave similarly to ensemble forecasts in real NWP, and the results obtained here also have relevance for the true atmosphere.

For the testing, two sets of data are required. First a 'historical database' was created by integrating Eq. (2.1a) and (2.1b), sampling every 0.15 time units, and

thus simulating a sequence of daily analyses (Wilks 2006a). The integration was performed over 10000 'days'. To test the influence of the dataset size, sequences of the first 50, 100, 200, 500, 1500 and 3500 'days' of the database are also used.
Additionally, a verification dataset of size n=10000 was created, again by integrating the model equations sampling 10000 sets of 6 points (T=0,1,..,5), the 6 points spaced by 1 time unit respectively. The sets are separated by 50 time units to ensure that they are independent.
'Climate' is given by the 'historical database' (figure 2.2 shows the 'climatological' distribution of $X_{k\in\{1,...,K\}}$). Probabilistic forecasts are made for six categories, separated by five quantiles of the climatological distribution of the predictand $X_{k\in\{1,...,K\}}$. The specific values of the quantiles are: $q_{1/10} = -2.867$, $q_{2/3} = 1.2886$, $q_{1/2} = 3.5338$, $q_{2/3} = 6.0279$ and $q_{9/10} = 10.9403$.

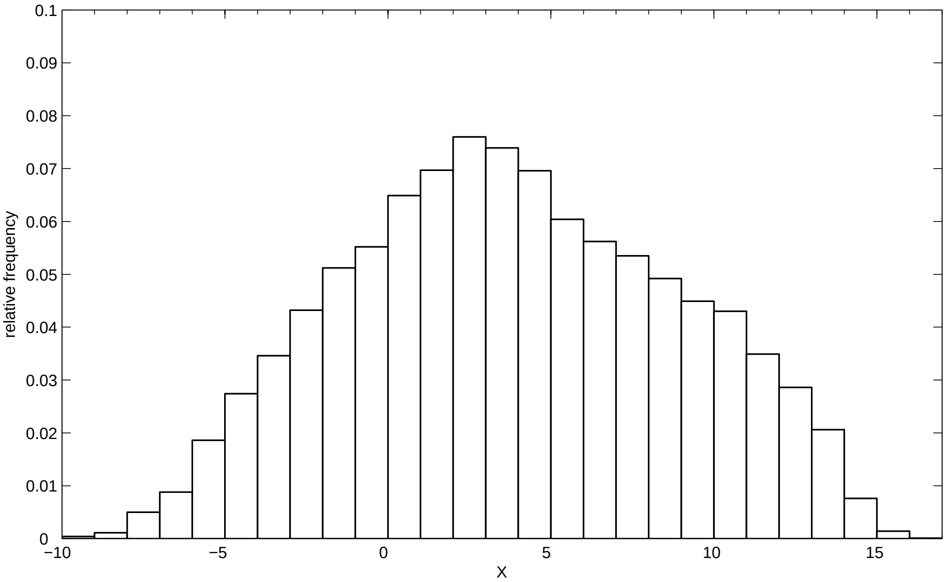


**Figure 2.2:** Histogram of the X values from the training data, illustrating the climatological distribution of the predictand X

### 2.1.2 Initialization of ensembles

According to Anderson (1996), in a good initialization ensemble members need to lie on the attractor of the system (points in phase space that can occur in a system).

Investigation of the structure of the attractor can be very difficult, however extremely long integration of the model can reveal the nature of its attractor (Anderson 1996). Therefore, Eq. (2.1) was integrated through a very large number of timesteps to find at least 100 analogs for all of the 20000 (10000 for the training data and 10000 for the test data) initial points. In this context, analog means that each $X_{k,analog}$ has to lie within an interval of 5% of the climatological range (which is the same for every $X_k$: 1.2) centered on the $X_k$ of the initial point ($X_k - 0.6 < X_{k,analog} < X_k + 0.6$ for all $k = 1, ..., K$) (Wilks 2005). Of these analogs, a (8*8) covariance matrix

$$[S_{local}]_{i,j} = \frac{1}{n_{analogs} - 1} \sum_{i=1}^{n_{analogs}} (X_i - \bar{X}_i)(X_j - \bar{X}_j) \tag{2.4}$$

was computed for each initial point. Moreover, matrices $[S_{init}]$ can be computed through

$$[S_{init}] = \frac{0.05^2 \sigma_{clim}^2}{\frac{1}{8}\sum_{k=1}^{8} \lambda_k} [S_{local}] \tag{2.5}$$

where $\lambda_k$ are the eigenvalues of $[S_{local}]$. These matrices $[S_{init}]$ shall be the covariance matrices of the initial ensemble distribution and have the same correlations and eigenvectors as $[S_{local}]$, but scaled such that the standard deviation in each of the K directions is 5% of the climatological standard deviation ($\sigma_{clim} = 5.07$)(Wilks 2005). To simulate the imperfection of the analysis, the 'true' values of $X_k$ first have to be perturbed. This is realized with:

$$\mathbf{X}_{anal} = \mathbf{X}_{true} + chol([S_{init}])\mathbf{z} \tag{2.6}$$

for all initial points. Here $\mathbf{X}$ denotes a vector with K=8 elements $X_{k \in \{1,..K\}}$, $chol([A])$ the Cholesky decomposition of a matrix $[A]$ ($[A] = [L][L]^T$, $[L] = chol([A])$) and $\mathbf{z}$ a K=8 dimensional vector of independent Gaussian random draws, which is different for each initial point.

Centered on this analysis, ensembles are initialized by:

$$\mathbf{X}_j = \mathbf{X}_{anal} + chol([S_{init}])\mathbf{z}_j, \quad j = 1, ..., n_{ens} \tag{2.7}$$

To create an 'historical' archive of ensemble forecasts (reforecast archive) and a dataset of ensemble forecasts, in order to test the ensemble MOS and analog methods, ensembles of sizes $n_{ens} = 5, 10, 25, 51$ and 100 were initialized for each of the

20000 (10000 for the training data and 10000 for the test data) initial points. Each member was advanced through Eq. (2.2) with a Runge Kutta *2*nd order integration scheme with timestep 0.005. Five day ensemble forecasts were simulated by sampling every 0.15 time units.

Hereafter, these ensemble forecasts are termed dynamical ensemble forecasts, in order to discern them from statistically obtained ensembles.

## 2.2 Performance measures

To estimate and compare the performance of the various methods, several measures are used. These measures are described in this section.

### 2.2.1 Brier (skill) score

To compare the quality of different probabilistic forecast methods, a scalar measure is very convenient. For two-category forecasts (an event can or cannot occur) the *Brier score* (Wilks 2006b) can be used:

$$BS = \frac{1}{n}\sum_{k=1}^{n}(y_k - o_k)^2 \tag{2.8}$$

n is the number of events available for testing (here: 10000), $y_k$ denotes the predicted probability of the k-th event to occur and $o_k$ becomes 0 if the event is not observed, and 1 if it is. Because $0 \leq y_k, o_k \leq 1$, the *Brier score* can only take on values in the range $0 \leq BS \leq 1$ (Wilks 2006b). A perfect forecast attains a *Brier score* of 0.
The *Brier skill score* specifies the performance of one method relative to another:

$$BSS = \frac{BS - BS_{ref}}{0 - BS_{ref}} = 1 - \frac{BS}{BS_{ref}}, \tag{2.9}$$

where $BS_{ref}$ indicates the *Brier score* of the reference method. If $BSS > 0$, the tested method is better than the reference method and if $BSS < 0$, it is worse. In the present study, mostly the climatology (i.e. climatological relative frequencies) is used as reference.

### 2.2.2 Ranked probability (skill) score

Since the forecasts in this study have more than 2 categories, one single *Brier (skill) score* cannot fully characterize a forecast method. To get an overall scalar measure for a method, the *ranked probability score* (RPS) (Wilks 2006b; Epstein 1969) can be used:

$$(RPS)_k = \sum_{m=1}^{J}(Y_m - O_m)^2 \tag{2.10a}$$

$$Y_m = \sum_{j=1}^{m} y_j, \quad m = 1, ..., J; \tag{2.10b}$$

$$O_m = \sum_{j=1}^{m} o_j, \quad m = 1, ..., J; \tag{2.10c}$$

Here J is the number of categories, $y_j$ the probability of the verification to fall in the j-th category and $o_j$ becomes 1 if the observation falls in the j-th category, and 0 otherwise.

The forecast categories that have been used here are $-\infty$ to $q_{1/10}$, $q_{1/10}$ to $q_{1/3}$, $q_{1/3}$ to $q_{1/2}$, $q_{1/2}$ to $q_{2/3}$, $q_{2/3}$ to $q_{9/10}$ and $q_{9/10}$ to $\infty$, where $q$ are the climatological quantiles of the predictand X.

The *ranked probability score* denoted by Eq. (2.10a) only gives a measure for a single forecast event (subscript k denotes the RPS of the $k$th event). For a set of n (here 10000) events, the $(RPS)_k$ are simply averaged:

$$RPS = \frac{1}{n} \sum_{k=1}^{n} (RPS)_k \tag{2.11}$$

For our purposes the advantage of a method over a reference method is sometimes more meaningful (*ranked probability skill score*):

$$RPSS = \frac{RPS - RPS_{ref}}{0 - RPS_{ref}} = 1 - \frac{RPS}{RPS_{ref}}, \tag{2.12}$$

where $RPS_{ref}$ is the *ranked probability score* of the reference method. The climatology is used as reference in this study. Its predicted probabilities of the verification (V) being smaller than a quantile q per definition equal the subscripts of the quantile ($Y_1 = p(V < q_{1/10}) = 1/10$, $Y_2 = y_1 + y_2 = p(V < q_{2/3}) = 2/3$,...)

### 2.2.3 Reliability diagram

Unlike single-number scores, such as the *Brier score* or the *ranked probability score*, the reliability diagram shows the full joint distribution of forecasts and observations (Wilks 2006b). Like the *Brier score* it can only be used for binary predictands (2 categories). It consists of 2 elements, the calibration function and the refinement distribution (Wilks 2006b).

For the calibration function, the conditional probability $p(o|y)$ is plotted against the predicted probability y for the event to occur. The conditional probability is the probability of positive observations (event occurs), given that the predicted probability is y. For a perfect forecast, y and $p(o|y)$ should be equal. The calibration function of a perfect forecast is thus a straight line with slope 1 (slashed line in subsequent figures).

For the practical realization, the probabilities have to be divided into categories. Here I=21 categories are used, with probabilities rounded to the nearest multiple of 0.05. The conditional probability for one category is then assessed through the relative frequency of positive observations for forecasts of this category. See Figure 2.3a for examples of calibration functions.

In the second part of the reliability diagram, the refinement distribution, the frequency of use p(y) of a probability forecast y is shown. Its dispersion can indicate the overall confidence of the forecaster. Because the probabilities of binary predictands are regarded, little dispersion (forecasts frequently near the climatological relative frequencies) reflects low confidence, and forecasts with frequently extreme values (probabilities close to 0 or 1) exhibit high confidence (Wilks 2006b).

Multiplying the calibration with the refinement function leads to the joint distribution of forecasts and observations.

$$p(y_i, o) = p(o|y_i)p(y_i) \tag{2.13}$$

In addition to the two functions, the values of the "Reliability" and "Resolution" terms of the algebraic decomposition of the *Brier Score* (BS; Wilks 2006b; Murphy 1973) are displayed in the reliability diagrams here.

$$BS = \underbrace{\frac{1}{n}\sum_{i=1}^{I} N_i(y_i - \overline{o}_i)^2}_{\text{"Reliability"}} - \underbrace{\frac{1}{n}\sum_{i=1}^{I} N_i(\overline{o}_i - \overline{o})^2}_{\text{"Resolution"}} + \underbrace{\overline{o}(1-\overline{o})}_{\text{"Uncertainty"}} \tag{2.14}$$

Here $N_i$ denotes the number of forecasts predicting $y_i$, $\overline{o}_i = p(o|y_i)$ is the conditional probability of a positive observation, given a forecast y and $\overline{o} = \frac{1}{n}\sum_{i=1}^{I} N_i\overline{o}_i$ is the overall climatological relative frequency of o.

The reliability in Eq. (2.14) is a weighted average of squared vertical distances between the calibration function and the 1:1 reference line (calibration function of a perfect forecast). Because smaller *Brier scores* are better, smaller values of reliability are desirable. The weights $N_i/n = p(y_i)$ (how often $y_i$ was forecasted) are shown in the reliability diagram as refinement distribution. So if the refinement distribution is small, a calibration function that differs to a greater extent from the reference line does not imply a bad reliability. Likewise, for a 'reliable' forecast method, the calibration function has to be closer to the 1:1 line, if the refinement distribution takes on high values (Bröcker and Smith 2007).

The resolution term in Eq. (2.14) is a weighted average of squared differences between the calibration function and the overall climatological relative frequency $\overline{o}$. Since the resolution term is subtracted in Eq. (2.14), higher values are better. Therefore, for a given reliability, steeper calibration functions are desirable. As with reliability, the squared differences are weighted with $p(y_i)$, shown by the refinement distribution.

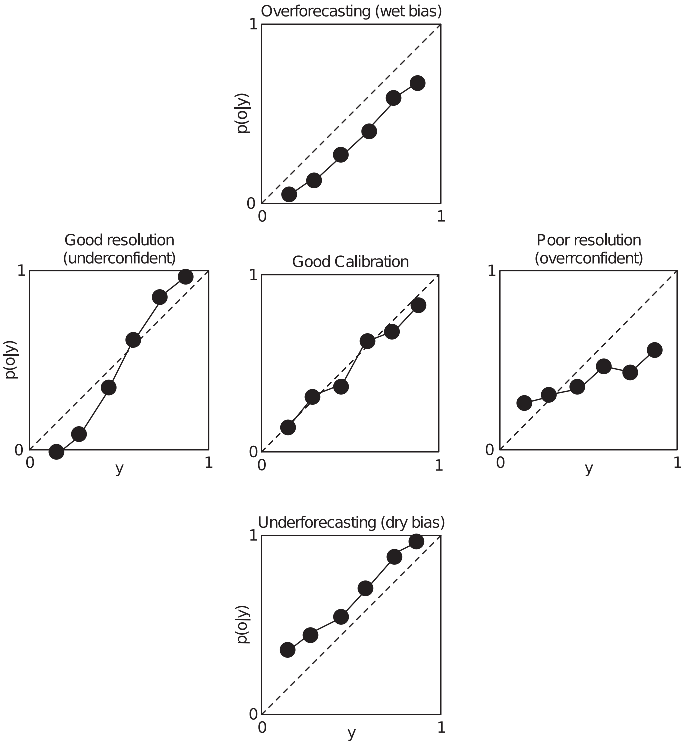
Figure 2.3 shows example calibration and refinement functions.

