universität wien

MASTERARBEIT / MASTER'S THESIS

Titel der Masterarbeit / Title of the Master's Thesis

Verification in Complex Terrain with Ensemble-Analysis

verfasst von / submitted by

Simon Wolfgang Kloiber, BSc

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien, 2017 / Vienna, 2017

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	A 066 614
Studienrichtung lt. Studienblatt / degree programme as it appears on the student record sheet:	Metorologie
Betreut von / Supervisor:	o. Univ.-Prof. Dr. Reinhold Steinacker

Abstract

The aim of the Mesoscale Verification Inter-Comparison over Complex Terrain (MesoVICT) project is to compare spatial verification methods with each other to gather information about proper ones. This is done for deterministic and ensembles forecasts. They are compared with a certain observation. However, the MesoVICT project provides two observation ensembles, which are computed from the University Vienna. Here, the Vienna Enhanced Resolution Analysis (VERA) is the basis for ensembles were station observation uncertainties and spatial patterns are represented. They are used for a so far little discussed topic, the verification from two ensembles against each other. Since this issue is complex, the spatial part is postponed and a point to point verification method inter-comparison, in the spirit of the MesoVICT project, is done.

The forecast ensemble is provided by the weather service of Emilia-Romagna (COSMO-LEPS). As parameters the wind speed (V), the potential temperature (Θ), precipitation (RR) and the pressure (p, p_{sf} or p_{md}) are chosen.
Before starting with the evaluation, distribution tests are performed for the ensembles (Lilliefors-, Jarque-Bera- and Anderson-Darling-Test). The conclusion is that non parametric statistics are useful, because the Gaussian distribution is not valid all the time.
The verification is done with distance measurements from probability density functions (PDF), cumulative distribution functions (CDF) and via histograms. Six methods are selected. The integrated quadratic distance (IQD) uses two CDF's and measures the sum of squared differences. The perfect receiver operating characteristic comparison (PROCC) needs the observation and forecast ensemble represented as PDF's. This score is rather complex, but based on a comparison with a perfect result. Probability vectors are required for the Kullback-Leibler Divergence (KLD). It provides the difference of those. The Euclidean, Seorgel and Lorentzian distances are based on histograms and the related probabilities in each bin. The used methods can also be applied for a comparison of, for example, two forecast models.
For satisfying the non parametric statistics restriction kernel density estimates are applied. This approach fits the PDF's and CDF's to the data without losing any previous properties.
The ensemble verification with different parameters showed, that it is reasonable to use wind speed and potential temperature. The precipitation and the pressure have to be modified or evaluated with different methods. A table provides an overview of the used verification methods and defines a judgmental interval for each score. The kernel density estimate improves as well as deteriorates the results.

Additionally, the *score specific uncertainty* and the *uncertainty between different observations* are calculated. Those estimates interact with each other. If the *score specific uncertainty* is small, the location difference between the observation ensembles are more important for the *uncertainty between different observations*. In the other hand, a big extent, due to the *score specific uncertainty* suppresses the location influence from the observations in the *uncertainty between different observations* and the *score specific uncertainty* is more prominent.

Finally, different observation ensemble possibilities, a brief overview for a model comparison and extreme value approaches are shown.

Zusammenfassung

Das Ziel vom Mesoscale Verification Inter-comparison over Complex Terrain (MesoVICT) Projekt ist es, räumliche Verifikationsmethoden miteinander zu vergleichen, um Informationen über die sinnvollsten Verfahren zu erhalten. Dies wird mittels Deterministischer- und Ensemblevorhersagen gemacht, welche gegen eine Beobachtung verifiziert werden. Das MesoVICT Projekt stellt allerdings zwei Beobachtungsensembles, von der Universität Wien, zur Verfügung. Dabei wird die Vienna Enhanced Resolution Analysis (VERA) als Basis für Ensembles, welche Stationsunicherheiten und räumliche Muster beinhalten, verwendet. Diese werden für die, bis jetzt wenig beachtete, Verifikation zweier Ensembles gegeneinander verwendet. Da dieses Thema sehr komplex ist, sind die räumlichen Verifikationsmethoden hinten angestellt. Zuerst wird eine Punktverifikationen durchgeführt, bei der der Methodenvergleich im Vordergrund steht.

Die Vorhersage wird vom italienischen Wetterdienst Emilia-Romagna (COSMO-LEPS) bereitgestellt. Als Parameter sind Windgeschwindigkeit (V), potentielle Temperatur (Θ), Niederschlag (RR) und Druck (p,p_{sf} oder p_{md}) vorhanden.
Bevor mit den Vergleichen begonnen wird, muss die Verteilung der Ensembles durch Tests festgestellt werden (Lilliefors-, Jarque-Bera- und Anderson-Darling-Test). Die Ergebnisse zeigen keine andauernde Gaußsche-Verteilung, was nicht parametrische Verfahren für die Verifikation notwendig macht.
Die Evaluierung selbst wird als Abstandsmessung für probability density functions (PDF's), cumulative distribution functions (CDF's) und Histogrammen durchgeführt. Insgesamt sind sechs Methoden als sinnvoll befundet worden. Die integrated quadratic distance (IQD) verwendet zwei CDF's und misst die Summe der quadratischen Differenzen. Der perfect receiver operating characteristic comparison (PROCC) benötigt das Beobachtung- und das Vorhersageensembles repräsentiert als PDF's. Dieser Score ist ziemlich kompliziert. Am Einfachsten kann er als Vergleich mit einem perfekten Ergebnis dargestellt werden. Wahrscheinlichkeitsvektoren sind notwendig für die Kullback-Leibler Divergence (KLD). Hier wird die Differenz von diesen gebildet. Die Euclidean, Seorgel und Lorentzian distance basiert auf Histogrammen und den damit verbundenen Wahrscheinlichkeiten in den Bins. Die gefundenen Methoden können zum Beispiel auch für den Vergleich zweier Vorhersageensembles verwendet werden.
Um der nicht parametrischen Statistik Rechnung zu tragen wird der kernel density estimate angewandt. Dieser Ansatz passt die PDF's und CDF's an den Datensatz an, ohne jedoch ihre Eigenschaft zu verändern.
Die Verifikation mit den verschiedenen Parametern hat ergeben, dass die Windgeschwindigkeit und die potentielle Temperatur am brauchbarsten für den hier durchgeführten Ensemblevergleich sind. Der Niederschlag und der Druck müssen mit anderen Verifikationsmethoden betrachtet werden. Mittels einer Tabelle wird ein Überblick der Verifikationsmaße angeboten. Diese gibt Entscheidungshilfen, wie das Ergebnis jeder Methode bewertet werden kann, an. Der kernel density estimate beeinflusst das Ergebnis sowohl positiv als auch negativ.

Zusätzlich werden die *score specific uncertainty* und die *uncertainty between different observations* berechnet. Beide Verfahren interagieren miteinander. Dadurch wird bei einer kleinen *score specific uncertainty* die Lage der Beobachtungsensembles zueinander wichtiger für die Berechnung der *uncertainty between different observations*. Umgekehrt beeinflusst eine breite *score specific uncertainty* die *uncertainty between different observations* mehr, als die Lage der Beobachtungen zueinander.

Abschließend sind verschiedene Beobachtungsensembles, ein Überblick für eine mögliche Modellverifikation und die Extremwert-Theorie Thema.

Inhaltsverzeichnis

Tabellenverzeichnis

Abbildungsverzeichnis

1. Introduction

1.1 Motivation

One big and important part in meteorology is the weather forecasting. Whenever a prediction about the future state of the atmosphere is needed, meteorologists, or others, can use two kinds of weather models. The most common approach yet is the deterministic. Here, past and present information are used to calculate one solution for the future. The ensemble approach on the other hand, gives multiple results for the upcoming weather. Therefore, the input variables getting perturbed and provide a set of solutions. But no matter what approach is used, it is always important to have knowledge about the accuracy of these weather models. That is why both approaches are compared with a certain observation or, occasionally, reanalysis. This process is called verification. Common methods to verify deterministic models are the Root Mean Square Error (RMSE), Mean Absolute Error (MAE) or the use of contingency tables. For the ensemble forecast the Continuous Ranked Probability Score (CRPS), the Brier Score (BS) or the reliability diagram are present adequate approaches (Wilks, 2011, pp. 301).

In this thesis a new verification approach is used. Instead of having one observation, multiple realizations for one observation, an ensemble, are available. There are no common methods for verifying both kind of weather models with an observation ensemble. That is why the first part of the thesis is about finding useful methods in order to make a verification possible (Section 3).

The topic, observation uncertainty in the verification, is often ignored with the argument that the random observational errors are significantly smaller than the standard deviation of random errors (Gorgas and Dorninger, 2012a). But with the availability of more detailed weather models the interest of having knowledge about the observation uncertainty is growing. Now, with the observation ensemble on hand, a calculation of uncertainty due to the observation is possible. This topic is discussed in Section 4.1. The aim is to get a useful method to estimate the uncertainty for common and verification methods from the first part of the thesis.
The production of observation ensembles is highly experimental and no claim to be right is made, because of that, a comparison method for verification results calculated with different observational inputs is necessary (Section 4.2). The results from both uncertainty estimates (Sections 4.1 and 4.2) getting compared with each other (4.3).

At the end of the thesis a short overview of recent options for observation uncertainty, an information about high impact weather verification and a forecast product outlook is provided in Section 5.

1.2 MesoVICT

Along with the improvement of weather models, the number of verification methods are growing too. New features like the comparison of objects have been developed within the process of verifying, for example, precipitation. That is the reason why the Spatial Forecast Methods Inter-Comparison Project (ICP) was founded. The aim was to understand and quantify different spatial verification methods with predefined fields in a simple environment for deterministic precipitation forecasts (Gilleland et al., 2009; Dorninger and Gorgas, 2013). After finishing that phase, the Mesoscale Verification Inter-Comparison over Complex Terrain (MesoVICT) project was established (Dorninger et al., 2013). This is referred to as the second phase and has multiple aims:

- Use real cases, not only predefined data and investigate the feasibility of using other parameters and ensemble forecasts.
- Understand the influence of complex terrain itself and in combination with different weather scenarios.
- Investigate the influence of observation uncertainty in the verification.
- Provide specific information about methods and motivate weather services to use new and high resolution numerical forecasts.

To achieve them effectively the project is spited into different tiers (see Figure 1).

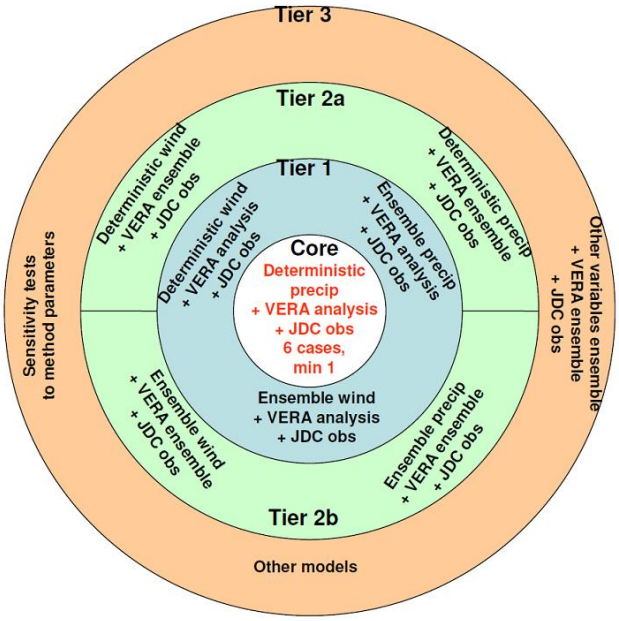


Figure 1: Schematic overview from the MesoVICT project. Tier 2 and 3 have a special interest for the the thesis (Dorninger et al., 2013).

The test area is Central Europe, including the Alps and parts of the Mediterranean Sea, which fits the requirements for a complex terrain quite perfectly. Furthermore, two big measurement campaigns took place between June and November 2007 in this area, which provides a huge set of observations additional to the Global Telecommunication System (GTS) stations. As a result, the Joint D-PHASE and COPS (JDC) data set was created (Section 1.3.1). This big data collection is used to verify forecast models and also provides more information for observational analysis products. To verify the behavior of certain verification methods during the measurement campaigns, six periods (cases) with specific meteorological conditions are chose.