The center panel of Figure 2.3a depicts a well calibrated forecast. With small differences between the 1:1 line and the calibration function, it obtains small values for the reliability, and thus achieves a good *Brier score.* The top and bottom panels show typical calibration functions of unconditionally biased forecasts. If the predicted probability is frequently too high, the calibration function has the same slope as the 1:1 line but is offset downwards. If, for example, forecasts are made for precipitation, the predicted precipitation is stronger than the actual precipitation (wet bias).

If the predicted probability is frequently too low, the calibration function is offset upwards (bottom panel). In the case of forecasting precipitation, the predicted weather is drier than the actual weather (dry bias).

The right and left panels of Figure 2.3a show characteristic calibration functions of forecasts exhibiting conditional biases. Forecasts with calibration functions similar to the one in the left panel exhibit overforecasting for smaller and underforecasting for larger forecast probabilities. Because the averaged squared difference between the function and the overall climatological frequency is large, the resolution term becomes large (good resolution) and thereby the *Brier score* becomes small. In

(a) Example Calibration Functions



(b) Example Refinement Distributions

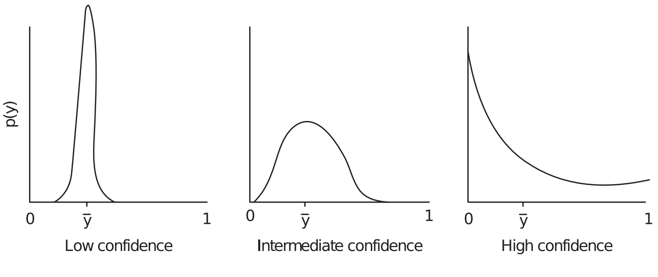


**Figure 2.3:** Example characteristic forms for the two elements of the reliability diagram. (a) calibration function and (b) refinement distribution taken from (Wilks 2006b)

contrast, the averaged squared distance to the 1:1 line is large, which enlarges the reliagibity and thus the *Brier score*.

In the right panel of Figure 2.3a, low forecast probabilities are associated with underforecasting, while high forecast probabilities are associated with overforecasting. For both reasons, because the resolution term is small and the reliability term large, such a forecast reaches bad *Brier scores*.

Typically, a calibration function with a shallower slope than the 1:1 reference line (right panel in Figure 2.3a) is associated with a refinement distribution similar to the one shown in the right panel of Figure 2.3b (high confidence), because extreme probabilities are predicted too often. Conversely, forecasts with a calibration function steeper than the 1:1 line mostly have refinement distributions exhibiting low confidence (left panel of Figure 2.3b).

# Chapter 3

# Forecast methods

In this chapter, several methods of probabilistic forecasting are described. In the first section, the ensemble MOS approaches tested by Wilks (2006a) are presented. Subsequently, the methods using analogs are explained.

## 3.1 Ensemble-MOS approaches

Wilks (2006a) showed that the methods *ensemble dressing*, *logistic regression* and *non-homogeneous Gaussian regression* work best in the Lorenz96 model. A modified new *Bayesian model averaging* method is also tested here and *direct model output* is used as a reference method. These five approaches are described in this section.

### 3.1.1 Direct model output (DMO)

*Direct model output*, as the name suggests, is not a MOS method, however, it is used here as a reference method. A reasonably efficient MOS method should give better results than this approach (Wilks 2006a).

It is assumed that the distribution of the ensemble reflects the probability distribution of the predictand. The simplest approaches for computing the cumulative distribution function are:

$$Pr(V \leq q) = \frac{Rank(q)}{(n_{ens} + 1)} \tag{3.1}$$

$$Pr(V \leq q) = \frac{Rank(q) - 1}{(n_{ens})} \tag{3.2}$$

$Pr(V \leq q)$ is the probability of the verification $V$ to be smaller or equal the quantile $q$, and $Rank(q)$ specifies the rank of the quantile q in the ensemble ($Rank(q) = 1$ if all ensemble members are greater and $Rank(q) = n_{ens} + 1$ if all ensemble members are smaller than the quantile $q$)

Note that in Eq. (3.1), $Pr(V \leq q) = 1$ if all $n_{ens}$ members are smaller than $q$ and that in Eq. (3.2) additionally $Pr(V \leq q) = 0$ if all members are greater than $q$. Since these extreme probabilities are not desired, different approaches are needed. Several estimates for constructing cumulative frequency distributions, known as plotting positions, are described in Wilks (2006b). Here, the so called Tukey plotting position is used

$$Pr(V \leq q) = \frac{Rank(q) - 1/3}{(n_{ens} + 1) + 1/3}, \tag{3.3}$$

which approximates the median of the sampling distribution (Wilks 2006b). Thus, the cumulative probability can take on values between $2/(3n_{ens} + 4)$ and $(3n_{ens} + 2)/(3n_{ens} + 4)$.

### 3.1.2 Ensemble dressing

In the *ensemble dressing* approaches of Roulston and Smith (2003) and Wang and Bishop (2005), a statistically-derived daughter ensemble is applied to each member of a dynamical ensemble forecast. The probability distribution of the resulting greater ensemble should better estimate the distribution of truth given the forecast than the raw dynamical ensemble.

The daughter ensemble is obtained from historical forecast errors. In Roulston and Smith (2003), errors of the best members (members that are closest to the verification in phase space) are used to dress the ensemble. After the best members are identified in the historical database, their errors are resampled to obtain a daughter ensemble for each ensemble member. Normally, only best member errors, of one member itself, are used to dress this. However, in ensembles where no member is superior to any other (like in the Lorenz96 ensemble [1] ), best member errors of all

---

[1] the first member is slightly more likely to be the best member, because it is initialized with the analysis (i.e. best approximation of the current state). However, the difference is marginal and therefore disregarded.

members can be applied to them.

Wang and Bishop (2005) argued, that this best member approach of Roulston and Smith (2003) does not guarantee a 'reliable' (Wilks 2006b) ensemble (i.e. it is not mathematically constrained to be drawn from the same distribution as the verification). Hence, they developed an approach that forces the variance of the dressed ensemble to be the same as the error variance of the ensemble mean in the training data.

In contrast to Wang and Bishop (2005), instead of dressing the ensemble with daughter ensembles, Wilks (2006a) proposed using Gaussian dressing distributions (kernels) and averaging them. Using Gaussian kernels is appropriate for the use of the Lorenz96 model, however for the real atmosphere, some variables (e.g. precipitation) might need different distributions. The variance of the Gaussian distributions is given by:

$$\sigma_D^2 = \sigma_{\overline{X}-V}^2 - (1 + \frac{1}{n_{ens}})\overline{\sigma}_{ens}^2, \tag{3.4}$$

where the first term on the right-hand side is the historical error variance of the (debiased) ensemble mean, and the second term denotes the ensemble variance. The factor $(1+\frac{1}{n_{ens}})$ is needed to account for the fact that the ensemble mean in the training data has been debiased (Wang and Bishop 2005). The debiasing is realized through:

$$\tilde{X}_i^t = a_t + b_t X_i^t, \tag{3.5}$$

where the parameters $a_t$ and $b_t$ minimize the function

$$\sum_{i=1}^{n} \sum_{j=1}^{n_{ens}} \sum_{k=1}^{K} (a_t + b_t X_{i,j,k}^t - V_{i,k}^t)^2 \tag{3.6}$$

for the training data (n=training data size, K=number of 'gridpoints'). $a_t$ and $b_t$ are computed separately for each timestep (t=1,2,..,5) and are equal for all members.

The cumulative distribution function of the forecast can then be computed, dressing all ensemble members with a Gaussian kernel with variance $\sigma_D$ and averaging them:

$$Pr(V \leq q) = \frac{1}{n_{ens}} \sum_{i=1}^{n_{ens}} \Phi[\frac{q - \tilde{x}_i}{\sigma_D}] \tag{3.7}$$

where $\Phi[\bullet]$ is the standard Gaussian cumulative distribution function

$$\Phi[h] = \frac{1}{2\pi} \int_{-\infty}^{h} exp(-\frac{t^2}{2})dt \tag{3.8}$$

and $\tilde{x}_i$ the debiased value of the $i$th ensemble member.

The approaches of Wang and Bishop (2005) and Roulston and Smith (2003) are limited to underdispersive ensembles, because by dressing all members symmetrically, they increase the dispersion in all cases. Because underdispersion is a common problem of current dynamical ensemble systems (Wang and Bishop 2005; Bishop 2008), including the Lorenz96 model used in this study, this is for most cases not a limitation.

In the approach of Fortin et al. (2006), the members are dressed and weighted differently, dependent on their rank. This makes their method also usable for overdispersive ensembles (Fortin, Favre, and Said 2006). However, due to the underdispersion of the Lorenz96 ensemble system, this enhancement did not display better results, so the simpler approach of Wang and Bishop (2005) is used here

### 3.1.3 Logistic regression

Another method to postprocess a dynamical ensemble forecast is *logistic regression.* With this, the probability of a binary predictand (2 categories) falling into one category can be estimated (Wilks 2006b). Hamill et al. (2004) used only the ensemble mean as predictor, because they found the correlation between ensemble mean error and ensemble spread to be low in their data. However, for the Lorenz96 system this spread skill relationship is sufficiently strong (Wilks 2006a), so the use of the ensemble standard deviation as a second predictor increases the forecast accuracy of this method. According to Wilks (2006b and 2006a), the distribution function is given by:

$$Pr(V \leq q) = \frac{exp(b_0 + b_1\overline{x}_{ens} + b_2\overline{\sigma}_{ens})}{1 + exp(b_0 + b_1\overline{x}_{ens} + b_2\overline{\sigma}_{ens})} \tag{3.9}$$

The parameters $b_0$, $b_1$ and $b_2$ are computed, maximizing the log likelihood function

$$ln(\Lambda) = \sum_{i=1}^{n} \{I(V_i \leq q)[b_0 + b_1\overline{x}_{ens,i} + b_2\overline{\sigma}_{ens,i}] - ln[1 + exp(b_0 + b_1\overline{x}_{ens,i} + b_2\overline{\sigma}_{ens,i})], \tag{3.10}$$

where n is the number of training events and the indicator function I(•)=1 if the argument is true and I(•)=0 if it is not. Eq. (3.10) was maximized iteratively using the simplex search method (Lagarias, Reeds, Wright, and Wright 1998), as implemented in Matlab® 2007a. The regression parameters are fitted separately for each quantile and lead time.

### 3.1.4 Non-homogeneous Gaussian regression (NGR)

Gneiting et al. (2005) presented a different regression method for predictands that are normal distributed. A predictand $v$ can be estimated using linear regression

$$v = a + b\overline{x}_{ens} + \epsilon \tag{3.11}$$

where $\epsilon$ is an error term that averages to zero. Inversely, once the regression coefficients $a$ and $b$ are estimated using historical data, a probabilistic forecast can be made, setting $\epsilon$ as a Gaussian distribution. To account for the spread skill relationship, the variance of this Gaussian distribution can be estimated by

$$\sigma_\epsilon^2 = c + d\sigma_{ens}^2 \tag{3.12}$$

(Gneiting, Raftery, Westveld, and Goldman 2005). This estimation is certainly only appropriate if the error term $\epsilon$ is distributed normally .

As proposed by Gneiting et al. (2005), the regression parameters a,b,c and d have been derived by minimizing the continuous ranked probability score (CRPS) (Hersbach 2000; Matheson and Winkler 1976; Wilks 2006b)

$$\overline{CRPS} = \frac{1}{n}\sum_{i=1}^{n}(c + d\sigma_{ens,i}^2)^{1/2}(z_i[2\Phi(z_i) - 1] + 2\phi(z_i) - \pi^{-1/2}) \tag{3.13a}$$

$$z_i = \frac{v_i - (a + b\overline{x}_{ens,i})}{(c + d\sigma_{ens,i}^2)^{1/2}}, \tag{3.13b}$$

where $\Phi(\bullet)$ and $\phi(\bullet)$ are the cumulative distribution function (CDF) and the probability density function (PDF) of the normal distribution, respectively. Eq. (3.13) is minimized iteratively for each timestep with the same regression parameters a through d for each quantile.

With the parameters a,b,c and d, the distribution function is given by

$$Pr(V \leq q) = \Phi[\frac{q - (a + b\overline{x}_{ens})}{(c + d\sigma_{ens,i}^2)^{1/2}}]. \tag{3.14}$$

### 3.1.5 Bayesian model averaging (BMA)

*Bayesian model averaging* (BMA) has already been used for a long time for statistic models (Leamer 1978; Hoeting et al. 1999). Raftery et al. (2005) were the first who proposed also using BMA for dynamical models. Like in *ensemble dressing*, kernel distributions are generated around each member of the ensemble. The main difference is that the dressing distributions are different for each member. Additionally, the different members are weighted and debiased separately, which is quite reasonable, especially for multi model ensembles. Like the *ensemble dressing* method used here, BMA is only appropriate for underdispersive ensembles.

First, each member is debiased separately and dressed with its own kernel distribution, using its historical errors for the occasions where it was the best member. Then the probability of each member being the best member is obtained once again using historical data. These probabilities of being the best of all members add up to one, and can thus be seen as weights ($w_{k\in\{1,\dots,n_{ens}\}}$). The probability distribution of truth given a forecast is then expressed by:

$$\rho(v|x_1,\dots,x_{n_{ens}}) = \sum_{k=1}^{n_{ens}} w_k \rho_D(v;\tilde{x}_k,\sigma_k^2), \tag{3.15}$$

where $\rho_D(\bullet;\tilde{x}_k,\sigma_k^2)$ denotes the dressing distribution with mean $\tilde{x}_k$ and standard deviation $\sigma_k$. For the Lorenz96 model, a Gaussian distribution can be used as dressing function. The tilde again denotes that the value has been debiased through Eq. (3.5)

Bishop (2008) found that both the BMA approach of Raftery et al. (2005) and the dressing methods (Wang and Bishop 2005; Roulston and Smith 2003; Fortin et al. 2006) overestimate the probability of climatologically extreme events, by increasing the spread of the ensemble symmetrically in every forecast. They argued that this failure follows the incorrect assumption that the raw ensemble forecast estimates a distribution of truth given the forecast.

In the following, they discovered that the distribution obtained with Eq. (3.15) is a better estimator for the distribution of the ensemble mean, provided the verification equals the current ensemble mean forecast.

$$\rho(\overline{\tilde{x}}|v = \overline{\tilde{x}}_c) = \sum_{k=1}^{n_{ens}} w_k \rho_D(\overline{\tilde{x}}; \tilde{x}_k, \sigma_k^2) \tag{3.16}$$

With Bayes' theorem

$$\rho(v|\overline{\tilde{x}}_c) = \frac{\rho(\overline{\tilde{x}}_c)\rho(v)}{\rho(\overline{\tilde{x}})} \tag{3.17}$$

and some conversion, the probability of the verification being smaller the quantile q can be computed by

$$Pr(V \leq q) = \sum_{k=1}^{K} w_k^a \Phi[\frac{q - \mu_k^a}{\sigma_{ak}}] \tag{3.18a}$$

$$\sigma_{ak}^2 = (\frac{1}{\sigma_k^2} + \frac{1}{\sigma_{clim}^2})^{-1} \tag{3.18b}$$

$$\mu_k^a = (\frac{\tilde{g}_k}{\sigma_k^2} + \frac{\mu_{clim}}{\sigma_{clim}^2}) \tag{3.18c}$$

$$\tilde{g}_k = \overline{\tilde{x}}_c - (\tilde{x}_k - \overline{\tilde{x}}_c) \tag{3.18d}$$

$$w_a^k = \frac{w_k \phi[\frac{\tilde{g}_k - \mu_{clim}}{\sqrt{\sigma_k^2 + \sigma_{clim}^2}}]}{\sum_{k=1}^{K} w_k \phi[\frac{\tilde{g}_k - \mu_{clim}}{\sqrt{\sigma_k^2 + \sigma_{clim}^2}}]} \tag{3.18e}$$

Eq. (3.18a) appears quite complicated. However, it simply gives higher weights to members that are closer to the climatological mean. Thus, the resulting probability distribution is shifted towards the mean of the climatological distribution. Thereby, the probability of climatological extreme events is reduced to a realistic extent.

Because in the Lorenz96 ensemble all members are computed with the same model equations, all members basically have equal likelyhood of being the best member. However, because the first member is initialized with the analysis (i.e. best approximation of the current state), it is assumed to have a slightly higher probability of being the best member. Hence, it should be weighted more strongly than the other members. Thus, the first member is weighted with its relative frequency of being the best member in the training data ($w_1$), and the weights for the other members are equally $(w_k)_{k \in \{2,\ldots,n_{ens}\}} = (1 - w_1)/(n_{ens} - 1)$.

## 3.2 Approaches using analogs

In this section, 3 basic methods of forecasting using analogs are first described. Because the quantification of similarity is an issue for all of these methods, the used analogy criteria are subsequently discussed.

### 3.2.1 No-forecast-model (NFM)

As the name suggests, this very simple approach provides a probabilistic forecast without the use of any numerical forecast model. It is assumed that similar patterns of specific meteorological variables (e.g. geopotential height, temperature, ...) progress similarly in time.

To obtain a forecast for an arbitrary lead time T=t, a historical database is scanned for synoptic situations (analyses) analog to the one at the initial time (T=0). The $n_{analogs}$ analyses that are the most similar to the current analysis are taken and their progress in time is extracted. Finally, for each lead time a set of $n_{analogs}$ situations that are likely to recur is available. This set can be interpreted as an ensemble with $n_{analogs}$ members, and thus with Eq. (3.3) a probabilistic forecast can be computed. A schematic illustration of this method is shown in Figure 3.1.

Note that in addition to the current analysis (T=0), it is also possible to take account of previous analyses (e.g. T=-1). However, this did not improve results for the Lorenz96 model, so it will not be further pursued here.

### 3.2.2 Analogs of a deterministic forecast (ADF)

For this method, first proposed by Hamill et al. (2006), a current +t deterministic forecast is compared with all +t forecasts in the reforecast-archive. The $n_{analogs}$ historical dates with the most similar reforecasts are extracted ($D_{j\in\{1,\ldots,n_{analogs}\}}$). The corresponding recorded analyses ($D_j + t$) then form an ensemble, which can be shaped into a probabilistic forecast using Eq. (3.3).

As an example, imagine that the deterministic forecast made on the 13th of June 1999 ($D_1$) for the 14th (t=24h) is very similar to the current forecast for tomorrow. Then the analysis of the 14th of June 1999 (23th + 24h) is taken as one member. If

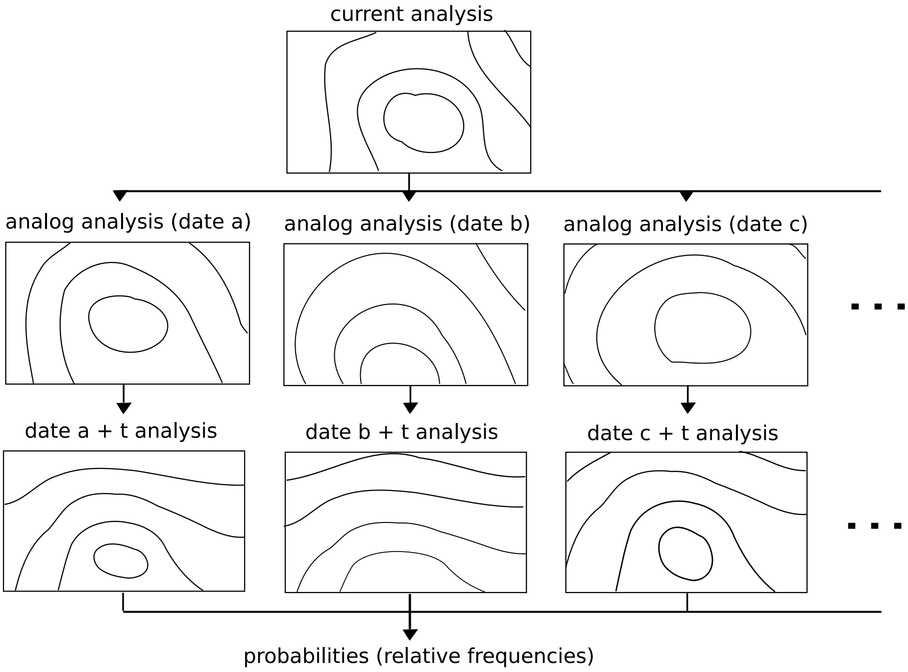


**Figure 3.1:** Schematic illustration of the *no-forecast-model* method

several similar forecasts are found, the corresponding analyses form an ensemble that with Eq. (3.3) provide a probabilistic forecast. Figure 3.2 illustrates this method graphically.

In this study, the ensemble mean is used as the deterministic forecast.

### 3.2.3 analog dressing

As suggested by Roulston et al. (2003), a combination of a dynamical and a statistical ensemble unites the advantages of both. Because in fact the *analogs of a deterministic forecast* approach uses a statistical ensemble, dressing each ensemble member with this method is quite reasonable. Hamill and Whitaker (2006) already tested this method for precipitation forecasts over the USA.

For a current +t ensemble forecast ($n_{ens}$ members), a meteorological archive is scanned for analog +t forecasts to each member.

If the members are computed differently (e.g. multi model ensembles), only historical forecasts of the member itself can be used. For the Lorenz96 model, where all

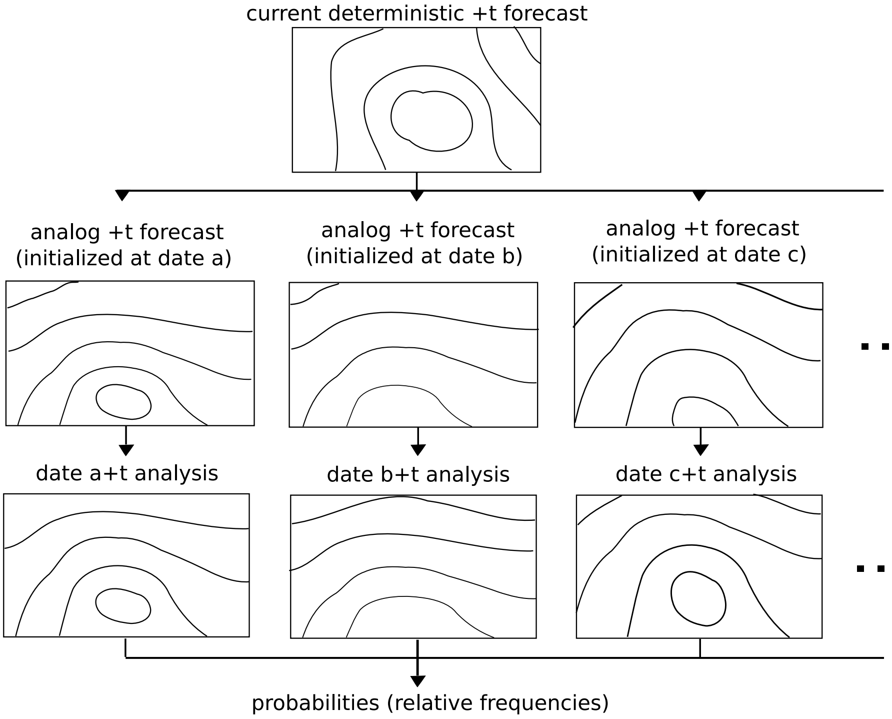

current deterministic +t forecast
analog +t forecast (initialized at date a)
analog +t forecast (initialized at date b)
analog +t forecast (initialized at date c)
date a+t analysis
date b+t analysis
date c+t analysis
probabilities (relative frequencies)

**Figure 3.2:** Schematic illustration of the *analogs of a deterministic forecast* method

members are equally calculated [2], forecasts of all members can also be compared among each other.

Thus, each member of the ensemble forecast is dressed with the corresponding analyses of the $n_{dens}$ most similar forecasts in the training data. The resulting ensemble with $n_{ens} * n_{dens}$ members should then be drawn from a similar distribution as the truth.

Like the approaches of Wang and Bishop (2005) and Roulston et al. (2003), this method is only appropriate for underdispersive ensembles. However, sorting the members from the smallest to the largest value and weighting them, like Fortin et al. (2006) proposed in their approach, is also feasible for this method. Thus, the method could also be used for overdispersive ensembles. However, because the Lorenz96 ensemble exhibits underdispersion, this has not been tested.

---

[2] except the first member (initialized with the analysis). However, the difference is marginal and therefore disregarded.

### 3.2.4 Analogy criterion

Regardless of whether the current analysis is compared to past analyses, or a current forecast is compared to a set of reforecasts, finding an appropriate criterion for the analogy is one of the main problems of all of these methods.

Given an infinite set of analyses/reforecasts, nearly identical analyses/reforecasts to the current one could be found. However, since the database is limited, there is a low chance of finding even similar states of global weather. However, some approximations can be made in order to obtain enough meaningful analogs (e.g. using smaller regions or fewer variables; Hamill and Whitaker 2006).

However, in global weather forecasting, because of both its high dimensionality (pressure, temperature, humidity,...) and the high resolution (high amount of gridpoints), an extremely large number of possibilities exist to test the similarity (choice of region, choice of variables,...).

In contrast, only one resolved variable exists in the Lorenz96 model, however, even in this simple model it is practically impossible to find a perfect analogy criterion. For example, if a forecast is to be made for the variable $X_1$, the analogy of the neighbour-gridpoints ($X_2$, $X_8$) might be more important than the analogy of farther gridpoints (e.g. $X_5$; see Figure 2.1). Furthermore, for the *no-forecast-model* method, the gridpoints 'upstream' might have to be examined more closely than the ones 'downstream' (Lorenz (1996) found, that the structures of X slowly progress 'westwards'). A good technique to account for these differences, would be to weight the absolute differences between $X_k$ (current) and $\hat{X}_k$ (historical) differently on each gridpoint. However, finding adequate weights leads to an extremely complicated optimization problem, for which at most local optima can be found. Thus, to retain simplicity and generality, in this study only two simple analogy criteria were tested:

1. a simple root mean square (rms) of the differences between $X_k$ (analysis or forecast) and the $\hat{X}_k$ (historical analysis or reforecast) to be compared.

$$AC(\mathbf{X}, \hat{\mathbf{X}}) = \sqrt{\sum_{k=1}^{K} (X_k - \hat{X}_k)^2} \tag{3.19}$$

   where $\mathbf{X}$ and $\hat{\mathbf{X}}$ are K-dimensional vectors with elements $X_{k\in\{1,...K\}}$ and $\hat{X}_{k\in\{1,...K\}}$ Smaller values of AC signify greater similarity.

2. Hamill and Whitaker (2006) found that in their data the rms criterion led to an underforecasting bias. A similar problem is also present in the Lorenz96 model. Because of the approximately normal distribution of the predictand X, it can be assumed that generally more analogs are found that are closer to the climatological mean. Thus, forecasts using these analogs are shifted towards the climatological mean. In Figure 3.3 this problem is illustrated schematically.