The main topic, the so called core, is about verifying a deterministic precipitation forecast for all six meteorological cases. The next topic or tier one is about an additional wind verification, including the ensemble forecast approach for wind and precipitation. Again all six cases are part of the analysis. The second and third tier are important for this thesis. Here, the observation ensemble is in use and different parameters are verified. Now, only the first case, or core case, for the meteorological conditions, is available. That is because the observation ensemble was only produced for this time period (Section 1.3.1).
The core case took place from 20-22 June 2007. Central Europe and the Alpine region is just behind the center of a ridge, which leads to warm air advection and convective events on the 20th of June. On the 21st a cold front is approaching the Alps from northwest. The front triggers prefrontal convective

precipitation. Together with the passage throughout the 22nd strong westerly winds occur. As a result, the front is to weak for crossing the Alps and disperses.

MesoVICT is a community project about verification methods inter-comparison and not a model inter-comparison. The point is not to determine which method is best, it is about which methods are suitable for which types of data, forecasts and application field.

1.3 Data

As already mentioned in the previous Section 1.2 two big measurement campaigns are the reason for choosing this area and time. The collected data can be used as station data (irregular distributed) and interpolated on a regular grid. Since Numerical Weather Prediction (NWP) systems using a regular grid, it is better to have organized observations. You have to keep in mind, that the verification should be carried out with the same grid size. Four parameter are used in the evaluation process and specific cities and areas are investigated in detail. More about that in Section 1.4.

1.3.1 Observation

JDC

The GTS provides the latest weather information on a well covered network around the world, but for mesoscale analysis more observations are necessary. A lot of national and regional weather services have more surface observations than used in the GTS-network and together with some additional stations during the two measurement campaigns a very dense network is created (see Figure 2).

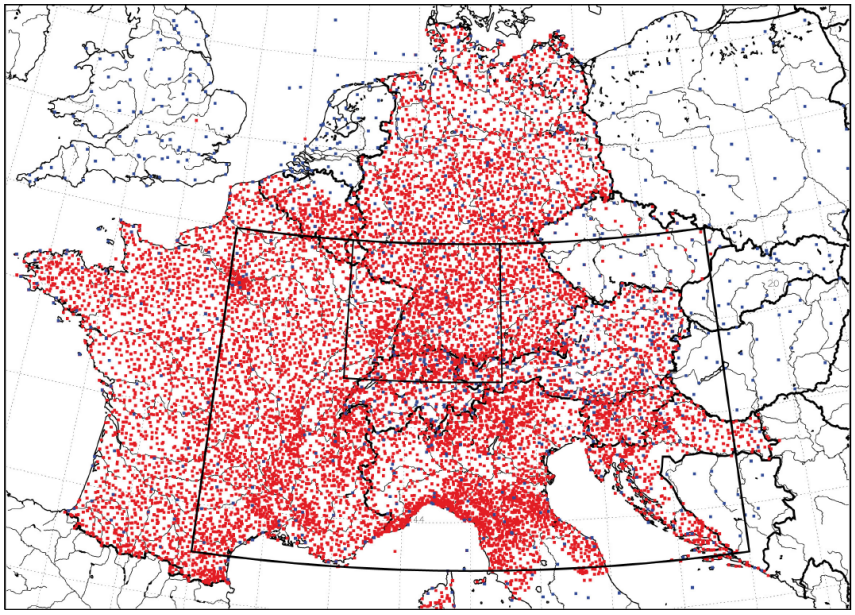


Figure 2: The blue dots are GTS stations and the red dots Non-GTS stations. Together they represent the JDC data set. The bigger frame shows the D-PHASE and the smaller one the COPS region (Dorninger, M. et al., 2009).

The two big campaigns are the Demonstration of Probabilistic Hydrological and Atmospheric Simulation of flood Events in the Alpine region (D-PHASE) (June-November 2007) as part of the Mesoscale

Alpine Program (MAP) and the Convective and Orographically-induced Precipitation Study (COPS) (June - August 2007). Combined they create the JDC data set (Dorninger, M. et al., 2009). As the name already suggests, the aim of the MAP D-PHASE was the evaluation of different weather simulations during precipitation and flooding events in the Alps. The COPS study is slightly smaller and covered only the Schwarzwald region. However, the aim was also to develop a high-resolution observation network. Those two data sets are used for model validation as well as data assimilation in their campaigns. Altogether around 12000 stations are available, which is equal to a mean station distance of 16 km. Not every station covers all parameters, some of them, due to the purpose of the COPS and D-PHASE campaign, cover only precipitation. Despite that fact, the amount of data for the other parameters still exceeds the normal one. That is why the JDC data set is used for the analysis and forecast fields in the MesoVICT project and consequently in this thesis.

VERA

The Vienna Enhanced Resolution Analysis (VERA) (Steinacker et al., 2000) provides model independent gridded surface observations. It is a real-time mesoscale analysis developed at the Department for Meteorology and Geophysics at the University of Vienna. Normal station observations are irregular distributed in space. This is problematic both representation- and verification-wise, because NWP outputs are regular distributed. To get the observation data on a regular grid an interpolation is necessary. The VERA mesh size is 8 km.

The VERA uses the thine-plate spline method. Here, the integral of the squared second spatial derivative is minimized. That idea assumes the natural behavior of the lowest cost-lost ratio. Meteorological parameters follow this principle quite well, for the simple reason that they want to have the lowest curvature between two points. Before an interpolation between two points is possible, an interpolation to the points has to be executed. Therefore, the same approach can be used. Now, the station observations are on a regular grid and the meteorological fields are properly interpolated.
Nonetheless, sometimes surface observations have an unusual value, which results in a wrong analysis. To prevent the usage of stations with a variety of errors, a quality control (QC) scheme is applied (Steinacker et al., 2011; Mayer et al., 2012). The QC filters values which exceeding physical or climatological thresholds.
Another important feature is the fingerprint technique (Steinacker et al., 2006). In complex terrain some mesoscale phenomena are not resolved properly. The fingerprint technique provides additional physical knowledge beyond the scale of the station density and includes these phenomena. Alternative sources for downscaling can also be topic of fingerprint techniques. Like the use of satellite - and radar data or information about snow coverage and land/sea distribution.

The following steps summarize the order in which the VERA is produced:

1. **Data Quality Control:** Only useful station data is used.

2. **Interpolation via splines:** No first guess or background flied is needed.

3. **Fingerprint:** Resolve mesoscale phenomena.

Note, that in the verification process the normal VERA is often named VERA reference.

VERA Ensemble

The purpose of an analysis is to give information about the "true" state of the atmosphere. Since errors occur in every analysis, no matter if it is model-depended or independent, an estimate for their uncertainty is required. A common approach in the weather prediction but new in the context of analysis is the use of ensembles.

Gorgas and Dorninger (2012a) created observation ensembles by using the VERA. Therefore, several requirements had to be fulfilled. First of all, the method had to be feasible for an analysis where no error estimate is provided automatically. This is the case in VERA. Furthermore, the observation and interpolation error had to be represented in the ensemble spread. Besides, the spatial uncertainties had to be included. Note, temporal errors were not assumed directly in this study. Finally, only station data should be used as input. In a later stage, radar or satellite data, as two examples, could be applied.

To satisfy all requirements and including spatial and observational errors the production was divided into two parts.
For getting the representation, or also called observational, error different rudiments are necessary (see Figure 3, middle level). For this thesis two different residuals are computed, which, in further consequences, are responsible for the two observation ensembles. One is done with the standard deviation of the bias corrected VERA QC residuals of a one-month period (std). The second one is a reference estimate of the equal perturbation magnitudes for all stations (equ).
Additionally, information about the spatial patterns are requested (see Figure 3, upper level). Therefore, the 2D wavelet transformation is reasonable. A Principal Component Analysis (PCA) is also possible, but not performed for the ensembles in use. In the process of getting those spatial patterns random Gaussian perturbations are applied.
A combination of the gathered residuals are used as perturbations.

The following guideline provides the order in which the VERA ensembles are produced:

1. The station observations after the quality control without any perturbations are available.
2. The residuals "std" and "equ" for the observation uncertainty are calculated. See the text above for a detailed explanation of "std" and "equ".
3. For the spatial pattern points 3 – 7 are needed. At the beginning the 2D wavelet transformation is applied on the operational gridded VERA (see Section 1.3.1/VERA). The results are multiple wavelet-coefficients.
4. The most eminent scale of those coefficients is getting disturbed randomly. This is repeated 50 times with a random Gaussian generator. The results are 50 different fields of wavelet-coefficients.
5. After that, the perturbed wavelet-coefficients are used for the inverse transformation, back on the gridded VERA. As a consequence, 50 differently disturbed analysis fields (VERA) are made.
6. The difference between VERA undisturbed (condition at point 3, before the wavelet transformation) and the VERA fields after point 5 gives the spatial pattern. This pattern is still on the VERA grid. The aim is to get it for every station observations, like in point 1.
7. That is why the spatial patterns are getting interpolated on the observation locations.
8. Now, two observational uncertainties (point 2) and the spatial patterns (50 various) from point 7 on the station observation scale are available. One observation ensemble set up includes all 50

spatial patterns and a observation uncertainty (“std” or “equ”). Those are multiplied with each other to 50 different perturbations. This method is also repeated for the second observation uncertainty. The, so fund, two sets of perturbations are added separately to the station data (point 1). Now, two set of 50 different observations (ensemble members) are produced.

9. However, the QC can be applied on the new station observations. If this is the case “-qc” is added to the set up name. After that, the VERA is interpolated normally on the regular grid and available fingerprints are used (see Section VERA).

The main advantage is that, the spatial patterns and the observation uncertainty are included in the ensemble procedure. This very complex technique is the same for every parameter. Here, the VERA set up from 2008 is used in the calculation process. The results are two different VERA ensembles (“std” and “equ-qc”) with 50 members on the VERA grid (8 km) for multiple parameters in an hourly interval. The ensembles together with the VERA reference are free to download from the MesoVICT homepage (`http://www.ral.ucar.edu/projects/icp/`[02/01/2017]).

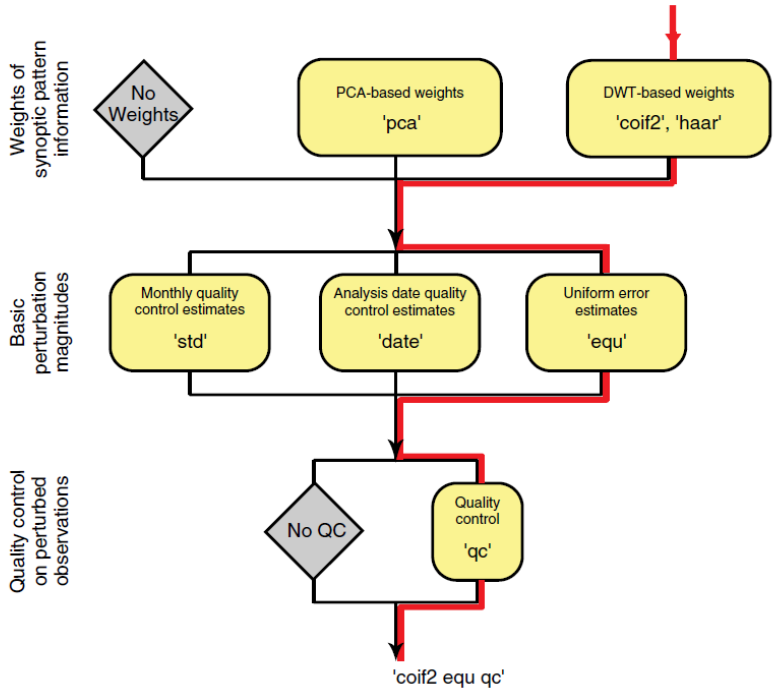


Figure 3: Schematic VERA Ensemble production. The first level shows possibilities for including the spatial patterns. Gorgas and Dorninger (2012a) used the Discrete Wavelet Transform (DWT). The middle level introduces observation errors. The bottom level pictures if a quality control is done (Gorgas and Dorninger, 2012a).

In Figure 3 an example for the “equ-qc” ensemble set up is shown (thick red line). The “std” one uses also the DWT branch, but is emphasized differently. Additionally, no quality control is applied.

During the study for the thesis one error was discovered. Member number 40 for the “equ-qc” set up on the 06/21/2007 at 7 Coordinated Universal Time (UTC) is missing. To avoid a complex programming solution, member number 39 was duplicated and is used as member 39 and 40.

The shown ensemble creation process is just one way how perturbations can be included. Sperka and Steinacker (2011) use a bias-correction method to collect uncertainty information. Those results could also be used in the just explained process. However, many ways of creating ensembles are possible. A short summery about different observation analysis is included in Section 5.1.