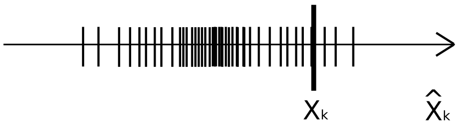


**Figure 3.3:** Problem of the rms analogy criterion. The small bars represent the historical data $\hat{X}_k$ and the large bar the current value of $X_k$. It is clear that in this example more analog $\hat{X}_k$ smaller $X_k$ than greater $X_k$ are found

As a solution, Hamill and Whitaker (2006) proposed an analogy criterion that operates with rank differences. For each gridpoint k, the rank of the current values $X_k$ when pooled with n (training data size) historical values $\hat{X}_k$ is derived. This rank of the current value is then compared with the ranks of the historical values $\hat{X}_k$. The smaller the sums of absolute rank differences, the more similar $\mathbf{X}$ and $\hat{\mathbf{X}}$ are.

$$AC(\mathbf{X}, \hat{\mathbf{X}}) = \sum_{k=1}^{K} |rank(X_k) - rank(\hat{X}_k)| \tag{3.20}$$

Furthermore, this analogy criterion is also applied locally to $n_r$ neighbour gridpoints of the gridpoint to be forecasted.

$$AC(X_m, \hat{X}_m) = \sum_{k=m-n_r/2}^{m+n_r/2} |rank(X_k) - rank(\hat{X}_k)| \tag{3.21}$$

If, for example, a forecast for $X_3$ has to be made, the sum of rank differences of $X_2$, $X_3$, and $X_4$ is used as analogy criterion. Thus, the effect of different 'regions' is tested. 'Region' sizes of $n_r = 0, 2, 4$ were used (*rank difference* analogy criterion with 'region size' 4 is termed *rankdiff4* in subsequent figures).

# Chapter 4

# Results

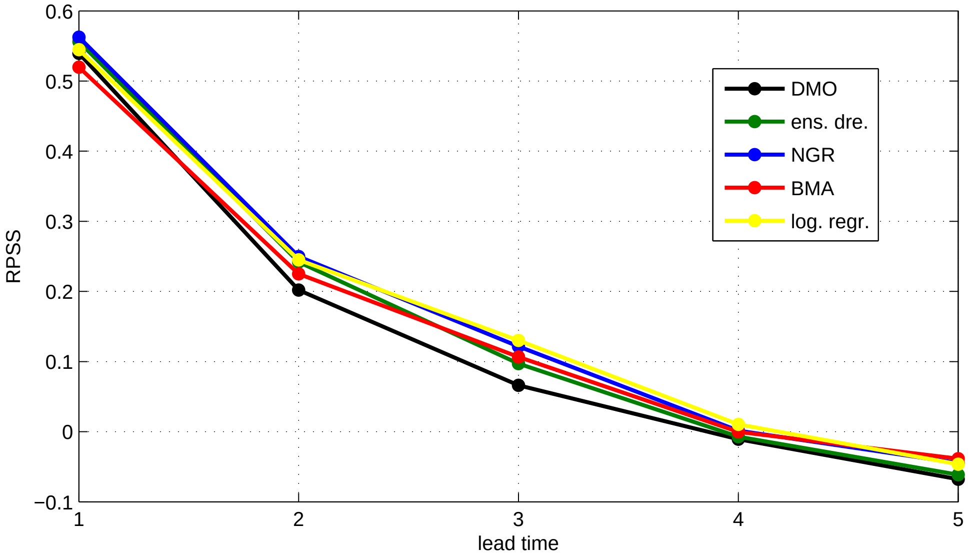
Wilks (2006a) tested several ensemble MOS methods (hereafter termed ’traditional’ ensemble MOS methods) in the same idealized Lorenz96 setting as the one used here. In this chapter, his results are first summarized. Subsequently, the new approaches using analogs are tested and compared with these ’traditional’ ensemble MOS Methods.

## 4.1 ’Traditional’ ensemble MOS methods

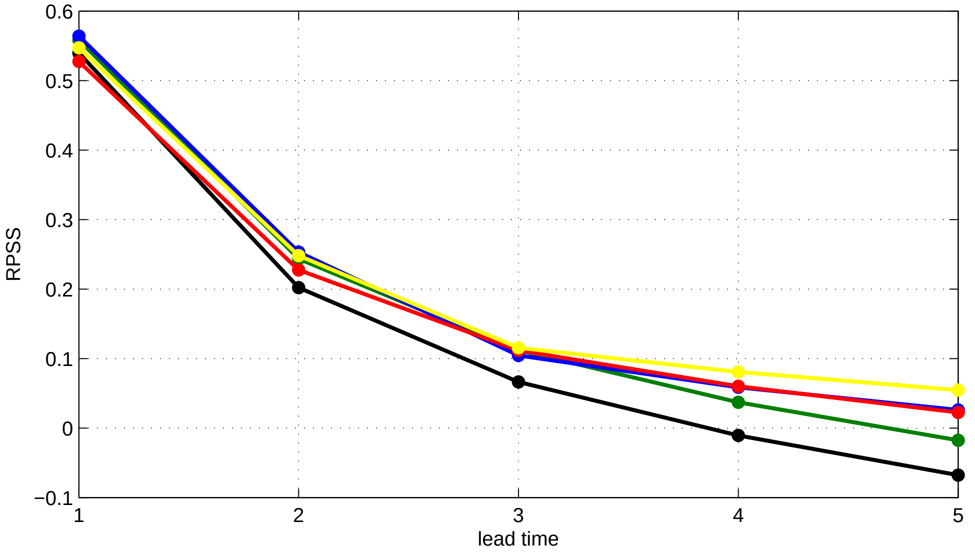
Figures 4.1a and 4.1b show the *ranked probability skill scores* (RPSS; section 2.2.2), relative to the climatology, for *direct model output* (DMO) and the ensemble MOS methods *ensemble dressing*, *logistic regression*, *non-homogenous Gaussian regression* (NGR) and *Bayesian model averaging* (BMA).

The training sample size used for Figure 4.1a is $n = 50$. This relatively small size shall simulate the small amount of historical data used in a continuously updated MOS system that can cope with NWP model changes or seasonal variations of climate (Wilks 2006a; Mao et al. 1999; Wilson and Vallée 2002). Figure 4.1b is computed using a training data length of 1500, which is comparable to the size of historical databases in traditional deterministic MOS settings (Wilks 2006a; Carter et al. 1989; Hamill et al. 2004). In both figures the ensemble size is $n_{ens} = 51$.

For both lengths of training data it can be seen that, except BMA at lead time



**(a)**



**(b)**

**Figure 4.1:** *Ranked probability skill scores* (RPSS) relative to the climatological relative frequencies for the six-category probabilistic forecasts, defined by the climatological quantiles $q_{1/10}$, $q_{1/3}$, $q_{1/2}$, $q_{2/3}$ and $q_{9/10}$. Ensemble size is $n_{ens} = 51$ and training data length is (a) n=50 and (b) n=1500. RPSS are shown for *direct model output* (DMO), *ensemble dressing*, *non-homogeneous Gaussian regression* (NGR), *Bayesian model averaging* (BMA) and *logistic regression*. The exact values of (b) are also shown in Table 4.1.

$T = 1$[1], all ensemble MOS methods improve more or less significantly over DMO. Using the larger training sample, the RPSS of the regression methods (NGR and *logistic regression*) and BMA actually remain positive (superior to climatology) throughout the whole forecast period (Figure 4.1b). For shorter lead times ($T = 1, 2$) NGR performs best, while *logistic regression* best estimates the probability distributions at longer lead times ($T = 3, 4, 5$).

Since a larger number of regression parameters have to be fitted for *logistic regression* (separate regressions for each quantile and forecast lead time), a larger training data set is needed. Consequently, with the small training sample used in Figure 4.1a, the improvement over the other approaches at longer lead times is smaller or even disappears.

*Ensemble dressing* shows good scores for all lead times and both training sample sizes.

Finally, the new BMA approach of Bishop (2008) performs poorly for shorter lead times ($T = 1, 2$), even worse than DMO at $T = 1$. However, for longer lead times, it can keep up with the regression methods. For $T = 5$ and the smaller training sample it even receives the best RPSS of the tested ensemble MOS methods.

Figure 4.1 shows the *Brier skill scores* (BSS; section 2.2.1) of the ensemble MOS methods, relative to DMO, for the climatological quantiles $q_{1/10}$, $q_{1/3}$ and $q_{1/2}$. Because of the climatological symmetry of the Lorenz96 model, the *Brier skill scores* for the quantiles $q_{2/3}$ and $q_{9/10}$ are expected to be the same as for $q_{1/3}$ and $q_{1/10}$. The training data set with length $n = 1500$ and an ensemble size of $n_{ens} = 51$ is used.

Consistent with the *ranked probability skill score*, *logistic regression* and *non-homogeneous Gaussian regression* generally receive best *Brier skill scores* for all quantiles and lead times. *Non-homogeneous Gaussian regression* performs slightly better at shorter and *logistic regression* is preferred at longer lead times. Because in *ensemble dressing* the members are dressed symmetrically with the same distribution for all forecast occasions, extreme events are overestimated (Fortin

---

[1] 1 time unit corresponds to $6.\dot{6}$ 'days', see section 2.1.1

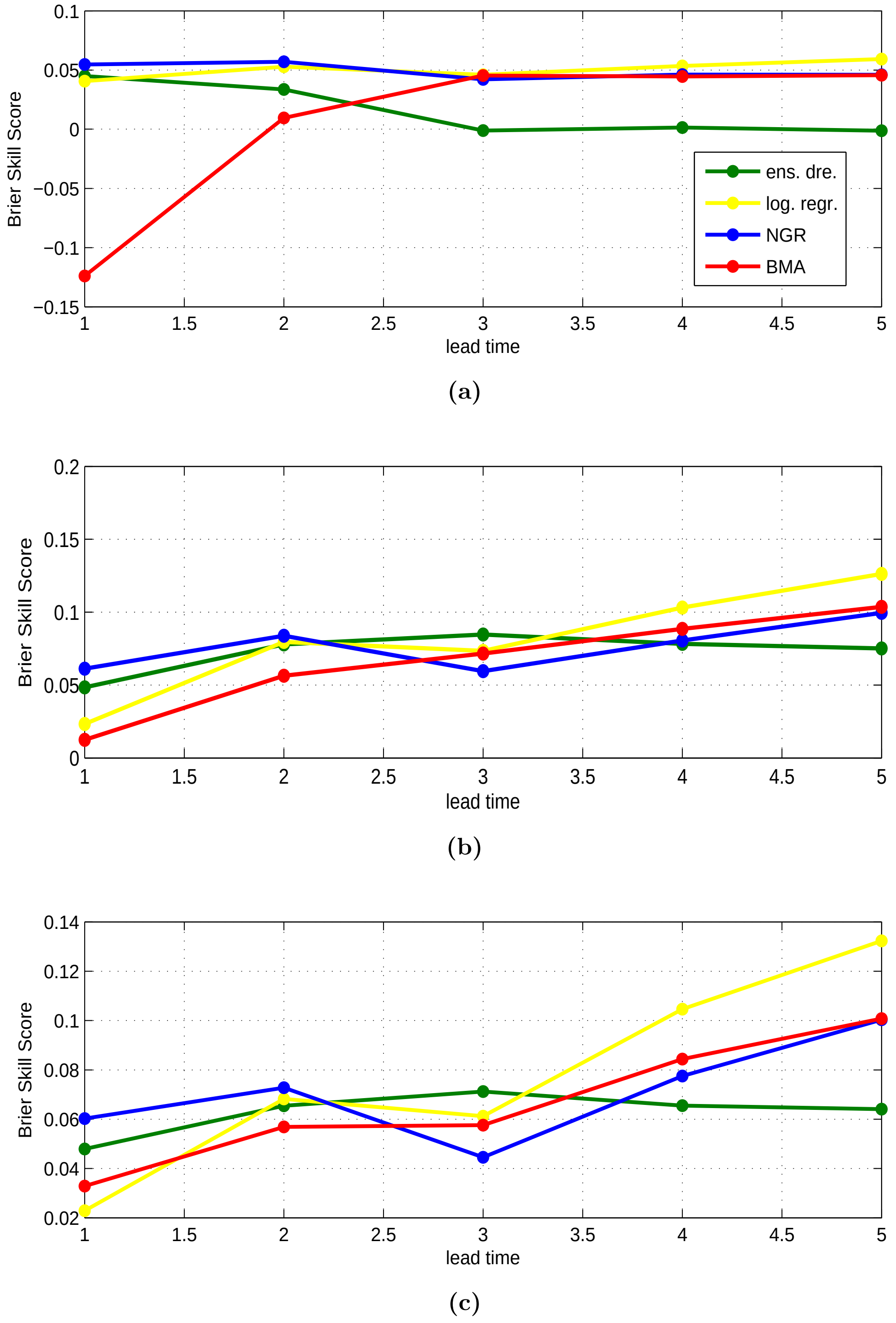
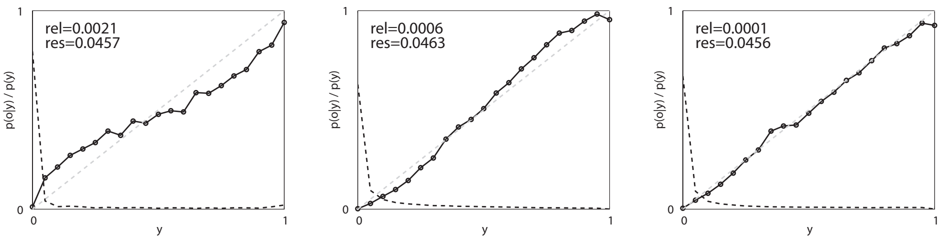
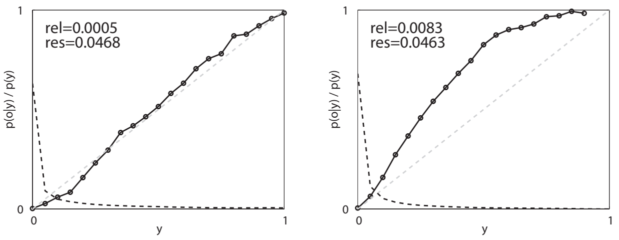


**Figure 4.2:** *Brier skill scores* (BSS), relative to *direct model output* (DMO), for the probabilities of the predictand not to exceed the climatological quantiles (a) $q_{1/10}$, (b) $q_{2/3}$ and (c) $q_{1/2}$. Training data size used is n=1500 and ensemble size is $n_{ens} = 51$. BSS are shown for *ensemble dressing*, *non-homogeneous Gaussian regression* (NGR), *Bayesian model averaging* (BMA) and *logistic regression*.

et al. 2006; Bishop 2008). Therefore, these methods do not improve over DMO for $q_{1/10}$ and longer lead times ($T = 3, 4, 5$). The main feature of the new BMA approach of Bishop (2008) is to correct this overestimation of extreme events. As can be seen in Figure 4.2a, this correction works well for longer lead times, while for $T = 1$ there is no improvement over the old BMA method of Raftery et al. (2005) (not shown).



**(a)** DMO **(b)** ens. dressing **(c)** log. regression

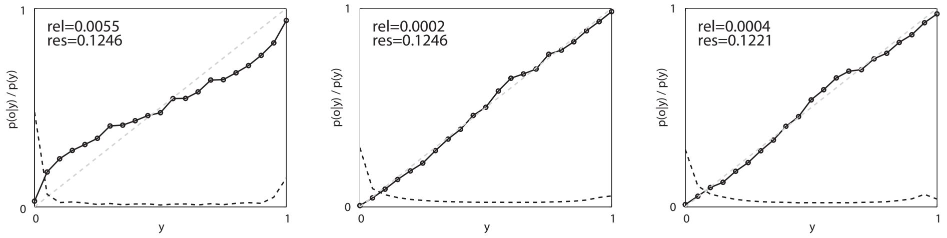


**(d)** NGR **(e)** BMA

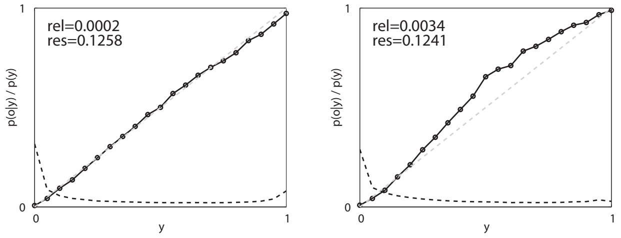
**Figure 4.3:** Reliability diagrams for $q_{1/10}$ at lead time $T = 1$. Ensemble size is $n_{ens} = 51$ and training data length $n = 1500$. The solid lines are the graphs of the calibration function, the dashed gray lines (1:1 line) characterize the calibration function of a perfect forecast and the dashed black lines specify the refinement distributions. See section 2.2.3 for further information. Reliability diagrams are shown for (DMO), *ensemble dressing*, *non-homogeneous Gaussian regression* (NGR), *Bayesian model averaging* (BMA) and *logistic regression.*

Figures 4.3, 4.4, 4.5 and 4.6 show the reliability diagrams (section 2.2.3) of DMO and the ensemble MOS methods for different quantiles and lead times.

The calibration function (heavy line) of DMO is less steep than the 1:1 reference line (dashed line - calibration function of a perfect forecast) for all quantiles and lead times. This overconfidence (Wilks 2006b) results from the underdispersion of the ensemble and is also reflected in the higher values of reliability (REL).

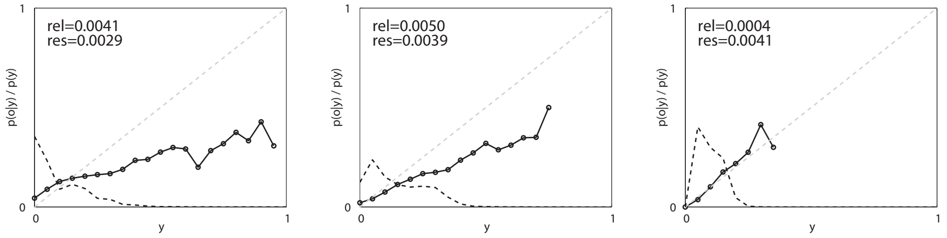


**(a)** DMO **(b)** ens. dressing **(c)** log. regression

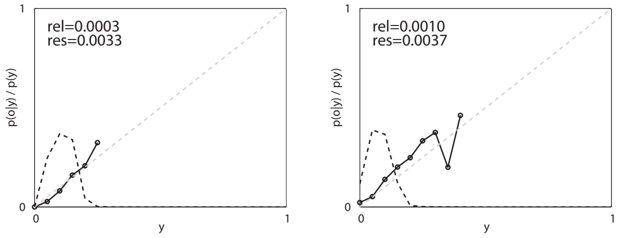


**(d)** NGR **(e)** BMA

**Figure 4.4:** As Figure 4.3 but for $q_{1/3}$

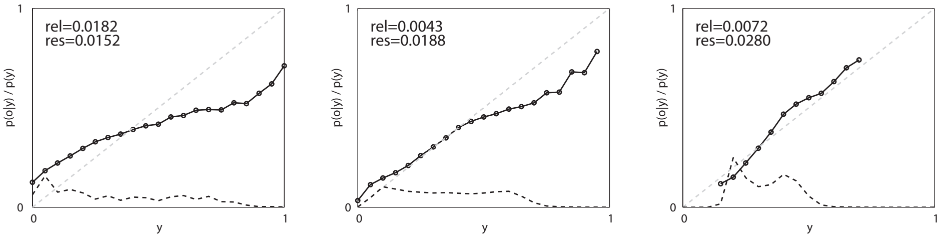


**(a)** DMO **(b)** ens. dressing **(c)** log. regression

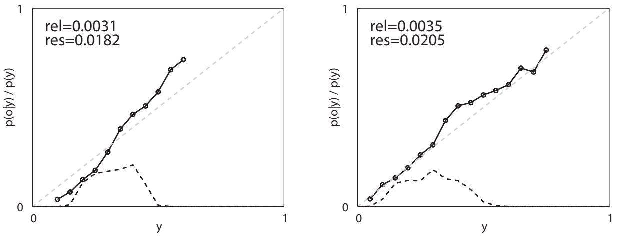


**(d)** NGR **(e)** BMA

**Figure 4.5:** As Figure 4.3 but for $q_{1/10}$ and T=4



**(a)** DMO **(b)** ens. dressing **(c)** log. regression



**(d)** NGR **(e)** BMA

**Figure 4.6:** As Figure 4.3 but for $q_{1/3}$ and T=4

Except BMA at $T = 1$ for $q_{1/10}$ and *ensemble dressing* at $T = 1$ for $q_{1/10}$, the other methods are all better calibrated (smaller value of reliability). *Logistic regression* and *non-homogeneous Gaussian regression* achieve the best reliability in all cases, which is consistent with their good *Brier scores* (Figure 4.1). The overforecasting of rare events ($V \leq q_{1/10}$) by *ensemble dressing* can be seen in its calibration function, shifted downwards in Figure 4.5b. Because of this, its reliability for $q_{1/10}$ and $T = 4$ is even smaller than that of DMO. This can also be seen its *Brier scores* (Figure 4.2a). In Figure 4.3b it is shown that this overforecasting is hardly present at shorter lead times.

Because the BMA method corrects this overforecasting, it reaches good reliability at longer lead times. However, for $T = 1$, Figure 4.3e shows that this correction is too strong, leading to a strong underforecasting bias (calibration function shifted upwards). This can explain the poor *ranked probability skill scores* and *Brier skill scores* for $q_{1/10}$ at this lead time.

## 4.2 Analog methods

In this section the results of the analog methods are presented and discussed. In Figure 4.7 the *ranked probability skill scores* (RPSS), relative to the climatology, of selected analog methods, DMO and *logistic regression* are shown. The RPSS of all ensemble MOS and analog methods tested can be found in Table 4.1. Because generally, *logistic regression* achieves the best, whereas DMO obtains the worst RPSS of the 'traditional' ensemble MOS methods (Figure 4.1b), these methods are chosen to indicate the boundaries of the previously tested approaches. Because the training data length for the analog methods should be significantly larger than the ensemble size, only the RPSS for the larger training sample (n=1500) are shown.

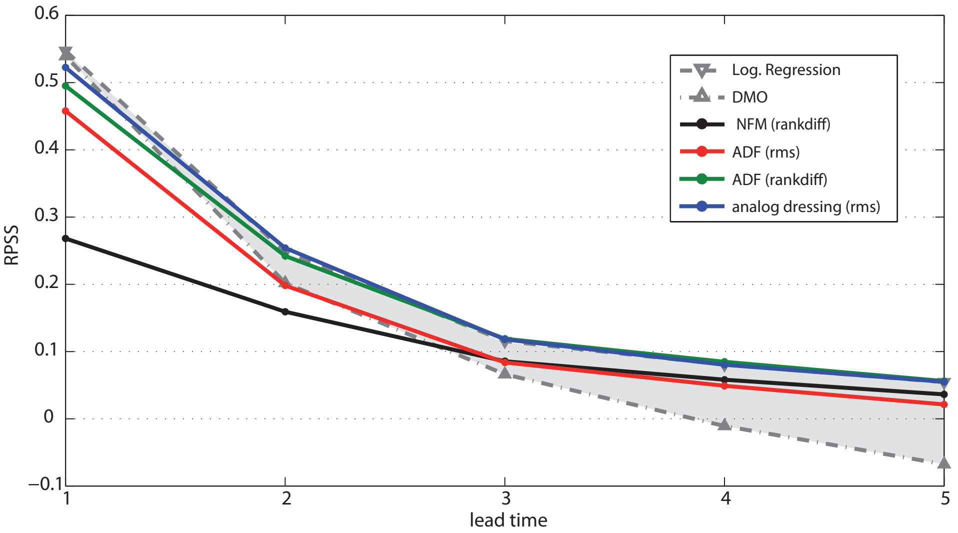


**Figure 4.7:** *Ranked probability skill scores* (RPSS) relative to the climatology for the six-category probabilistic forecasts defined by the climatological quantiles $q_{1/10}$, $q_{1/3}$, $q_{1/2}$, $q_{2/3}$ and $q_{9/10}$. Ensemble size is $n_{ens} = n_{analogs} = 51$, daughter ensemble size is $n_{dens} = 15$ and training data length is n=1500. RPSS are shown for *logistic regression*, *direct model output* (DMO), *no-forecast-model* (NFM), *analogs of a deterministic forecast* (ADF) and *analog dressing*. Area between DMO and logistic regression is shaded. Terms in brackets denote the used analog criteria *rank difference* (rankdiff) or RMS.

Overall, the analog methods achieve relatively bad scores at short lead times ($T = 1$). However, for longer lead times ($T = 3, 4, 5$), they all improve significantly over DMO and even have scores comparable to or better than *logistic regression*.

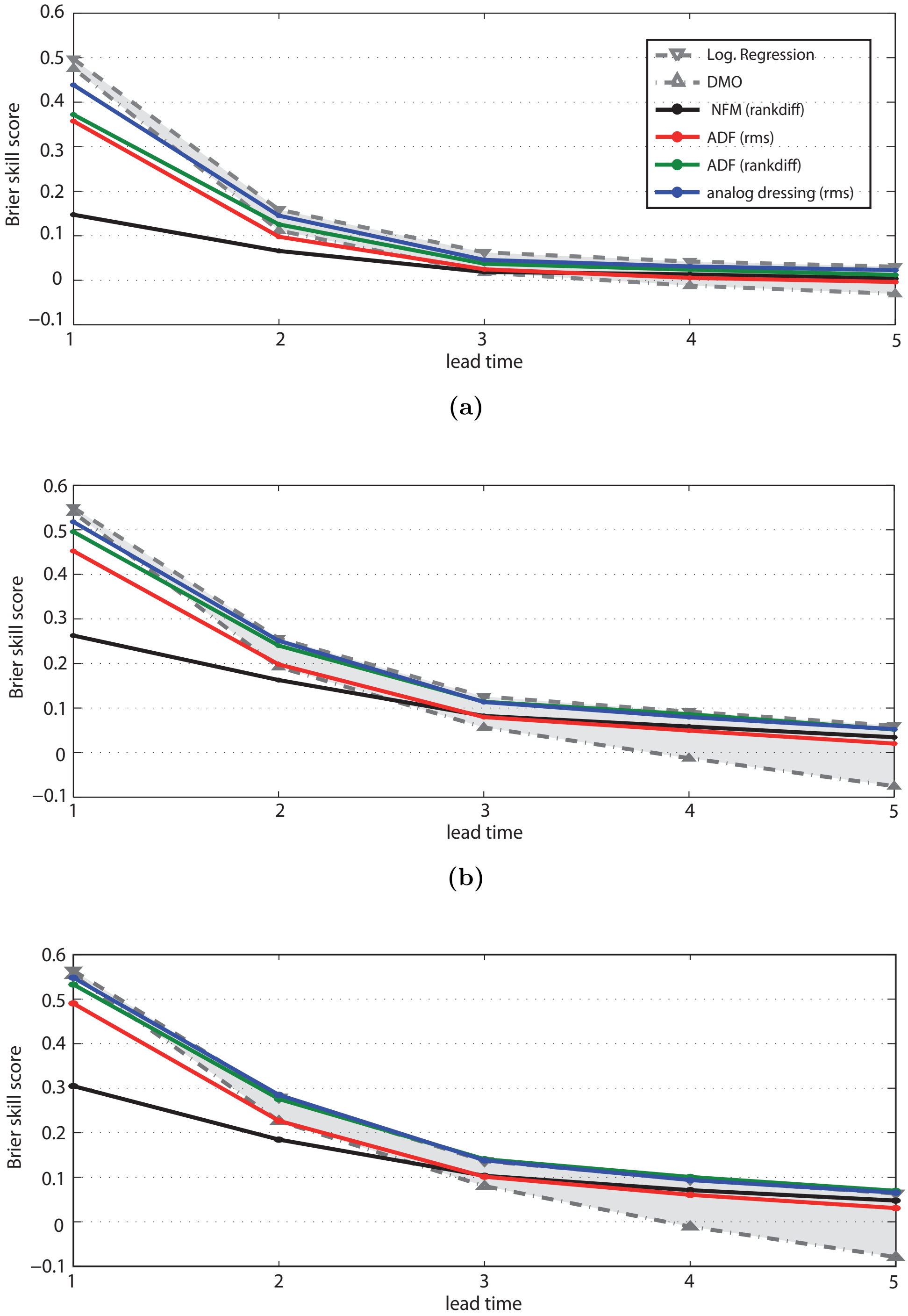
| | Forecast Lead Time, T | | | | |
|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 |
| DMO | 0.5394 | 0.2020 | 0.0662 | -0.0108 | -0.0677 |
| Logistic Regression | 0.5471 | 0.2480 | 0.1155 | 0.0806 | 0.0545 |
| NGR | **0.5639** | 0.2530 | 0.1043 | 0.0586 | 0.0260 |
| ensemble dressing | 0.5581 | 0.2435 | 0.1118 | 0.0371 | -0.0175 |
| BMA | 0.5274 | 0.2275 | 0.1117 | 0.0600 | 0.0224 |
| NFM (rms) | 0.2681 | 0.1592 | 0.0855 | 0.0581 | 0.0363 |
| NFM (rankdiff) | 0.2693 | 0.1600 | 0.0864 | 0.0587 | 0.0369 |
| NFM (rankdiff4) | 0.2159 | 0.1462 | 0.0743 | 0.0547 | 0.0316 |
| ADF (rms) | 0.4579 | 0.1982 | 0.0837 | 0.0490 | 0.0213 |
| ADF (rankdiff) | 0.4951 | 0.2420 | 0.1190 | 0.0847 | 0.0560 |
| ADF (rankdiff4) | 0.5127 | 0.2490 | **0.1239** | **0.0869** | **0.0572** |
| analog dressing (rms) | 0.5225 | 0.2540 | 0.1180 | 0.0803 | 0.0550 |
| analog dressing (rankdiff) | 0.5203 | 0.2536 | 0.1179 | 0.0801 | 0.0551 |
| analog dressing (rankdiff4) | 0.5330 | **0.2582** | 0.1184 | 0.0790 | 0.0536 |