1.3.2 Forecast

The forecast model is provided by the Hydro-Meteo-Climate Service of the Environmental Agency of Emilia-Romagna (ARPA-SIMC), Italy and is also available via download from the MesoVICT homepage. ARPA-SIMC is part of the Consortium for Small-scale Modeling (COSMO) and calculates an Limited-Area Ensemble Prediction System (LEPS/COSMO-LEPS [CLEPS or CLE]). Each LEPS member is nested on selected members of the European Centre for Medium-Range Weather Forecasts (ECMWF) ensemble prediction system (Marsigli et al., 2005). Like the VERA, the CLEPS uses the JDC-data set as input. The forecast model gives 16 ensemble members on a mesh size of 10 km and a forecast range of 132 hours (3-hourly). However, the VERA uses a 8 km grid, which makes an interpolation necessary.
In Section 3 comparisons against a deterministic forecast are made. Since no control (main) run or deterministic model from the same provider are present, the mean of all 16 ensemble members is used as deterministic forecast.

1.4 Experiment setting

The model domain is smaller than the VERA domain. This is not a restriction, because the chosen verification range is nested in both domains (see Figure 4). The chosen frame has a size of 44-50° Nord and 8-18° East. However, the emphasis is on the complex topography. Embedded in this selection there are six subsections to specify error amplitudes to a region. The six subsections are South Germany (SG), Great Vienna (GV), West Alps (WA), East Alps (EA), Po Valley (PV) and the Adriatic (A) area. They are not shown in Figure 4. Nine cities are selected in order to study a specific error pattern over time (red dots in Figure 4). The cites are Vienna, Graz, Klagenfurt, Innsbruck (all Austria), Munich (Germany), Bolzano, Milano, Venezia (all Italy) and Zagreb (Croatia). Most of the results are referred to Vienna.

VERA and the CLEPS supply multiple parameters. In coordination with the MesoVICT tiers (Figure 1) precipitation and wind speed are investigated. Additionally, due to the meteorological importance, pressure and the potential temperature are verified too. Those four parameters are a direct model output and no post processing is needed. They are used to check, how and if an ensemble approach is useful.
The verification is done every three hours, because the model output is not available every hour, like it is with the VERA ensemble. This is why the precipitation RR measuring unit is millimeters per three hours (mm/3h). The wind speed V ($\mathrm{m\,s^{-1}}$) is calculated from the 10 m wind components u and v. The wind speed is a scalar V (see Equation 1.1).

$$V = \sqrt{(u^2 + v^2)}. \tag{1.1}$$

where:

u = 10m wind component along x-axis [$\mathrm{m\,s^{-1}}$]
v = 10m wind component along y-axis [$\mathrm{m\,s^{-1}}$]

The potential temperature Θ normally with K, here °C = K - 273,15, is calculated (1.2):

$$\Theta = T\left(\frac{p_0}{p}\right)^{\kappa}. \qquad with \qquad \kappa = \frac{R_L}{c_p} \approx \frac{2}{7}. \tag{1.2}$$

Figure 4: Verification area. Embedded in the red frame are six subsections (not shown) and nine cities (red dots). Those marks are used for a space and time verification. Vienna is additionally marked with an arrow. `https://upload.wikimedia.org/wikipedia/commons/3/38/Europe_topography_map_en.png` *[02/01/2017]*

where:

T = Temperature at certain level [K]
p_0 = Reference level (1000 hPa)
p = Pressure at certain level [hPa]
κ = Isentropic exponent
R_L = Gas constant for dry air [J kg^{-1} K^{-1}]
c_p = Specific heat for dry air [J kg^{-1} K^{-1}]

The VERA gives the MSL pressure p (hPa). The forecast, in the other hand, creates two solution for the pressure. One reduces the pressure to the sea level with the standard pressure reduction formula p_{sf}, the same as used in the VERA, see Equation 1.3 (Raith and Bauer, 2001, pp. 162). The other solution applies a model formula, which is different depending on the forecast model (p_{md}), and provides the pressure on the lowest model layer. This fact is important, if multiple models are compared. Because a verification against the standard reduction offers knowledge about the accuracy of a model. For the forecast it is essential to verify the specific method, due to the fact that these pressure values are shown on weather charts. Since this thesis is not about a model inter-comparison, but about verification inter-comparison both forecast pressure fields are only checked for consistency and not for accuracy.

$$p_{sf} = p \cdot \exp\left(\frac{\mathbf{g} \cdot z}{R_L \cdot T_v + R_L \gamma \frac{z}{2}}\right). \tag{1.3}$$

where:

p_{sf} = Mean sea level (MSL) pressure with standard formula [hPa]
$\mathbf{g}$ = Gravity acceleration [$\mathrm{m\,s^{-2}}$]
z = Absolute height [m]
T_v = Virtual temperature [K]
γ = Temperature gradient [$\mathrm{K\,m^{-1}}$]

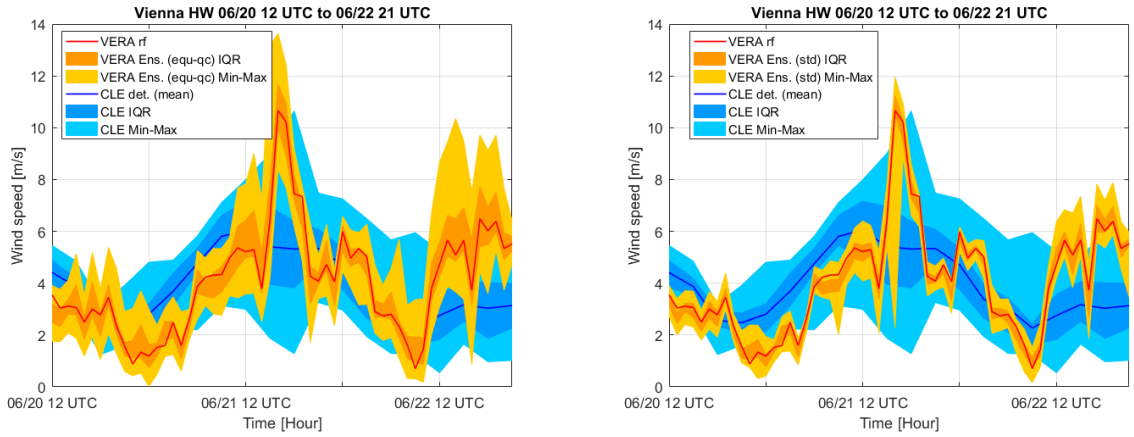
To summarize, for this study the MesoVICT core case 20-22 June 2007 is used. The aim is to find appropriate verification methods for comparing an observation- with a forecast ensemble (see Table 1). Furthermore, uncertainty estimates due to the observation and based on different observation models should be provided. As parameters precipitation (RR), wind speed (V), pressure (p,p_{sf} or p_{md}) and the potential temperature (Θ) are chosen.

Table 1: Summary from the properties of the observation (VERA) and forecast (CLE) ensemble.

Ensemble	Name	Member	Interval [h]	Resolution [km]
Observation	VERA ("std" and "equ-qc")	50	1	8
Forecast	CLEPS	16	3	10 interpolated on 8

2. Determine distribution

Before doing the verification a closer look at the data is necessary. At first the spread from both observation and forecast ensembles, together with the reference run from VERA and the mean (deterministic) CLE, are plotted (see Figure 5). The observation segments in those plots are yellow- and the forecast parts are blueish. The CLE starting time is similar to the start of the time-axis, 06/20/07 12 UTC. Wind speed is shown.



(a) Time series for the VERA ("equ-qc") and the CLE (initial time: 06/20 12 UTC) wind speed ensembles.

(b) Time series for the VERA ("std") and the CLE (initial time: 06/20 12 UTC) wind speed ensembles.

Figure 5: Time series of the VERA and CLE ensembles.The observational parameters are in yellowish and forecast ones blueish. The VERA has an hourly and the CLE a three hour interval.

The most prominent difference is the spread extent between those three ensembles. The CLE has the largest one, which implicates a lack of reliability. Both VERA ensembles are narrower than the CLE. Especially the "std" set up produces a very narrow ensemble. This leads to a nearly philosophical question. How much observation-wise spread is wanted? Is it better if the ensemble is wider, maybe even close to the forecast spread, or is a narrow one requested? However, the measurement, interpolation, representation and many more uncertainties should be represented properly. This question can not be answered yet, which makes the uncertainty estimate between different observations from Section 4.2 even more important.
Another distinction is the time interval. If a comparison between the one hourly VERA data and the three hourly CLE output is made, the forecast ensemble seems more smooth. Actually, the three hour interval makes the CLE more inertial. This is an important point for the verification discussion (see Section 3.3).
This example was realized with the wind speed. The potential temperature behaves quite the same. However, there are some exceptions concerning the precipitation and pressure, which are discussed in Section 3.

In this thesis the verification between two ensembles should be executed with histograms or Probability Density Functions (PDF's). For a histogram no distribution is assumed, the PDF in the other hand assumes a Gaussian or also called normal distribution, see Equation 2.1 (Wilks, 2011, pp. 87).

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right). \tag{2.1}$$

where:

x = Data
μ = Mean from x
σ = Standard deviation from x
σ^2 = Variance from x

The data is represented in the well known bell-shape, where the area under the curve is equal to one. Another important property is the standard deviation. Here, information about the amount of variation of the sample is provided. A standard deviation of plus/minus one from the mean indicates, that 68.27% of all data is in this interval. Plus/minus two and three σ suggests 95.45% and 99.73%.

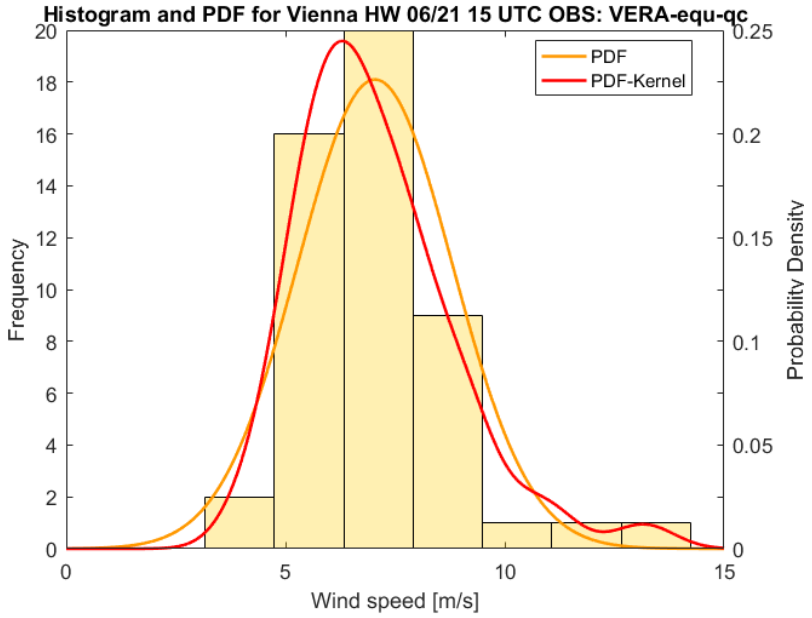


Figure 6: The histogram uses a Gaussian binwidth approximation and is compared with the organge curve. This is the PDF derived from the same data set. There is a sight disagreement between those visualizations products. The red curve is the fitted PDF, also calles Kernel density estimate (Nonparametric statistic, see Section 2.3). Here, the agreement is much better. The used grid point is Vienna at 06/12 at 15 UTC. The parameter is the wind speed.

The histogram is a graphical tool. The range of a sample is separated into equal intervals (bins). The entries of values in each interval are counted. Plot 6 shows an histogram-PDF-comparison for the wind speed at the grid point Vienna. A disagreement between the histogram (no direct assumption about the distribution) and the PDF is visible. The main issue when constructing a histogram is the binwidth (Wilks, 2011, pp. 33). If the interval is to width the histograms lean to be too smooth and misses important details. In the other hand, a too narrow bin makes an interpretation difficult. However, a good approach for the binwidth h is Equation 2.2 which is also used in Figure 6.

$$h \approx \frac{cIQR}{n^{1/3}}. \tag{2.2}$$

where:

c = Constant in the range 2.0-2.6 (Here, 2.6 for Gaussian bell-shape)
IQR = Interquartile range (IQR see Section 3.1.5)
n = Number of entries

This approach creates an indirect connection between the histogram and the Gaussian distribution. Despite all preparations, there is no perfect fit between the histogram and the PDF.