**Table 4.1:** As Figure 4.7 but tabulated for all tested methods. Best method for each lead time is highlighted.

For shorter forecast periods, *analog dressing* performs best, while for longer lead times, the *analogs of a deterministic forecast* (ADF) method achieves the best RPSS of all tested approaches. The *no-forecast-model* (NFM) method performs poorly at short lead times. For longer forecast periods, however, it can keep up with the other ensemble MOS and analog methods. The reason for this will be discussed later.

For the NFM method, the *rank difference* analogy criterion works best, though the RMS criterion is hardly worse. Applying the *rank difference* criterion to a smaller 'region' (rankdiff4) diminishes the skill scores of this method (Table 4.1). This is because the development of the predictand is influenced by the X-values on all gridpoints. Hence, analogy on all gridpoints is required.

For the ADF approach, the *rank difference* criterion is considerably superior to the RMS similarity measure. With the smaller 'region', this method can be improved additionally. At shorter lead times (T=1,2,3), the use of fewer gridpoints also improves the *analog dressing* method, however, for longer lead times (T=4,5) it slightly worsens its skill scores. The differences between the *rank difference* and the RMS analogy criterion are small for this approach.



**(a)**

**(b)**

**(c)**

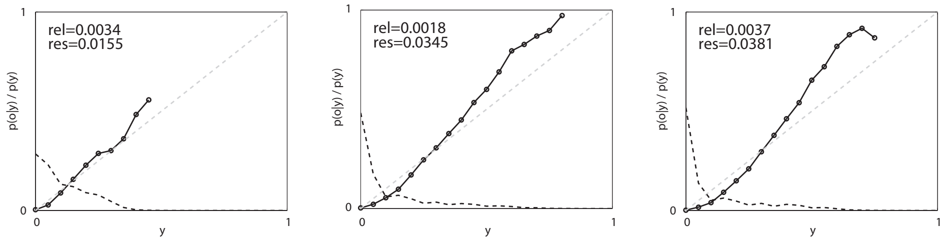
**Figure 4.8:** *Brier skill scores* (BSS), relative to climatology, for the probabilities of the predictand not exceeding the climatological quantiles (a) $q_{1/10}$, (b) $q_{2/3}$ and (c) $q_{1/2}$. Ensemble size is $n_{ens} = n_{analogs} = 51$, daughter ensemble size is $n_{dens} = 15$ and training data length is n=1500. BSS are shown for *logistic regression*, *direct model output* (DMO), *no-forecast-model* (NFM), *analogs of a deterministic forecast* (ADF) and *analog dressing*. Area between DMO and logistic regression is shaded. Terms in brackets denote the used analog criteria *rank difference* (rankdiff) or RMS.

In Figure 4.8 the *Brier skill scores*, relative to the climatological relative frequencies, are plotted for selected analog methods, DMO and *logistic regression*. Similar to the RPSS, the analog approaches perform quite well at longer lead times and relatively poorly for shorter forecast ranges. In comparison to the 'traditional' ensemble MOS methods, all analog methods are slightly less suitable for rarer ($V \leq q_{1/10}$) than for common ($V \leq q_{1/2}$) events. This is because in the training data these events appear less frequently. Therefore, the analogs found for extreme (rare) events are generally less similar than for common events. This leads to an overdispersive ensemble and hence to underconfident forecasts for these events. This can also be seen in the reliability diagrams for the $q_{1/10}$-quantile (Figure 4.9; calibration function steeper than the 1:1 reference line). Using a larger training data set reduces this problem (see Figure 4.15), because more close analogs can be found.

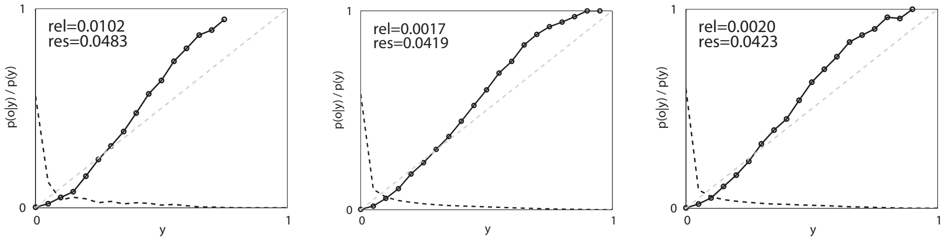
For the ADF method, the differences (of the *Brier skill scores*) between the two analogy criteria are considerably smaller for $q_{1/10}$ than for the other quantiles. This leads to the assumption that the just described problem of overdispersion is somewhat intensified with the *rank difference* analogy criterion (Figure 4.9b and c). This is easy to understand, because the *rank difference* criterion partially forces the analog historical forecasts to be more extreme than the current one. Then the found analogs are still less similar, because few of them exist overall.

Comparing the reliability diagrams of the analog (Figures 4.9, 4.10 ,4.11 and 4.12) and the 'traditional' MOS methods (Figures 4.3, 4.4 ,4.5 and 4.6), one can see that overall the reliability is good (small) for the analog methods, while the resolution is generally better (large) for the other approaches. Bad resolutions signify that the forecasts poorly discern between different events (Wilks 2006b). As an extreme example, the climatological relative frequencies are the same for all forecast occasions and achieve a reliability of 0 (perfect) but only a resolution of 0 (worst).

Better values of resolution can be achieved, if the *rank difference* or the 4-neighbor *rank difference* analogy criterion (rankdiff4) is used. With the RMS criterion, the

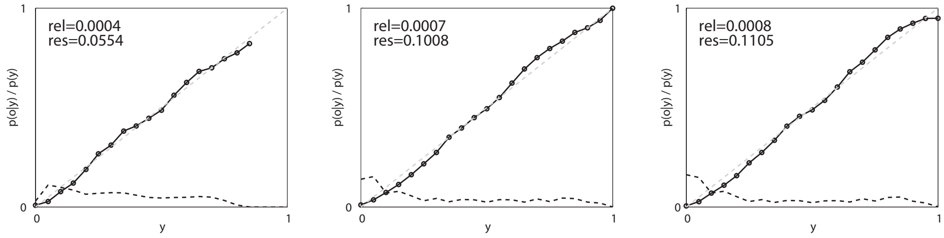


**(a)** NFM (rankdiff) **(b)** ADF (rms) **(c)** ADF (rankdiff)

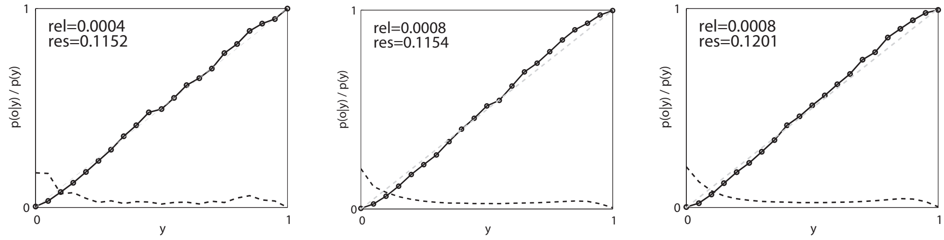


**(d)** ADF (rankdiff4) **(e)** analog dressing (rms) **(f)** analog dress. (rankdiff4)

**Figure 4.9:** Reliability diagrams of the analog methods for $q_{1/10}$ and lead time $T = 1$. Training data length is $n = 1500$. Ensemble size is $n_{ens} = n_{analogs} = 51$ and the size of daughter ensembles for *analog dressing* is $n_{dens} = 15$

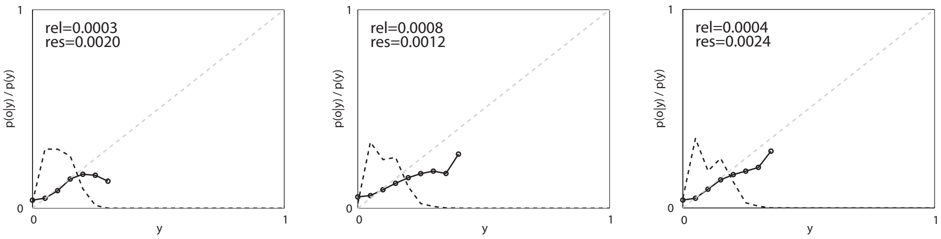


**(a)** NFM (rankdiff) **(b)** ADF (rms) **(c)** ADF (rankdiff)

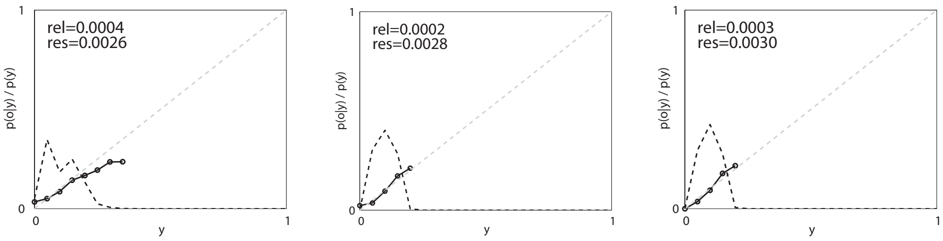


**(d)** ADF (rankdiff4) **(e)** analog dressing (rms) **(f)** analog dress. (rankdiff4)

**Figure 4.10:** As Figure 4.9 but for $q_{1/3}$ and T=1

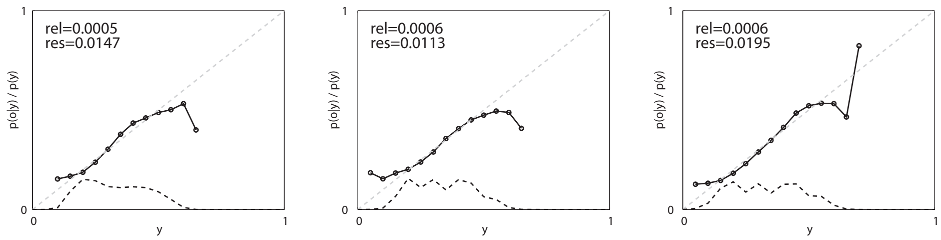


**(a)** NFM (rankdiff) **(b)** ADF (rms) **(c)** ADF (rankdiff)

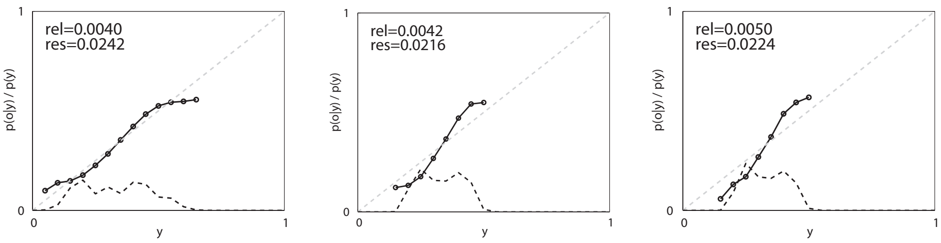


**(d)** ADF (rankdiff4) **(e)** analog dressing (rms) **(f)** analog dress. (rankdiff4)

**Figure 4.11:** As Figure 4.9 but for $q_{1/10}$ and T=4



**(a)** NFM (rankdiff) **(b)** ADF (rms) **(c)** ADF (rankdiff)



**(d)** ADF (rankdiff4) **(e)** analog dressing (rms) **(f)** analog dress. (rankdiff4)

**Figure 4.12:** As Figure 4.9 but for $q_{1/3}$ and T=4

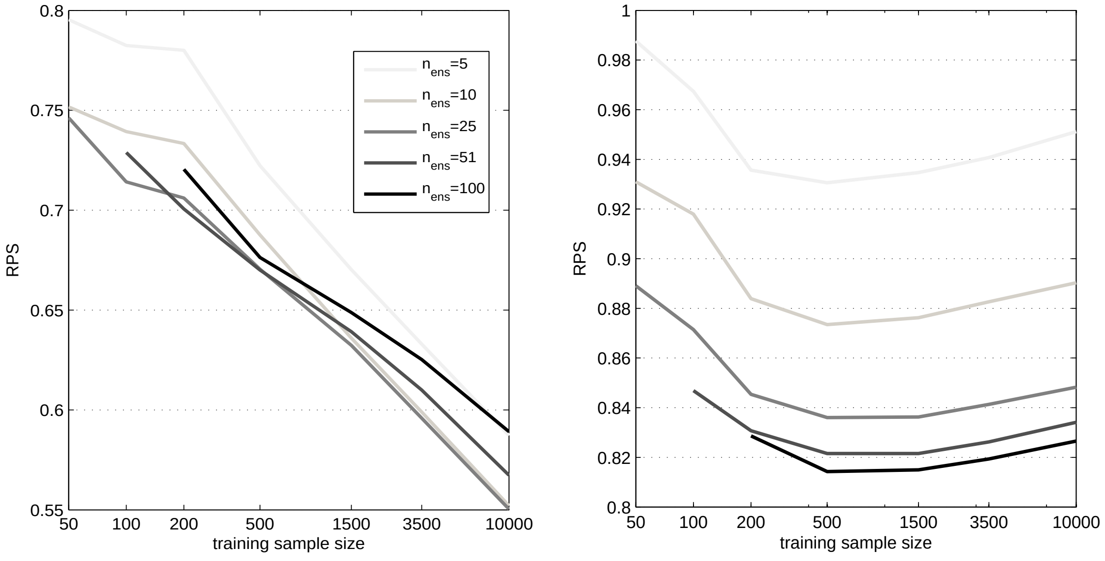
forecasts are typically shifted towards the climatological mean. It thereby converges slightly towards the climatology, which can explain the worse resolution. With the 4-neighbor *rank difference* criterion, the analogs found are locally closer, which further improves the resolution.

Intuitively, one would think that a larger training sample would increase the resolution. However, comparing Figures 4.9b, 4.10b, 4.11b and 4.12b with Figure 4.15, it can be seen that for the ADF method this is only true at short lead times ($T = 1$). This suggests that for longer lead times, enough close analogs can be found already with the n=1500 training dataset. Thus, an even longer training data set does not further improve this method at longer forecast periods. This problem will be discussed later.

The ensembles of the analog methods exhibit overdispersion at $T = 1$ and $q_{1/10}$. This can be seen in their calibration functions steeper than the 1:1 reference line in Figures 4.9 and 4.10. Therefore, the reliability of these approaches reaches significantly smaller (worse) values than of the 'traditional' ensemble MOS methods for this lead time and quantile. Together with the bad resolution, this causes the bad *Brier* and *ranked probability scores* of the analog methods at $T = 1$.

For *analog dressing*, all information of the dynamical ensemble is used. Since the raw ensemble performs quite well at $T = 1$ and 2, this method is best among the analog methods for these lead times, however it is still worse than DMO at $T = 1$. Better RPSS than DMO at $T = 1$ can be achieved, using a larger dataset (n=3500,10000) or smaller 'regions' for the analogy criterion (0- and 2-neighbor *rank difference* analogy criterion). For longer forecast periods, the ensemble mean (ADF) contains sufficient information, and using the entire ensemble (*analog dressing*) even worsens the performance.

Generally, it can be said that the *no-forecast-model* method works quite badly. So how can it perform better than DMO and other ensemble MOS methods at longer lead times? To explain this, the results of the ensemble MOS methods have to be examined more closely. For lead times $T = 4$ and 5, DMO produces forecasts that are worse than the climatology. The ensemble MOS methods can then only correct these forecasts and hence their RPS can becomes slightly superior to that of the

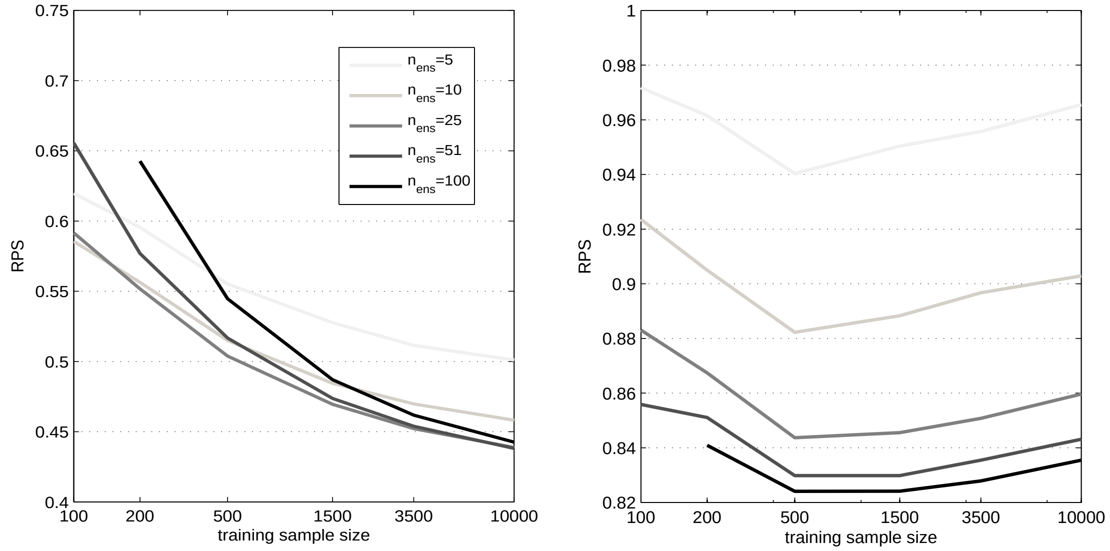


**(a)** **(b)**

**Figure 4.13:** *Ranked probability scores* (RPS; smaller values are better!) of the *no-forecast-model* (NFM) method for different ensemble and training sample sizes for lead times (a) $T = 1$ and (b) $T = 4$

climatology. On the contrary, the NFM method is completely independent from the dynamical ensemble and can maintain its advantage over climatology for the entire forecast period. This leads to the suggestion, that the good results of this method can not be traced back to the fact that it performs well, rather, it results from the bad performance of the dynamical ensemble and hence of the ensemble MOS approaches.

Figures 4.13 and 4.14 show the effects of different training sample and ensemble sizes for the NFM and the ADF methods respectively. As expected, there is a strong dependency on the training data length at $T = 1$, whereas the influence of the ensemble size is weak. However, at longer lead times the performance of these methods is affected strongly by the ensemble size, while the dependency on the training sample size is not clear. If the training data set is short, an increase improves the RPS. However, if the dataset is already large enough ($n > 500$), a further enlargement even worsens the score. A possible explanation is that the calibration functions of these methods at $T = 4$ are slightly less steep than the



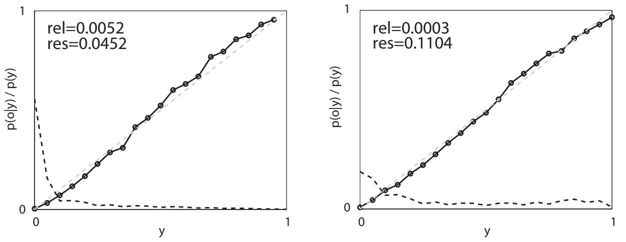
**(a)** **(b)**

**Figure 4.14:** Same as Figure 4.13 but for the *analogs for a deterministic forecast* (ADF) method

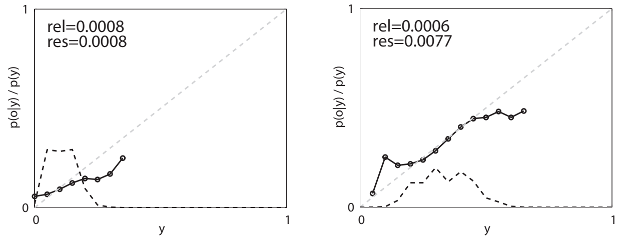
1:1 reference line (Figures 4.11b and 4.12b). This indicates an underdispersive ensemble and thus an overconfident forecast. With a greater training dataset, closer analogs can be found, intensifying this underdispersion (compare Figures 4.11b and 4.12b with Figure 4.15). In fact this should be visible in the value of reliability. However, differences exist only in the resolution term. The reliability is the same for both training data lengths (Figures 4.11b, 4.12b and 4.15). Therefore the reason for the increase of RPS with longer training data at longer lead times has to remain unclear. Furthermore, the differences are essentially tiny, so maybe this increase should not be overemphasized.

In Figure 4.16, a similar graphic is shown for *analog dressing*. Instead of the ensemble size, the effects of different daughter ensemble sizes are tested. The influence of the training data size is similar to the other methods, however, the dependency on the daughter ensemble size remains largely unclear. Only the smallest daughter ensembles ($n_{dens} = 1, 5$) can be identified to be worse.

Finally, Figure 4.17 depicts the effects of different 'region' sizes, to which the *rank difference* analogy criterion is applied. For the NFM method it can be seen that

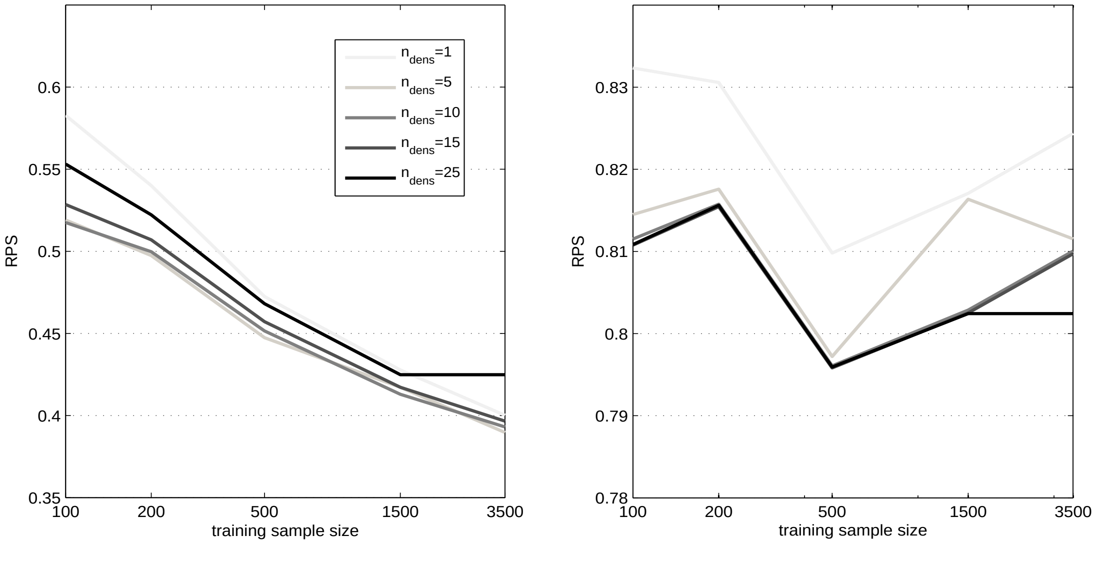


**(a)** $q_{1/10}$, T=1 **(b)** $q_{1/3}$, T=1



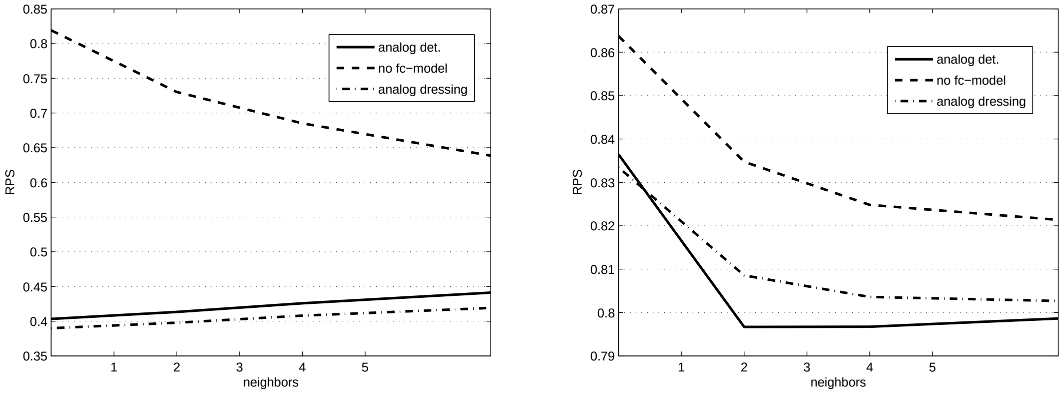
**(c)** $q_{1/10}$, T=4 **(d)** $q_{1/3}$, T=4

**Figure 4.15:** Reliability diagrams of the *analogs for a deterministic forecast* (ADF) method for different quantiles and lead times. The RMS analogy criterion is used. Ensemble size is $n_{ens} = n_{analogs} = 51$ and training sample size is n=10000



**(a)** **(b)**

**Figure 4.16:** *Ranked probability scores* (RPS) of the *analog dressing* method for different daughter ensenmble and training sample sizes for lead times (a) $T = 1$ and (b) $T = 4$. Size of the dynamical ensemble is $n_{ens} = n_{analogs} = 51$



**(a)** **(b)**

**Figure 4.17:** *Ranked probability scores* (RPS) of the 3 basic analog methods as function of the 'region'-size for the *rank difference* analogy criterion. The values on the abscissa specify the number of used neighbors.

irrespective of lead time, the best scores are achieved if all gridpoints are used. For the other two approaches, at $T = 1$ the 0-neighbors *rank difference* analogy criterion works best. This is evident, because then the forecasts made with these methods display the highest similarity to DMO. Since DMO generally performs better than these approaches at $T = 1$, this leads to improved scores. However, at $T = 4$, the results achieved with the 0-neighbor criterion are significantly worse . At this lead time, for the ADF approach, taking into account 2 or 4 neighbors is slightly better than using all gridpoints. For *analog dressing*, the best results are achieved by applying the analogy criteria to all gridpoints.

All in all, a medium number of gridpoints is best for both methods.

# Chapter 5

# Summary and conclusion

In this thesis, several methods of probabilistic forecasting were tested and compared in the idealized Lorenz96 model (Lorenz 1996). The main focus was on three basic methods that use 'analogs'.

For the first, imagine that an analysis very similar to the current one can be found in a historical database. Then it is quite probable that the weather of tomorrow is similar to the weather that was observed the day after this similar historical analysis. Thus, the analysis of the next day can provide a forecast for tomorrow. If not only the single best, but a set of analogs is taken, an ensemble can be formed to provide a probabilistic forecast. This method has been called *no-forecast-model*, because it does not use any numerical forecast model.