Another way to demonstrate a distribution is the empirical Cumulative Distribution Function (CDF). This measurement is related to the histogram and is displayed as a two-dimensional plot (see Figure 7). The shown test values are converted into probabilities. Those are summed up as far as the test values continue. The result is an irregular stair (Figure 7, blue line). The cumulative probability is on the y-axis. The related data is on the x-axis.

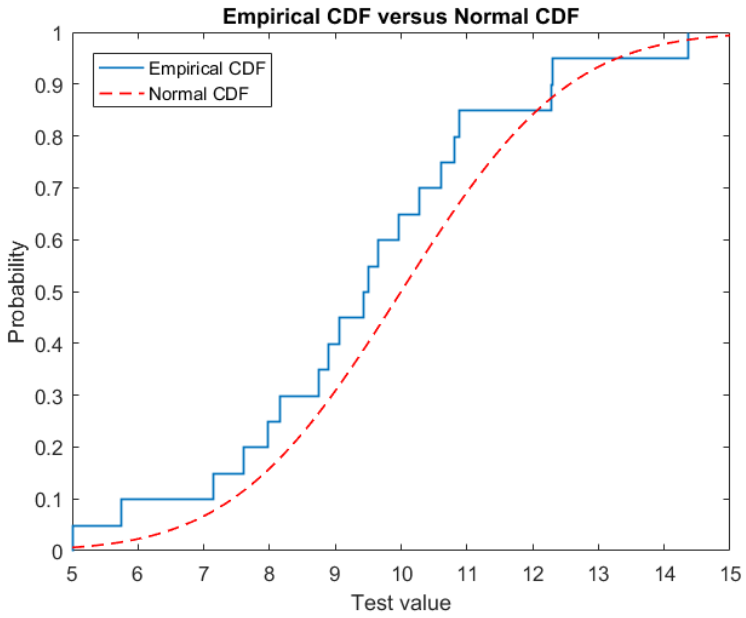


Figure 7: The Empirical CDF is displayed for a test value (blue). The Gaussian CDF is shown as a red dashed line.

There is also a second way to determine the CDF. In this case, the Gaussian assumptions is needed. After finding specific distribution parameters a smoother curve are produced (red dashed line in Figure 7). This one has an advantage over the empirical approach. Whenever two samples are about to get verified against each other and they do not have the same amount of entries, the empirical CDF produces a different number of steps. The comparison would not be clean. Since the observation and forecast ensemble did not have the same sample size (VERA 50 and CLE 16 members) the empirical approach is not used for the verification (see Section 3), but for nonparametric tests (see Section 2.1). Here, the uncertainty about the Gaussian assumption for both ensemble data is discussed.

2.1 Nonparametric tests

Nonparametric tests are used to get information about the distribution of a data set (Wilks, 2011, pp. 133). For this tests the programming language MATLAB® is applied. In this case, the sample in use has to be compared with a hypothetical distribution, also called *null hypothesis* H_0. If H_0, for this thesis a Gaussian distribution, is rejected, the data is not Gaussian. In that case the verification has to be done in a way that the distribution is not important.
Before doing those test, some parameters have to be specified. One of those is the *significance level* α. It gives the probability of the test for rejecting the *null hypothesis* correctly, in this case α is 0.05 or 5%. That means, if the test is done 100 times and the distribution is Gaussian, for five results the *null*

hypothesis is rejected anyway.
The *p-value* is an indicator for the probability that the given test result occur. If this value is too small the null hypothesis gets rejected. Note, that this is not the probability that H_0 is true, it is more like the confidence with which the *null hypothesis* is rejected or not.
In the process of getting the results from the tests in use (see Sections 2.1.2,2.1.3,2.1.4,2.1.5 and 2.1.6), the data is compared with a *critical value*. Here, the hypothesis that the data is normal distributed gets rejected if the test statistic is greater than the critical value obtained from a predefined table or via *Monte Carlo estimate* (Wilks, 2011, pp. 424). The *Monte Carlo estimation* is a numerical method to achieve results by repeated random sampling. Numerous repetitions are made until the standard error is as small as a given threshold. For Section 2.1.2,2.1.3 and 2.1.4 a predefined table is used where this algorithm has already been executed. A recalculation with a different threshold is possible.

2.1.1 Quantile-Quantile Plot (Q-Q Plot)

This visual product is based on the comparison between quantiles from a data sets and the *null hypothesis* (Wilks, 2011, pp. 115). Quantiles are the representation from a data sample, after it is divided into contiguous intervals with equal probabilities.This is a nonparametric measurement, because no knowledge about the distribution is necessary.

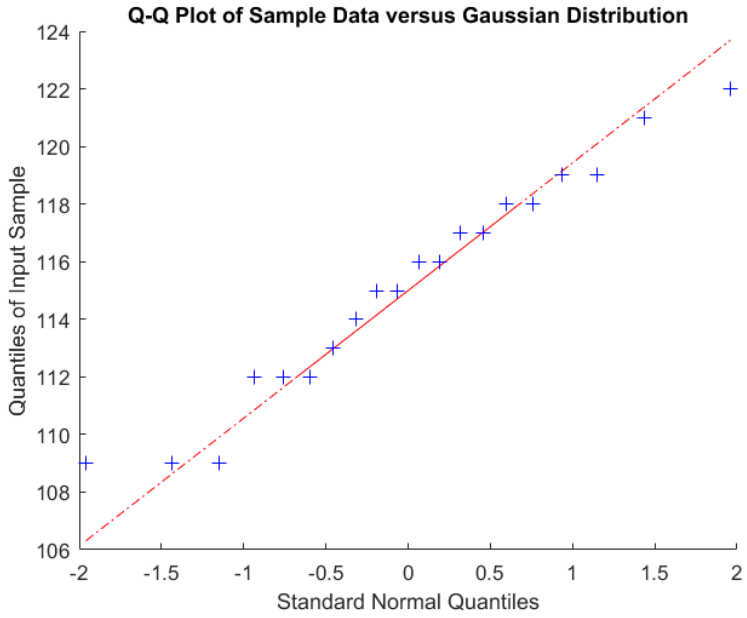


Figure 8: An example Q-Q plot. The red line shows the perfect agreement to the normal distribution.

Figure 8 shows an example for a parametric test. A data sample is verified against the *null hypothesis*, a Gaussian distribution. The perfect agreement would be at the red line. This method is comparable with the histogram, except that Q-Q plots not even an indirect connection to the normal distribution is made. However, there is still a problem, because no quantitative results about the data distribution is available. Therefore, some other tests are useful.

2.1.2 Kolmogorov-Smirnov-Test (KS-Test)

$$KS = \max|F_n(x) - F(x)|. \tag{2.3}$$

where:

$F_n(x)$ = Empirical CDF
$F(x)$ = Theoretical CDF

The one-sample Kolmogorov-Smirnov test (see Equation 2.3) identifies if the empirical CDF of the data is equal to the theoretical one (Wilks, 2011, pp. 151). The maximum difference between both is taken into count for the evaluation. KS gets compared with the *critical value*. If the result for KS is bigger than a given value, the *null hypothesis* gets rejected.

Strength:

- Stable, for continuous and discrete variables.
- Comparison not only for normal distribution, as long as continuous.

Weakness:

- Less selective if discrete (hypothesis less rejected).
- Not that powerful, Lilliefors-Test (see Section 2.1.4, adapted KS-test) and Anderson-Darling (see Section 2.1.6) better.

2.1.3 Chi-Square-Test (CS-Test)

$$\chi^2 = \sum_{i=1}^{N} \frac{(O_i - E_i)^2}{E_i}. \tag{2.4}$$

where:

i = Category
O_i = Observed in category
E_i = Expected in category

The chi-square test uses parameters estimated from the data to determine if a sample comes from a Gaussian probability distribution (Wilks, 2011, pp. 149). The test creates categories, where the observed and expected groups are compared (see Equation 2.4). If the matches are sufficiently large, the test statistic is valid. The *null hypothesis* gets rejected as soon as the test statistic is higher than a *critical value*. Therefore, the degree of freedom is important. It gets calculated by subtract the number of estimated parameters (μ, σ plus one by definition, so three) from the number of categories. The so found value and the test statistic result are needed in order to find the probability that the *null hypothesis* gets rejected.

Strength:

- Categorical scale.

Weakness:

- Must be independent.
- Bad if the sample size is less than 50.

2.1.4 Lilliefors-Test (L-Test)

$$L = \max|F_n(x) - F(x)|. \tag{2.5}$$

where:

$F_n(x)$ = Empirical CDF
$F(x)$ = Theoretical CDF

The Lilliefors test is an advancement to the one-sample Kolmogorov-Smirnov test (see Equation 2.5), as it got established especially for normal distributions (Wilks, 2011, pp. 151). To get a more accurate solution the *critical value* for this test has been computed by the *Monte Carlo* method. To find out how good the *Monte Carlo* estimate is, the standard error gets calculated (see Equation 2.6). The Lilliefors-Test in MATLAB® uses interpolated tables of the *critical values* pre-computed by the *Monte Carlo* simulation for a sample sizes less than 1000 and a *significance levels* between 0.001 and 0.50. A new computation can be made by changing the *Monte Carlo* Standard Error.

$$SE = \sqrt{\frac{p-(p-1)}{mcreps}}. \tag{2.6}$$

where:

p = Estimated *p-value*
$mcreps$ = *Monte Carlo* replications performed

Strength:

- Stable for continuous and discrete variables.
- Comparison not only for normal distribution as long as continuous.
- Good for a small sample.

Weakness:

- Jarque-Bera-Test (see Section 2.1.5) and Anderson-Darling (see Section 2.1.6) better.

2.1.5 Jarque-Bera-Test (JB-Test)

$$JB = \frac{n}{6}\left(s^2 + \frac{(k-3)^2}{4}\right). \tag{2.7}$$

where:

n = Sample size
s = Skewness
k = Kurtosis

The Jarque-Bera test uses the skewness and kurtosis to specify a normal distribution (see Equation 2.7). The test is applied to operate with parameters estimated from the data (Carlos M. Jarque and Anil K. Bera, 1987). For large sample sizes, the test switches to the chi-square distribution with two degrees of freedom. The *Monte Carlo* estimate computes the *critical value* (see Equation 2.6). Again some values are predefined.

Strength:

- Stable.
- Powerful.

– Good for small sample.

Weakness:

– Careful when using the JB test in the transition area between the Chi-Square test.

2.1.6 Anderson-Darling Test (AD-Test)

$$A_n^2 = -n - \sum_{i=0}^{n} \frac{2i-1}{n} [\ln(F(X_i)) + \ln(1 - F(X_{n+1-1}))]. \tag{2.8}$$

where:

n = Sample size
$\{X_1 < \ldots < X_n\}$ = Ordered sample data points

The Anderson-Darling test (see Equation 2.8) uses the gap between a hypothetical and an empirical CDF. It belongs to the family of quadratic empirical distribution function statistics (Stephens, 2006). However, the data has to be ordered. For making this test more sensitive to outliers and better at detecting irregularities from normality, the tails of the observation distribution are weighted.
To find out if the Gaussian assumption is valid the *p-value* for the hypothesis test with the specified *significance level* are compared with an already calculated table. This is the only test where no *critical value* is used. The table can be reproduced by using the standard error, which is estimated with Equation 2.6.

Strength:

– Valid for sample size at least n=8.
– Good for outliers.
– Powerful.

Weakness:

– Non know.

2.2 Results

The Q-Q plot is not used for testing the *null hypothesis* because a quantitative result is needed. Before applying the other nonparametric tests on both observation and the forecast ensemble, the tests themselves have to be checked for consistence. This has to be done since all tests are performed by a predefined MATLAB® function. Therefore, two normal distributed samples with the same number of entries as the two ensemble set ups (16 and 50 members) have to be created. The tests are performed 7900 times. This is the amount of points, which is included in the verification area (see Figure 4). The first sample has mean $\mu = 0$ and standard deviation $\sigma = 1$. The second one has $\mu = 10$ and $\sigma = 2$.
Different values for the *significance level* ($\alpha = 0.05$, $\alpha = 0.10$ or $\alpha = 0.20$) are tested. Furthermore, the necessity to recalculate the predefine tables for the L-, JB- and AD-tests are under investigation by using Equation 2.6 with the standard error ($SE = 0.1$ or $SE = 0.01$) .