For the second approach, termed *analogs of a deterministic forecast*, a reforecast dataset is needed. If deterministic forecasts similar to the current one can be found in the reforecast database, it should be appropriate to take the corresponding analyses (the day for which the forecast was made) as a forecast. Again, by taking a set of analogs, a probabilistic forecast is provided.

The third method is an extension of the *analogs of a deterministic forecast* approach. The same technique is applied separately to each member of a dynamical ensemble. Hence, a greater ensemble is formed, which again provides a probabilistic forecast. This method has been called *analog dressing*, because of the similarity to *ensemble dressing*.

To determine which analyses/forecasts are more similar than others, an analogy criterion is needed. Besides a simple root mean square (RMS) of the differences on all gridpoints, a measure that uses rank differences was introduced. Furthermore, different 'region' sizes (number of gridpoints) to which these criteria are applied were compared.

For comparison, other approaches that statistically postprocess ensemble forecasts were tested. *Ensemble dressing*, *logistic regression* and *nonhomogeneous Gaussian regression* (NGR) were chosen, because Wilks (2006b) showed that they perform best in the Lorenz96 system. Since *Direct model output* (DMO) is the method commonly used in operational weather prediction, this also was compared. Moreover, a new *Bayesian model averaging* (BMA) approach, which Bishop (2008) proposed, was tested here in the Lorenz96 model.

The *no-forecast-model* method generally performs poorly, with the exception of longer lead times, for which it can keep up with the other methods. This was found not to result from the good quality of the method, but rather from the bad performance of the dynamical ensemble, and thus, the other methods that postprocess this dynamical ensemble.

The two other analog methods showed very promising results. The *analogs of a deterministic forecast* method even performs better than all other tested methods at longer lead times. However, for shorter forecast ranges the applicability of this method is limited. For these cases, it is even inferior to *direct model output* (worse than the ensemble MOS methods).

*Analog dressing* achieves slightly better results at these short lead times. However, for longer forecast periods, its skills are somewhat inferior to the *analogs of a deterministic forecast* method. The *analog dressing* method uses the entire information of the dynamical ensemble, whereas for the *analogs of a deterministic forecast* method, all information is compressed into the ensemble mean. Hence, it can be said that if the raw ensemble forecast already predicts well, its entire information is useful. However, when the raw ensemble becomes worse (i.e. longer lead times), the ensemble mean provides better information than the whole ensemble.

For the *no-forecast-model* method, the RMS and the *rank difference* analogy criteria work quite similarly. Larger 'regions' to which the criteria are applied are definitely superior for this method.

The *analogs of a deterministic forecast* method showed better results when the *rank difference* similarity measure was used. For *analog dressing*, it does not matter which criterion is used.

For both methods, *analog dressing* and *analogs of a deterministic forecast*, fewer gridpoints for the criteria are better at short lead times. For longer forecast periods, a medium number of gridpoints showed the best skills for the *analogs of a deterministic forecast*, while the use of all gridpoints works best for *analog dressing.* All in all, a medium number should be used for both approaches.

At short lead times, all analog methods can be improved by increasing the training data size. On the contrary, for longer lead times, the approaches do not show better results when a larger training sample is used.

The new BMA approach of Bishop (2008), which was tested here, showed a significant improvement over the old BMA method of Raftery et al. (2005). However, at very short lead times, the ill treatment of climatological extreme events (Bishop 2008), which the old BMA method exhibits, is overcorrected. Therefore, this approach in fact performs worse than *direct model output* for these short forecast ranges.

For the most part, observations have significantly higher resolutions than model forecasts. Because in the final step, all analog methods use well resolved historical observations, their forecasts provide the same resolution as these observations. This downscale ability is one of the biggest advantages of the analog techniques. However, this advantage was not used in the Lorenz96 system. Therefore it is very probable that for variables that are resolved poorly in NWP models, these methods in fact perform better than in the present study. However, because of the considerable higher dimensionality of the true atmosphere, it might be harder to find enough meaningful analogs.

For precipitation forecasts over the USA, Hamill and Whitaker (2006) already obtained very promising results, especially with the *analogs of a deterministic forecast* method. Hence, further investigation should be carried out to test and improve these methods, including for other meteorological variables.

# Bibliography

Anderson, L.J., 1996: Selection of initial conditions for ensemble forecast in a simple perfect model framework. *J. Atmos. Sci.*, **53**, 22–36.

Bishop, C.H., 2008: Bayesian model averaging's problematic treatment of extreme weather and a paradigm shift that fixes it. *Mon. Weather Rev.*, **136**, 4641–4652.

Bröcker, J., and L.A. Smith, 2007: Increasing the reliability of reliability diagrams. *Weather and Forecasting*, **22**, 651–661.

Carter, G., J. Dallavalle, and H. Glahn, 1989: Statistical forecasts based on the national meteorological center's numerical weather prediction system. *Weather Forecast*, **4**, 401–412.

Epstein, E.S., 1969: A scoring system for probability forecasts of ranked categories. *J. Appl. Meteorol.*, **8**, 985–987.

Fortin, V., A.C. Favre, and M. Said, 2006: Probabilistic forecasting from ensemble prediction systems: Improving upon the best-member method by using a different weight and dressing kernel for each member. *Quart. J. Roy. Met. Soc.*, **132**, 1349–1369.

Gneiting, T., A.E. Raftery, A.H. Westveld, and T. Goldman, 2005: Calibrated probabilistic forecasting using ensemble model output statistics and minimum crps estimation. *Monthly Weather Review*, **133**, 1098–1118.

Hamill, T.M., 2001: Interpretation of rank histograms for verifying ensemble forecasts. *Monthly Weather Review*, **129**, 550–560.

Hamill, T.M., and J.S. Whitaker, 2006: Probabilistic quantitative precipitation forecasts based on reforecast analogs: Theory and application. *Monthly*

*Weather Review*, **134**, 3209–3229.

Hamill, T.M., J.S. Whitaker, and S.L. Mullen, 2006: Reforecasts: An important dataset for improving weather predictions. *Bulletin of the American Meteorological Society*, **87**, 33–46.

Hamill, T.M., J.S. Whitaker, and X. Wei, 2004: Ensemble reforecasting: Improving medium-range forecast skill using retrospective forecasts. *Monthly Weather Review*, **132**, 1434–1447.

Hersbach, H., 2000: Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather Forecast*, **15**, 559–570.

Hoeting, J.A., D.M. Madigan, A.E. Raftery, and C.T. Volinsky, 1999: Bayesian model averaging: A tutorial (with discussion). *Stat. Sci.*, **14**, 382–401.

Lagarias, J., J.A. Reeds, M.H. Wright, and P.E. Wright, 1998: Convergence properties of the nelder-mead simplex method in low dimensions. *SIAM Journal of Optimization*, **9**, 112–147.

Leamer, E.E., 1978: *Specification Searches*. Wiley.

Lorenz, E.N., 1996: Predictability - a problem part solved. *Proceedings, Seminar on Predictability ECMWF*, **1**, 1–18.

Lorenz, E.N., 2004: Designing chaotic models. *Journal of the Atmospheric Science*, **62**, 1574–1587.

Mao, Q., R. McNider, S. Mueller, and H.M. Juang, 1999: An optimal model output calibration algorithm suitable for objective temperature forecasting. *Weather Forecast*, **14**, 190–202.

Matheson, J.E., and R. Winkler, 1976: 1976. *Manage. Sci.*, **22**, 1087–1095.

Murphy, A.H., 1973: A new vector partition of the probability score. *J. Appl. Meteorol.*, **12**, 595–600.

Raftery, A.E., T. Gneiting, F. Balabdaoui, and M. Polakowski, 2005: Using bayesian model averaging to calibrate forecast ensembles. *Monthly Weather Review*, **133**, 1155–1174.

Roulston, M.S., D.T. Kaplan, J. Hardenberg, and L.A. Smith, 2003: Using medium-range weather forcasts to improve the value of wind energy production. *Renewable Energy*, **28**, 585–602.

Roulston, M.S., and L.A. Smith, 2003: Combining dynamical and statistical ensembles. *Tellus*, **55A**, 16–30.

Stephens, G.L., 2005: Cloud Feedbacks in the Climate System: A Critical Review. *Journal of Climate*, **18**, 237–273.

Wang, X., and C.H. Bishop, 2005: Improvement of ensemble reliability with a new dressing kernel. *Quart. J. Roy. Met. Soc.*, **131**, 965–986.

Wilks, D.S., 2005: Effects of stochastic parametrizations in the lorenz '96 system. *Quarterly Journal of the Royal Meteorological Society*, **131**, 389–407.

Wilks, D.S., 2006a: Comparison of ensemble-MOS methods in the Lorenz '96 setting. *Meteorological Applications*, **13**, 243–256.

Wilks, D.S., 2006b: *Statistical methods in the atmospheric sciences* (2nd ed.), Volume 91 of *International Geophysics Series*. Academic Press.

Wilks, D.S., and T.M. Hamill, 2007: Comparison of ensemble-mos methods using gfs reforecasts. *Monthly Weather Review*, **135**, 2379–2390.

Wilson, L.J., and M. Vallée, 2002: The Canadian Updateable Model Output Statistics (UMOS) System: Design and Development Tests. *Weather and Forecasting*, **17**, 206–222.

# List of Figures

2.1 Schematic illustration of the Lorenz96 system . . . . . . . . . . . . . 6

2.2 Climatological distribution of the predictand X . . . . . . . . . . . . 8

2.3 Example characteristic forms of reliability diagrams . . . . . . . . . . 15

3.1 Schematic illustration of the *no-forecast-model* method . . . . . . . . 25

3.2 Schematic illustration of the *analogs of a deterministic forecast* method 26

3.3 Problem of the rms analogy criterion . . . . . . . . . . . . . . . . . . 28

4.1 RPSS of DMO, *ensemble dressing*, NGR, BMA and *logistic regression* 30

4.2 BSS of DMO, *ensemble dressing*, NGR, BMA and *logistic regression* . 32

4.3 Reliability diagrams of DMO *ensemble dressing*, NGR, BMA and *logistic regression* for $q_{1/10}$ and $T = 1$ . . . . . . . . . . . . . . . . . . 33

4.4 Reliability diagrams of DMO, *ensemble dressing*, NGR, BMA and *logistic regression* for $q_{1/3}$ and $T = 1$ . . . . . . . . . . . . . . . . . . 34

4.5 Reliability diagrams of DMO, *ensemble dressing*, NGR, BMA and *logistic regression* for $q_{1/10}$ and $T = 4$ . . . . . . . . . . . . . . . . . 34

4.6 Reliability diagrams of DMO, *ensemble dressing*, NGR, BMA and *logistic regression* for $q_{1/3}$ and $T = 4$ . . . . . . . . . . . . . . . . . 35

4.7 RPSS of DMO, *logistic regression*, NFM (rankdiff), ADF (rms and rankdiff) and *analog dressing* (rms) . . . . . . . . . . . . . . . . . . . 36

4.8 BSS of DMO, *logistic regression*, NFM (rankdiff), ADF (rms and rankdiff) and *analog dressing* (rms) . . . . . . . . . . . . . . . . . . . 38

4.9 Reliability diagrams of NFM (rankdiff), ADF (rms, rankdiff and rankdiff4) and *analog dressing* (rms and rankdiff4) for $q_{1/10}$ and $T = 1$ 40

4.10 Reliability diagrams of NFM (rankdiff), ADF (rms, rankdiff and rankdiff4) and *analog dressing* (rms and rankdiff4) for $q_{1/3}$ and $T = 1$ 40

4.11 Reliability diagrams of NFM (rankdiff), ADF (rms, rankdiff and rankdiff4) and *analog dressing* (rms and rankdiff4) for $q_{1/10}$ and $T = 4$ 41

4.12 Reliability diagrams of NFM (rankdiff), ADF (rms, rankdiff and rankdiff4) and *analog dressing* (rms and rankdiff4) for $q_{1/3}$ and $T = 4$ 41

4.13 RPS of the NFM method for different ensemble and training sample sizes . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 43

4.14 RPS of the ADF method for different ensemble and training sample sizes . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 44

4.15 Reliability diagrams of the ADF method with training sample size n=10000 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 45

4.16 RPS of the *analog dressing* method for different daughter ensemble and training sample sizes . . . . . . . . . . . . . . . . . . . . . . . . 45

4.17 RPS of the 3 basic analog methods as function of the 'region'-size . . 46

# Acknowledgments

At this point I want to thank my supervisor Prof. Dr. Georg Mayr for his support, advice and inspiring discussions, which were essential for the progress of this diploma thesis.
Furthermore, I am indebted to all my friends and colleagues for the great time I had studying in Innsbruck.
And last but not least, I am deeply grateful to my family, especially to my parents Hemma and Klaus for their great support - not only financial. Thank you!

# Curriculum Vitae

**Name:** Jakob Messner
**Born:** 13 July 1985 in Rum, Austria
**Address:** Innrain 71, 6020 Innsbruck, Austria
**Email:** jakob_m@gmx.at

---

## Education and professional training:

| | |
|---|---|
| 2008–2009 | Diploma thesis under the guidance of Ao. Univ. Prof. Dr. Georg Mayr, Department of Meteorology and Geophysics, University of Innsbruck: *"Probabilistic forecasting using analogs in the idealized Lorenz96 setting"*. |
| 2008–ongoing | Bachelor degree course Technical Mathematics at the University of Innsbruck. |
| 2007–2008 | One semester abroad at the Complutense University of Madrid, Spain. |
| 2004–2009 | Diploma study at the University of Innsbruck. *Master of Natural Science (Magister rerum naturalium)* in Meteorology. |
| 2003–2004 | Civil service as paramedic, Red Cross Reutte. |
| 1995–2003 | Bundesrealgymnasium Reutte. *Matura.* |

---

## Teaching experiences:

| | |
|---|---|
| 2008–2009 | Tutor for 'Einführung in die Mathematik' I and II (Bachelor degree course Geo- und Atmospheric Science). |
| 2009 | Tutor for 'Stochastische Prozesse' (Master degree course Atmospheric Science). |