All five tests show the same properties when changing the *significance level*. In combination with the literature (Wilks, 2011, pp.133) $\alpha = 0.05$ will be used in this thesis. Now, the behavior from each test with respect to the different distributions are discussed. Only the predefined tables and no specific

Monte Carlo estimation is applied.
For the first data set ($\mu = 0$ and $\sigma = 1$) with 50 entries, only the CS-test rejects the *null hypothesis* by 1% too often. For 16 entries, in the other hand, the CS-test completely ignores the *significance level*. All the other tests perform the same and as they should. For the second data set ($\mu = 10$ and $\sigma = 2$) and 50 entries the KS-test fails. The result would suggest, that no normal distribution is valid, which is wrong. The same result is shown for 16 entries. Again, the CS-test underestimates the *null hypothesis* for 50 entries and fails for 16. The L-, JB- and AD-tests perform perfect no matter what distribution or how big the data size is.
As a conclusion, only the Lilliefors-, Jarque-Bera- and Anderson-Darling tests are reasonable. Here, the *Monte Carlo estimation* has to be reviewed. The results with the existing tables, which were used so far, and calculations with $SE = 0.1$ and $SE = 0.01$ are compared. It is not important which one of both distribution is in use. Since for those three tests the results for both distributions are the same. A small difference occurs when verifying the estimate for a sample with 16 members. Here, the $SE = 0.1$ set up for the L-test rejects the *null hypothesis* 1% more often in comparison with $SE = 0.01$ and the given tables. For 50 members no differences are seen. This leads to the conclusion that no *Monte Carlo estimation* is necessary. The *Monte Carlo* computation is very time consuming and no additional information is acquired. For $SE = 0.1$ and the L-test for 16 entries the result is even worse than with the predefined table.

Before doing the nonparametric tests for the parameters used in this thesis, a summary about the experimental setting is provided. The nonparametric tests are here to show if the data distributed in a normal way. Therefore, the Lilliefors-, Jarque-Bera- and Anderson-Darling tests are useful. The *significance level* α is 0.05. No discrimination between the rejection reasons are made. This could be caused by the exceeding of the *critical value* or a too low *p-value*. The percentage shown in table 2, 3 and 4 are the agreement rate that the assumption of a normal distribution is valid. Here the wind speed for 06/21 00 UTC is demonstrated.

Table 2: Non parametric test results for the wind speed at 06/21 00 UTC. For the whole verification area.

Ensemble [V]	L-Test [%]	JB-Test [%]	AD-test [%]
VERA “equ-qc”	89	90	86
VERA “std”	88	91	86
CLE (+012)	92	94	91

All three ensemble have a quite good agreement with the Gaussian distribution, but still around 4 – 9% of the data points are not normal distributed. The next results in table 3 are for the potential temperature.

Table 3: Non parametric test results for the potential temperature at 06/21 00 UTC. For the whole verification area.

Ensemble [Θ]	L-Test [%]	JB-Test [%]	AD-test [%]
VERA “equ-qc”	92	91	92
VERA “std”	81	92	83
CLE (+012)	92	94	91

Here, nearly the same pattern is observed, around 3 – 14% of the data points are not Gaussian distributed. The precipitation and the pressure evolving a bit different and the normal distribution is not valid in a bigger amount of cases. These tests are also performed at different times. An example is

provided for the potential temperature at 06/22 12 UTC. The deviation in this case is around 3 – 20% (see Table 4).

Table 4: Non parametric test results for the potential temperature at 06/22 12 UTC. For the whole verification area.

Ensemble [Θ]	L-Test [%]	JB-Test [%]	AD-test [%]
VERA "equ-qc"	90	92	87
VERA "std"	74	90	75
CLE (+048)	98	100	99

The overestimation for the forecast ensemble is suspicious and suggests an error in the calculation. The VERA"-std" set up has the biggest difference. This could be caused by the narrow spread.

All this information together leads to the conclusion, that the data is not normal distributed at least for some time steps and some grid points. Altogether, nonparametric verification methods should at least be used as comparison, because otherwise the error caused by the wrong assumption can be become to big. A partial solution for this problem is provided in Section 2.3.

2.3 Kernel density estimate

Since the applied data is not normally distributed all the time, a nonparametric method has to be found. As previously mentioned the histogram is a nonparametric statistic. Here, the bins representing the data distribution in equal, intervals (bins). The data values are rounded to the center of the corresponding bin (see Figure 9, left picture, black rug plot on the x-axis). All bins are represented in a rectangular shape. Now, a different approach is used. The restriction of the centering is dropped and all entries are discussed by themselves. Additionally, every data point gets his own smoother shape, the kernel (see Figure 9, right picture, dashed lines). The whole process is called kernel density smoothing or Kernel Density Estimation (KDE) (Wilks, 2011, pp.34).

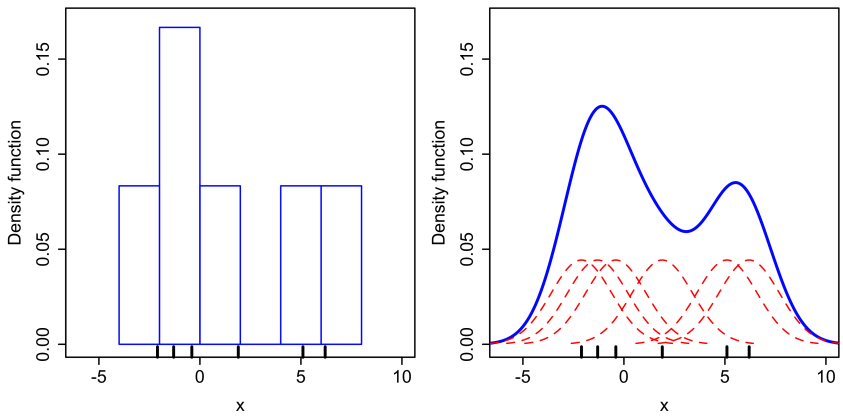


Figure 9: The left part shows a histogram (blue bins) with the data points explicitly shown on the x-axis (black dashes). On the right picture again the data points are on the x-axis. Here, the Gaussian kernels (red dashed lines) and the kernels density estimate for the data (blue line) is shown. `https://upload.wikimedia.org/wikipedia/en/4/41/Comparison_of_1D_histogram_and_KDE.png` *[02/14/2017]*

Those kernels are added to a curve, which has the same properties like a PDF (see Equation 2.9 and Figure 9, right picture, blue line). The shape of the kernels is provided by Equation 2.10 and is in this thesis Gaussian. Many other shapes are available as well, like the Triangular, Quartic (biweight) or Cosine. But the choice of the kernel type is normally less important than the declaration of the bandwidth h.

$$\hat{f}(x_0) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x_0 - x_i}{h}\right). \tag{2.9}$$

where:

x_i = Data
x_0 = Kernel estimated at a given value
h = Bandwidth (smoothing) parameter
K = Kernel (see Equation 2.10)
n = Number of kernels

$$K(t) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{t^2}{2}\right). \tag{2.10}$$

where:

$t = -\infty < t < \infty$

The smoothing parameter or bandwidth is chosen specifically for the Gaussian kernel (see Equation 2.11).

$$h = \frac{min\{0.9\sigma, \frac{2}{3}IQR\}}{n^{\frac{1}{5}}}. \tag{2.11}$$

where:

σ = Standard deviation
IQR = Interquartile Range (IQR)
n = Sample size

This approach can also be applied for the CDF. The so found PDF's and CDF's can be used in the verification without making any assumption about the distribution, but having the same properties like normal PDF's and CDF's. A way to avoid the kernel density estimate is to use applications where no distribution has to be assumed (e.g. histogram, Q-Q plot or a box plot [see Section 3.1.5]).

3. Verification

A weather forecast has to be checked against a observation or a good estimate of the true outcome (Wilks, 2011, pp. 301). This process is called verification or sometimes validation, or evaluation. The verification of weather models have already made since 1884. The aim is to control, improve and compare the forecast quality. But what makes a forecast good? Therefore, a list of attributes has been defined. The list is not exceptional, so differences depending on the source are existing.

1. The *Accuracy* is meant to specify the level of agreement between the forecast and the truth. The aim is to get a single number to quantify the quality of the forecast. The difference between the prediction and the truth, or observation, is the error. As lower an error is as greater is the accuracy.
2. The *Bias* (or *unconditional bias*) provides information about the connection between the mean forecast and mean observation. If the temperature is consistently too warm in comparison with the observation, the model has a positive *bias*.
3. The *Reliability* (or *conditional bias*) is the average agreement between the forecast and the observed distribution. If all predictions are compared together, then the overall *reliability* is the same as the *bias*. For specific values or categories, the *reliability* is equal to the *conditional bias*.
4. The *Resolution* refers to the availability of the forecast to resolve a set of events into subsets. If the outputs for a 10 °C and a 20 °C prediction are well different from each other the *resolution* is good. On the other hand, if these forecasts nearly have the same results the *resolution* is bad. Even if the forecast is not accurate, it is possible to have a good *resolution*.
5. The *Discrimination* shows the ability of the forecast to discriminate between observations. Whenever an observation (e.g. snow) occurs the model has to have a higher prediction frequency for that specific event (snow) than for a comparable one (sleet).
6. The *Sharpness* is an attribute just for the forecast alone. It describes the possibility of a model to predict extreme events. A forecast within the climatological range has a low *sharpness*. Extreme value prediction in the other hand provides a high *sharpness*. This attribute can be seen as good even though the forecast is wrong (like the *resolution*). Everyone can produce a sharp prediction; however, the difficulty is not to over-forecast such an event. In other words, not to make these extreme assumptions too often.
7. The forecast *Skill* refers to the relative accuracy of a forecast against some reference condition. Those references could be a random sample, persistence forecast or the climatology. The *skill* shows how much the forecast is improved in comparison with a reference. This has to be checked because “simple weather” is easier to forecast.
8. The *Uncertainty* gives information about variability of the observation. This is important since a big uncertainty makes a verification much trickier.

This last point is a major part of the thesis. Measurement errors are usually small in the process of recording, for example, the temperature or the pressure. Nevertheless, when using radar- or satellite-based measurements the topic should not be neglected. The representation error occurs when a scale mismatch is valid. This happens, for instance, when the rain is measured in a much smaller area (Rain Gauge) than the rain is predicted (Wilks, 2011, pp. 382). However, the VERA ensembles provide uncertainty estimates for numerous parameters.

Bowler (2006) tried to determine the effect of the observation uncertainty for a 2×2 contingency table. Therefore, he constructed an expected 2×2 table for an hypothetical error-free observation. The measured and the hypothetical tables are compared with each other. The errors in the observation make the calculated score less bad.
Ciach and Krajewski (1999) used radar-derived area-averaged data and divided the Mean Squared Error (MSE) in an MSE for the instrument error and one for the representativeness. The conclusion was, that the representativeness error was the dominant part of both MSE's and was more important at a shorter timescale, but it remained critical even for a longer accumulation time.
The observation uncertainty gained more interested when ensemble forecasts came into operation in the weather prediction. Whenever the observation error extent covers a big part of the ensemble spread an underdispersion is wrongly diagnosed. As remedy the observation error is simulated by random numbers and added to each ensemble member. The supplement has the characteristic of an error (Wilks, 2011, pp. 383).
Candille and Talagrand (2008) suggested to use probabilities or probability distributions for observations. The verification should still be done with methods like the Brier score, reliability diagram or the continuous ranked probability score (CRPS). This approach is the basis for this thesis (see Section 3.1). Note that, the appropriate observation probabilities still have to found, which is a difficult task and a comparison between different observation is necessary (see Section 4.2).

3.1 Verification methods

Since there are different forecast types the verification itself has to be adjusted. The methods are different if the forecast is probabilistic or deterministic. Now, another condition is taken into count, the observation uncertainty. In this thesis a probabilistic (ensemble) approach is used to represent the observational uncertainty.

Further the treatment between a space and/or time verification is an issue. Spatial methods compare the foretasted field with the observed one at a certain time. Whenever the comparison is done the next time step is under investigation. This is the main part oft the MesoVICT project (Gilleland et al., 2009). In Gilleland et al. (2009), no observation ensemble is used. The second approach is the time verification. A grid point or some station observations are verified over a time span. This is not the prime agenda for the project, but here observation ensembles are easier to include. In this thesis gridded data, to be more accurate, specific grid points are analyzed over a period of time. A spatial approach is needed as well. However, this topic is not part of the thesis. The third one, the space-time verification did not discriminate between space and time. This is quite often used in the climatology, for example, a monthly averaged global temperature anomaly. No research is done in that field by MesoVICT.
As already discussed in Section 2.1, verification methods are able to assume a distribution for the investigated data. If the assumption about the distribution is wrong, the used method would not be proper. In a fair amount of cases the normal distribution is accurate and no further specifications are needed. If this is not the case, nonparametric statistics must be applied (e.g. Kernel density estimates).

The easiest way to verify whether the forecast is true or not, is to look at the plotted data and compare the prediction with the actual condition. This is the subjective "eyeball" verification and can not be be done in an effective way. The solution is to use objective numeric values. The way of getting those numbers depends on the existing forecast and observation.

For a non-probabilistic prediction and observation the verification can be done continuously or categorical (discrete). A categorical measurement is the contingency table. There are different approaches, the first on is to identify an event and check if it occurs or not. An example of an event is precipitation above a threshold. The result is an entry in the linked 2×2 table (Wilks, 2011, pp. 306). It is also possible to expand the table to a $x \times x$ matrix for multiple categories. Continuous methods for example are the (mean) absolute error (see Section 3.1.1) and the box plot (see Section 3.1.5). The box plot is a bit of an exception, because it can also be used to show the distribution of an ensemble and is applied in their evaluation.
For the probabilistic forecast verification with discrete predictions against one observation the Brier Score (see Section 3.1.2) or the relaiabilty diagram (see Section 3.1.3) are important. It is also possible to adjust at least the reliability diagram for an observation ensemble (Newman et al., 2015). A quite often ignored approach in the probability verification is the comparison between continuous forecast and observation outputs (ensembles). Therefore, the distribution is represented as a PDF, CDF or histogram. The result is a evaluation or so also called distance measurement (see Section 3.1.6, 3.1.7, 3.1.8, 3.1.9), which is the main topic in this thesis.

3.1.1 (Mean) Absolute error (AE)

$$MAE = \frac{1}{n} \sum_{k=1}^{n} |p_k - o_k|. \tag{3.1}$$

where:

n = Number of forecast and observation pairs
k = Specific pair number k
p_k = Forecast
o_k = Observation

The MAE is a measurement for the *accuracy* and common for non-probabilistic continuous variables (Wilks, 2011, pp. 325). It is linear and shows the mean magnitude of the error, but it gives no information about the direction of the deviations. Equation 3.1 can be used as a spatial or time mean, but whenever this is not the case the deviation through n and the sum $\sum_{k=1}^{n}$ is neglected. The result is Equation 3.2.

$$AE = |p - o|. \tag{3.2}$$

This is the Equation which is used in the following verification. The perfect result is 0, whenever the discrepancy between the forecast and the observation grows the value of the AE grows as well.

3.1.2 Brier score (BS)

$$BS = \frac{1}{n} \sum_{k=1}^{n} (p_k - o_k)^2, \tag{3.3}$$

where:

n = Number of forecast and observation pairs
k = Specific pair number k
p_k = Forecast probability
o_k = Observation probability

For this measurement, probabilities in comparison with the mean squared error are applied (Wilks, 2011, pp. 331). This kind of mean square error uses a probabilistic forecast and a dichotomous observation as input. Therefore, the observation either not occur, zero in the calculation, or occur, one in this case. The perfect BS is 0 and the worst $BS = 1$, for a forecast that is always incorrect. Murphy (Murphy, 1973) showed that Equation 3.3 can be adjusted to 3.4.

$$BS = \frac{1}{n}\sum_{i=1}^{I} N_i(p_i - \bar{o}_i)^2 - \frac{1}{n}\sum_{i=1}^{I} N_i(\bar{o}_i - \bar{o})^2 + \bar{o}(1-\bar{o}). \tag{3.4}$$

where:

n	= Number of forecast and observation pairs
N_i	= Number of forecasts with the same probability category
i	= Specific pair number i
p_i	= Forecast probability
$\bar{o}_i$	= Observation probability
$\bar{o}$	= Climatology probability
$\frac{1}{n}\sum_{i=1}^{I} N_i(p_i - \bar{o}_i)^2$	= *Reliability*
$\frac{1}{n}\sum_{i=1}^{I} N_i(\bar{o}_i - \bar{o})^2$	= *Resolution*
$\bar{o}(1-\bar{o})$	= *Uncertainty*

This variation of the BS creates information about the *reliability*, *resolution* and *uncertainty*. One limitation is, that it is not useful to verify a single probabilistic forecast. Therefore, it is better to apply a set of probabilistic forecasts. The Brier score is sensitive to the climatology of the event. A disadvantage is that the BS is about to get too good without having any real skill. This can happens if an event is very rare.

3.1.3 Reliability diagram

The reliability diagram (see Figure 10) plots the forecast probability against the observed frequency from one event (Wilks, 2011, pp. 334). The forecast probability is divided into bins (for example, K is $0-5\%$, $5-15\%$, $15-25\%$ and so on).
The perfect *reliability* is achieved when the observation and the forecast produces a 45°-line. The deviation from the diagonal shows the *conditional bias*. If the curve lies below the perfect-reliability-line an over-forecasting (probabilities too high) is suggested. Vice versa if the points are above the line under-forecasting (probabilities too low) is valid. As closer the curve is to the climatology as less *resolution* it has. If the forecast represents the climatology (horizontal dashed line in Figure 10) it does not have any *resolution*. Sometimes an additional histogram is added. Here, each bin value is an indicator for the *sharpness* of the forecast.
The reliability diagram is restricted to the forecast. A good addition is the ROC (see Section 3.1.7), which is connected with the observations. Again, a single probabilistic forecast verification is not useful.
For the comparison of two ensembles against each other the observed frequency on the y-axis can be adjusted to the observed probability.

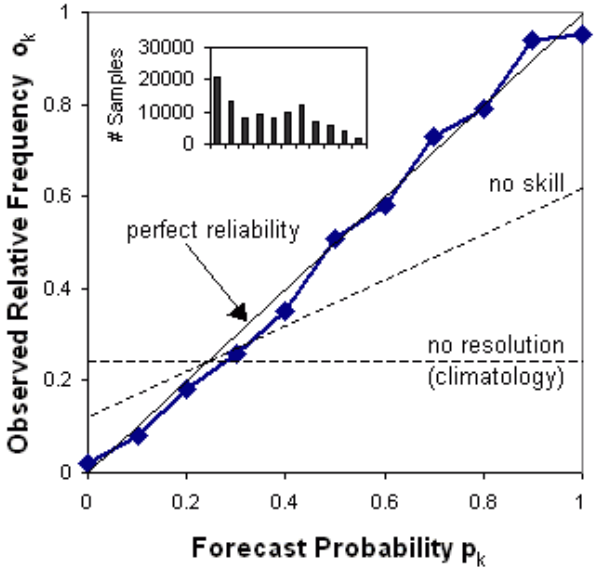


Figure 10: Reliability diagram example including a histogram with the sample size in each bin. The perfect result is marked with an arrow. The example line is blue. `http://www.cawcr.gov.au/projects/verification/verif_web_page.html` *[02/20/2017]*

3.1.4 Continuous ranked probability score (CRPS)

$$CRPS = \int_{-\infty}^{\infty} [P(x) - O(x)]^2 dx. \tag{3.5}$$

where:

$P(x)$ = CDF for the forecast
$O(x)$ = Dichotomous observation; occurrence = 1, not occurrence = 0

The CRPS corresponds with the (mean) absolute error, but not for a single valued forecast. It uses a continuous probabilistic forecast as input (Candille et al., 2007; Wilks, 2011, pp. 351), but still is a measurement for the *accuracy*. The forecast is presented as a cumulative distribution function (CDF). The observation is dichotomous and no distribution, it jumps from 0 to 1 at the point the observation appears (see Figure 14 the red line). The CRPS measures the sum of squared differences between the forecast CDF and the single value observation (see Equation 3.5). The perfect score is 0 and grows as further apart the two lines go. Since it is a quadratic measurement, bigger differences getting more severe represented in the score.

3.1.5 Box plot

The box plot or box and whisker plot is a graphical tool like the histogram. It is part of the nonparametric statistic and provides information about the used ensemble input (Wilks, 2011, pp. 29). The median is located in the center of the data, above and below there are 50% of the sample. One example should make it clear. The sample is $\{4,5,1,3,2\}$. This has to be ordered $\{1,2,3,4,5\}$ and, as a result, the median is 3. If the distribution is symmetric the median and the mean are equal, like in the example.
In Figure 11 also shown are the 3^{rd} (q_3) and 1^{st} (q_1) quartiles, also called as 75^{th} and 25^{th} percentiles.

Those two frame the interquartile range (IQR) where 50% of the data is located. The median is sometimes named the second quartile or the 50^{th} percentile. A quartile (percentile) cuts the data in contiguous intervals with equal probabilities (2.1.1). The third quartile, for example, separates the highest 25% of data from the lowest 75%.
The upper whiskers or maximum value shows how much of the sample is greater than $q_3 + 1.5 \times (q_3 - q_1)$. The lower whisker or minimum value is calculated with $q_1 - 1.5 \times (q_3 - q_1)$. This corresponds to approximately $\pm 2.7\sigma$ and 99.3 percent coverage if the distribution is Gaussian.
The points outside of the upper and lower whiskers are outliers. Sometimes they are divided in mild (between $1.5 \times IQR$ and $3 \times IQR$) and extreme outliers (above $3 \times IQR$). In this thesis no separation is used.
Additional information is gained with the use of notches. Within this thesis, the 95% confidence interval of the median is pictured.

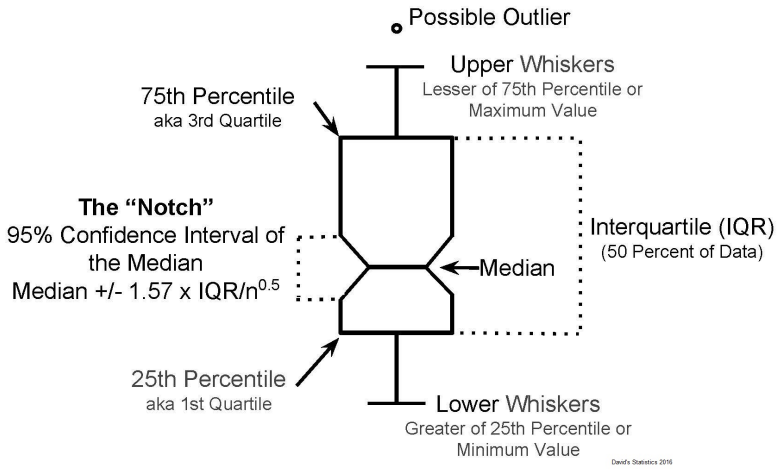


Figure 11: Explanation for a box plot with notches. Additionally the outliers can be separated into mild and extreme one. `https://sites.google.com/site/davidsstatistics/home/notched-box-plots` *[02/20/2017]*

An "eyeball" verification shows similarity between location, spread, and skewness of forecast and observed data. It can be used as an ensemble comparison.

3.1.6 Integrated quadratic distance (IQD)

$$d_{IDQ} = \int_{-\infty}^{\infty} [P(x) - O(x)]^2 dx. \tag{3.6}$$

where:

d = Distance measurement
$P(x)$ = CDF for the forecast
$O(x)$ = CDF for the observation

The IQD expands the CRPS (see Section 3.1.4) and is still an *accuracy* measure (Thorarinsdottir et al., 2013). For the normal CRPS a CDF is compared with a dichotomous observation. In this context, the observation is also probabilistic and represented as a CDF (see Equation 3.6). It measures the sum of squared differences, which means bigger errors are more relevant to the score than smaller ones. The

empirical cumulative distribution function is not used for the verification. The perfect result is zero and as bigger it gets, as worse it is.

3.1.7 Perfect receiver operating characteristic comparison (PROCC)

The Receiver Operating Characteristic or also called Relative Operating Characteristic (ROC) diagram is a diagnoses tool corresponding to the reliability diagram from Section 3.1.3 (Wilks, 2011, pp. 342). The ROC- in comparison to the reliability diagram is connected with the observations.
Here, for the PROCC a different approach is used. To avoid a possible disorder, no more references with the prior discussed ROC diagram are made. For this score, two PDF's are compared with each other in order to determine the *accuracy* (Metz, 1978). Therefore, the area under the two PDF's is spitted into four parts by a threshold, called "cutoff" (see Figure 12, plot above, black line). In Figure 12 in the upper plot two of those parts are marked. One is red striped. This is the "observation positive" area, because the red curve is an example for a PDF produced by an observation ensemble. The addition "positive" is used, due to the fact that it is on the positive side from the "cutoff" point of view. The same logic is applied for the blue striped area and the nomenclature of the "forecast positive" area. This PDF is from the connected forecast. The two still not named areas are on the other side of the "cutoff" value and not important in this thesis.
Now we have two suitable areas for one "cutoff" value. To avoid the arbitrary selection of thresholds, numerous "cutoffs" are moved over the range of the two PDF's. This results in numerous solutions for the "observation positive" and "forecast positive" areas.

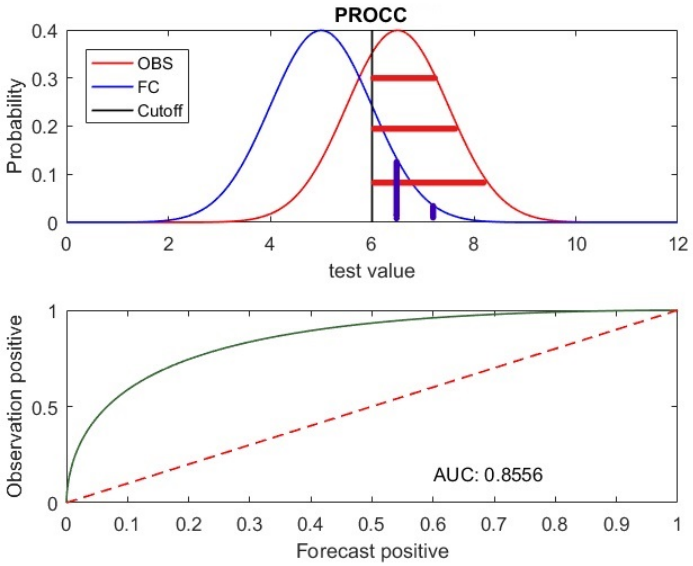


Figure 12: Perfect receiver operating characteristic comparison explanation. In the upper plot two PDF's are shown. The red one is an example for an observation ensemble and the blue one for a forecast. The black line is the "cutoff" and symbolizes a threshold. The red striped area is the "observation positive" and the blue one represents the "forecast positive". The plot below pictures the ROC curve (green). It is calculated out of the "observation positive" (y-axis) and the "forecast positive" (x-axis). The area under the (ROC) curve (AUC) is optimized when the red dashed line is the same as the ROC curve.

Those solutions are shown in the lower plot of Figure 12. The "forecast positive" is pictured at the x-axis and the "forecast positive" at the y-axis (Fawcett, 2006). Together they are adding up to the ROC curve (see the green line). The essential parameter for the verification is the Area under the (ROC)

cure (AUC). In this example $AUC \approx 0.86$.
The perfect result is achieved when $AUC = 0.50$ (see red dashed line). The procedure of getting the ROC curve and the AUC allows to get this result twice. To by-pass this problem the absolute difference between the actual AUC and the perfect $AUC_{Perf.} = 0.50$ is calculated (see Equation 3.7). It is named Area between perfect (ABP).

$$d_{ABP} = |AUC - AUC_{Perf.}|. \tag{3.7}$$

where:

AUC = Area under the curve
$AUC_{Perf.}$ = Perfect result for the area under the curve ($AUC_{Perf.} = 0.50$)

The so found measurement is bounded between 0 (best) and 0.5 (worst). The AUC discriminates well between good and bad PDF locations, but not for good against very good and bad against worse (Marzban, 2004). The ROC curve is a fundamental tool for medical decision making and has been used more often in machine learning and data mining research. The here used approach of calculating the absolute difference between a hypothetical value and the actual AUC is completely new and experimental. Additional information about the under- or over forecasting properties could be gathered by calculating the ABP as a "normal" difference. Then the sign is an indicator, if the forecast is two high or low. This adjustment is even more experimental and is not discussed here.

3.1.8 Kullback-Leibler Divergence (KLD)

$$D(\mathbf{O}||\mathbf{P}) = d_{KLD} = \sum_{i=1}^{n} \mathbf{O}(i) \log\left(\frac{\mathbf{O}(i)}{\mathbf{P}(i)}\right). \tag{3.8}$$

where:

n = number of entries in the probability vector
i = specific vector entire
$\mathbf{P}$ = Forecast probability vector
$\mathbf{O}$ = Observation probability vector

The Kullback-Leibler divergence is a non-symmetric distance measure for the *accuracy* and uses the difference between two probability vectors (Thorarinsdottir et al., 2013). A probability vector adds up to one and has no negative entries. It gives the probability that a specific value occurs. The log is applied as natural logarithm. Equation 3.8 can be transformed into a symmetric form (see Equation 3.10).

$$D(\mathbf{O}||\mathbf{P}) \neq D(\mathbf{P}||\mathbf{O}), \tag{3.9}$$

$$D_2(\mathbf{O}||\mathbf{P}) = D_2(\mathbf{P}||\mathbf{O}) = D(\mathbf{O}||\mathbf{P}) + D(\mathbf{P}||\mathbf{O}). \tag{3.10}$$

For the verification Equation 3.8 is used. The best result is zero and again a bigger value indicates less agreement between the forecast and the observation. The KLD can also be seen as source for the expected discrimination information (Kullback, 1978). This is not discussed in this thesis.

3.1.9 Distance between histograms

Histograms with the concept from Section 2 are produced. However, one exception is made. The binwidth is not calculated with Equation 2.11, the point of interest is the bin amount. Therefore, the binwidth is different in every histogram, but the bin amount, with ten, is constant. This is necessary since that approach calculates the distance between corresponding bins. The bin amount has to be equal. The entries inside each bin are divided by the number of all entries together, the result is a probability (vector) for each bin (Cha, 2007). One example is shown in Equation 3.11.

$$\mathbf{P_3} = \frac{H_3}{n}. \tag{3.11}$$

where:

$\mathbf{P_3}$ = Probability vector for the forecast bin three
H_3 = Entries in bin three
n = Sum of all entries

Those values are used to calculated different distance measurements between the bins. In this thesis Equation 3.12,3.13 and 3.14 are applied for getting information about the *accuracy*. In Cha (2007) 45 different measurements are shown (intersection etc.). The difference of the probability vectors here and for those who are used for the KLD is that here the vectors have much less entries, namely ten.

$$d_{Euc} = \sqrt{\sum_{i=1}^{d} |\mathbf{P_i} - \mathbf{O_i}|^2}. \tag{3.12}$$

Equation 3.12 is known as Euclidean distance. It is one of the most common distance measurements and describes that the shortest distance between two points is a line. It was derived by Euclid, a Greek mathematician who lived around 300 before the Christian era. The square penalizes bigger differences more. The perfect agreement is shown by a result equal to zero. If the score increases a growing differential between the forecast and the observation is suggested. The similarity in the IQD (3.1.6) calculation allows to compare the analogy in the result of both.

$$d_{Sg} = \frac{\sum_{i=1}^{d} |\mathbf{P_i} - \mathbf{O_i}|}{\sum_{i=1}^{d} max(\mathbf{P_i}, \mathbf{O_i})}. \tag{3.13}$$

The Soergel or Bray-Curtis distance (see Equation 3.13) is often used in the ecology. This score is bounded between zero and one, where zero is the best and one the worst result.

$$d_{Lor} = \sum_{i=1}^{d} ln(1 + |\mathbf{P_i} - \mathbf{O_i}|). \tag{3.14}$$

where:

n = number of bins (probability vector entries)
i = specific bin
$\mathbf{P}$ = Forecast bin (probability vector)
$\mathbf{O}$ = Observation bin probability vector

Equation 3.14 is called Lorentzian. Again log is applied as natural logarithm. Here, one is added to guarantee the non-negativity, which is needed due to the logarithm property. Zero is the perfect score and again grows when the distance is getting bigger. This score deviates quite good.

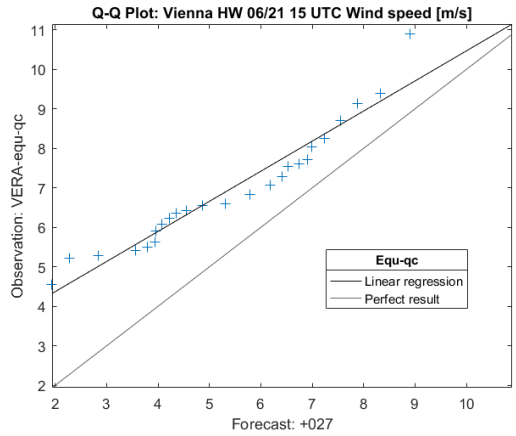
3.2 MesoVICT verification

The verification is separated into two parts. The first one is an example verification for the wind speed in Vienna at a certain time (see Section 3.2.1). The next one shows how the values evolve in time and at different grid points (see Section 3.2.2). Additionally, different parameters are used in this part. The forecast ensemble is verified, with one exception, against the observational one. This exception is made due to the CRPS, where the ensemble is compared with the VERA reference.

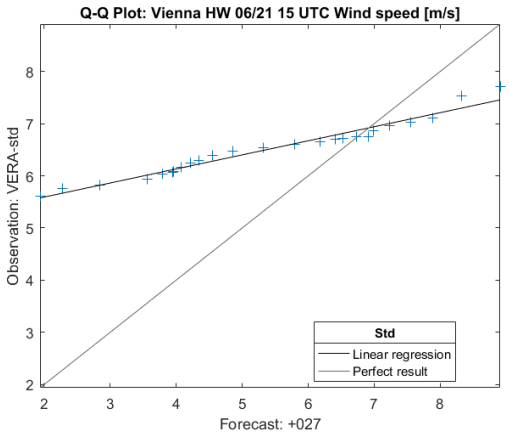
3.2.1 Ensemble-Verification for the wind speed in Vienna

Vienna is in the Eastern Alps and marked with a red dot in Figure 4. The verification time is 06/21 2007 15 UTC. Right around that time a cold front approaches Vienna. The used parameter is the wind speed.
The first verification method is the Q-Q plot (see Section 2.1.1). Here, no assumption about the distribution is needed. Figure 13a is produced by the "equ-qc" observation ensemble set up and a 27 hour forecast. The distribution is shown as 25 equal probability intervals (blue crosses). The perfect agreement is pictured with the gray line. The black one is the linear regression from all 25 probability marks. An "eyeball" verification suggests a close but no good match between those two distributions. The same method is applied for the "std" observation ensemble. The result is represented via Figure 13b. This set up shows a much worse agreement between the forecast and the observation. The narrow spread is the reason for that (see Figure 5b). It does not mean that it could not be a good estimate for the "true" observation uncertainty, but at least the forecast is not distributed like that, which leads to the suggestion of a bad prediction. Since it is an example verification and all methods should be explained in an understandable fashion.
Due to the fact that, the "std" set up has less agreement all further verifications are shown for the "equ-qc" one. Later an overview over the "std" results is provided (see Figure 17b).
A related comparison method to the Q-Q plot, is the box plot (see Section 3.1.5). As it is also an "eyeball" verification no examples are shown, because quantitative measurements are needed.

One of those objective measurements is the integrated quadratic distance (IQD). In Figure 14a the IQD for the "equ-qc" observation ensemble is calculated (blue line against the yellow) and compared with the CRPS (blue line against the red one). The CRPS compares the forecast ensemble with the VERA reference. It could be considered as a prestage IQD without an observation ensemble. The different results could be an indicator of how an observation ensemble changes the verification result. In this example the d_{IQD} is nearly half as big as the $CRPS$. Remember a value close to zero is worthwhile.
Since CDF's are used for the calculation, an assumption about the distribution is made. To avoid this matter the kernel density estimate (see Section 2.3) is used to fit the CDF's. The shown results are different depending on which score is investigated (see Figure 14b). The CRPS got increased by this approach. On the other hand, the IQD is smaller than without the kernel density estimate. As a result the difference between the CRPS and the IQD increases above 50%.

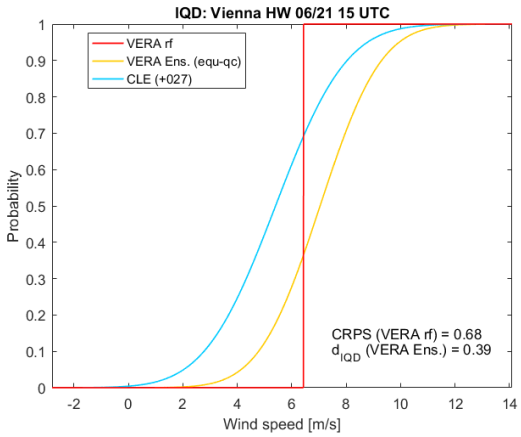


(a) Q-Q plot for Vienna at 06/21 2007 15 UCT the parameter is the wind speed. The observation ensemble is "equ-qc".

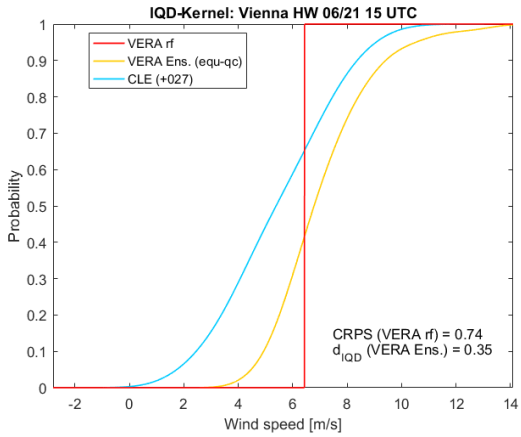


(b) Q-Q plot calulated for the same grid point and time, but with the "std" observation ensemble.

Figure 13: Both plots shows the comparison between an 27 hour ensemble forecast against an observation ensemble. The blue crosses represents the probability intervals. The gray line indicates the perfect agreement and the black one the linear regression calculated with the intervals.



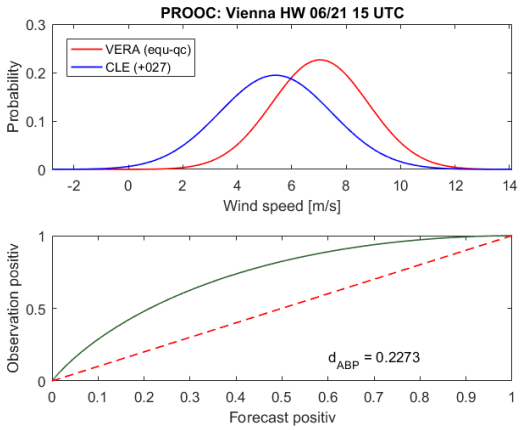
(a) IQD and CRPS represented for Vienna and the wind speed at 06/21 2007 15 UCT.



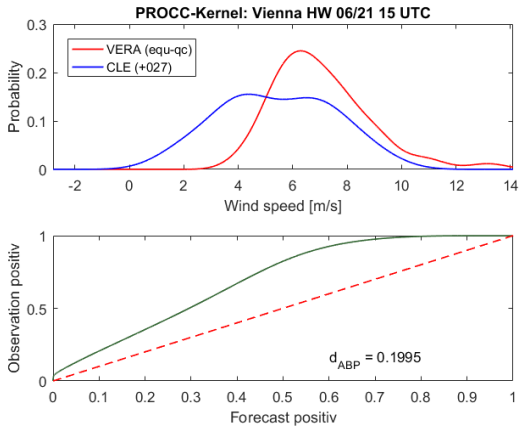
(b) Here, the kernel density estimate is valid. The CDF's are plotted without an assumption about the distribution.

Figure 14: The CRPS is calculated with the red line, which indicates the VERA reference, against the blue one. This is the forecast ensemble. The IQD uses the blue line and compares it with the corresponding yellow CDF ("equ-qc" observation). The right plot shows the same approach, but with the kernel density estimate.

Another quantitative method is the PROCC. Here, PDF's are a part of the measurement (see Figure 15). In Figure 15a the above plot shows two PDF's. The blue one is the PDF calculated with parameters collected from the 27 hour forecast ensemble. The red curve is from the "equ-qc" observation ensemble. The Area between perfect (ABP) is shown on the left below plot. It is the area between the green ROC curve and the red dashed line, which indicates the perfect result.
As it is in this case with the CDF's also with the PDF's assumptions about the distributions are made. In order to avoid that the kernel density is also used for the probability density functions. The result is depicted in Figure 15b. The comparison between the sores indicates that the kernel density approach makes the score smaller. It decreases from $d_{ABP} \approx 0.23$ to $d_{ABP} \approx 0.20$. The same behavior was observed at the IDQ.



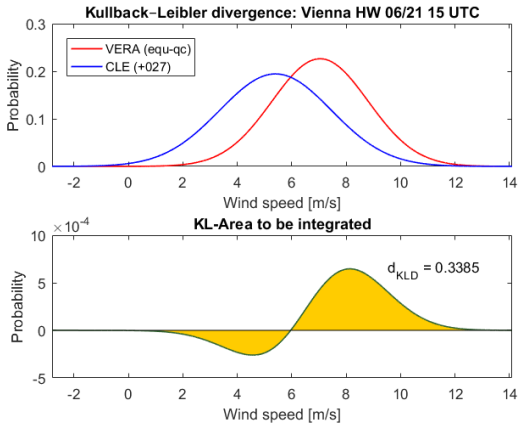
(a) The PROCC fo the grid point Vienna. Above two PDF's are shown. Below the ROC curve with the area between perfect is pictured.



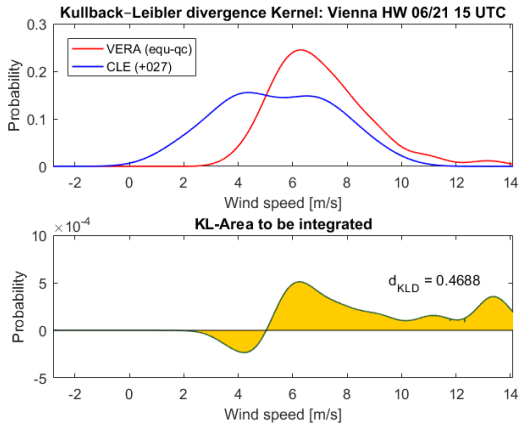
(b) The same as on the plots on the left side are shown, except the kernel density estimate is used.

Figure 15: This example verification is done for the grid point Vienna. The time is 06/21 15 UTC. The plot on the left side above shows an "equ-qc" observation (red line) and a forecast (blue line) PDF. The plot underneath displays the d_{ABP}, where the green curve is the ROC curve and the red dashed one is the perfect result. The area between those two is the ABP. On the right side the same is shown, but for the kernel density estimate.

The Kullback-Leibler divergence uses PDF's and splits them in probability vectors. The result is shown in Figure 16. The area which is important for the KLD is marked yellow. Again the ensemble forecast and the "equ-qc" observation are used. On the left side of the plot normal PDF's are used. On the right side the kernel density estimate is used. This time, in comparison with the IQD and the PROCC, the KDE increases the score from $d_{KLD} \approx 0.34$ to $d_{KLD} \approx 0.47$.



(a) The KLD for Vienna at 06/21 15 UTC.

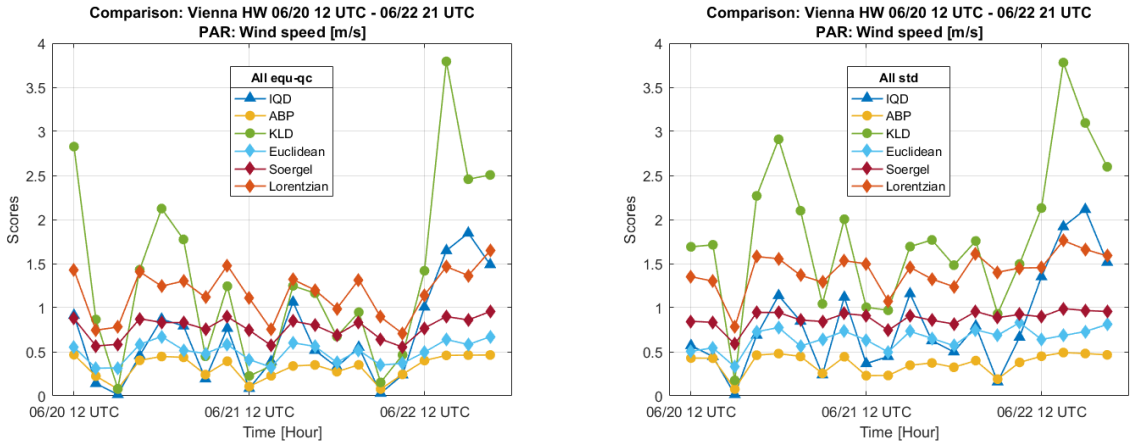


(b) The KLD for the same point and time but with the kernel density approach.

Figure 16: The KLD uses probability vectors to calculate the score (below plots yellow marked). Again the forecast is blue and the observation with the "equ-qc" set up is red. On the right side the kernel density estimate is applied.

The histogram scores are not explained here, but they are calculated as well and part of the time evaluation.
All together we have six different scores to compare the forecast with the observation ensemble. All

of them have zero as their perfect score. So far, we just evaluated one time step, but now we are comparing the scores for the whole verification period (see Figure 17a). Here the values for the "equ-qc" set up are pictured. It is easy to see that most of the time all the scores behave in the same way. This means that they are increasing and decreasing at the same time. The exceptions are more detailed discussed in Section 3.3.
In Figure 17b the "std" observation is used. All results are significantly higher. The reason for that is the more narrow observation ensemble spread. This makes the CLE forecast look worse in comparison with the "equ-qc" set up.



(a) The ensemble verification done for the grod point Vienna is pictured. Here the "equ-qc" set up is valid.

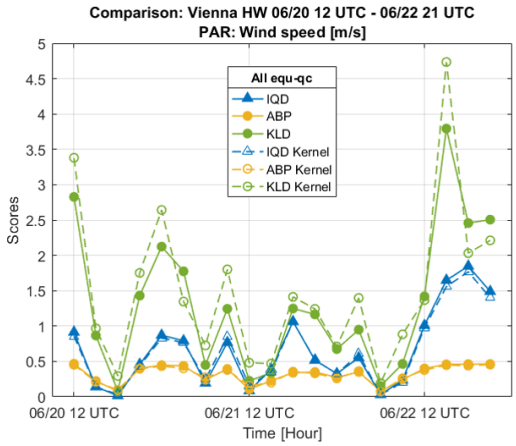
(b) Again all scores are represented but for the "std" observation ensemble. The forecast is the CLE.

Figure 17: All six ensemble verification methods are represented over the whole time period. The "equ-qc" set up is shown left and the "std" left. The parameter is the wind speed.

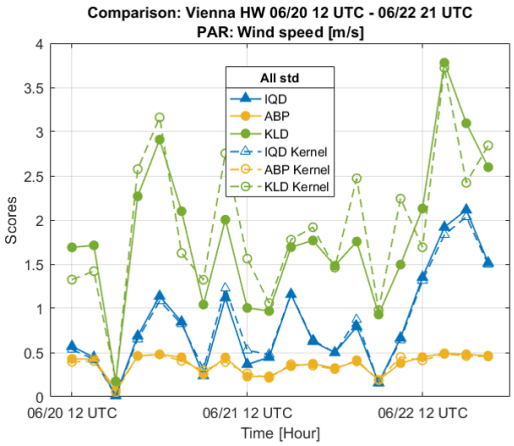
The IQD, PROCC (ABP) and KLD in Figure 17 using the Gaussian assumption. The conclusion in Section 2 proposes to apply the kernel density estimate (KDE). The results are shown in Figure 18. The left plot in Figure 18a represents the "equ-qc" observation ensemble. The biggest deviation between the kernel and not kernel values are seen for the KLD. Some deviation occurs also in the two others. The finding is that, the kernel density estimates effect the results in both directions. They could be more or less due to that approach. The same summery can be provided for the "std" set up (see Figure 18b). The KDE can in- as well as decrease the verification scores. Note the different axis scaling. Overall, the evaluation with the "std" observation ensemble provides worse value for the CLE.

3.2.2 Ensemble-Verification for multiple parameters over time at different grid points

The previous example verification was carried out with the wind speed. Now the potential temperature, the pressure (model and standard formula reduction) and the precipitation for the grid point Vienna are under investigation. To make an overview more easy to capture, only the IQD is shown (see Figure 19). This is proper since all verification methods discussed in Section 3.2.1 are evolving simultaneously.
The biggest standout is the precipitation. As long as no rain occurs the verification score is very good, which is not surprising, but whenever precipitation appears the amount is not detected that well. The second noticeable parameter is the potential temperature (theta). A first look at the verification results suggests that, there are bigger disagreements in the ensemble spread. Nevertheless, a closer look at the distribution declares that the high verification results more frequently came from a bad



(a) IQD, ABP and KLD time evaluation for Vienna and the wind speed. Kernel density estimation used.



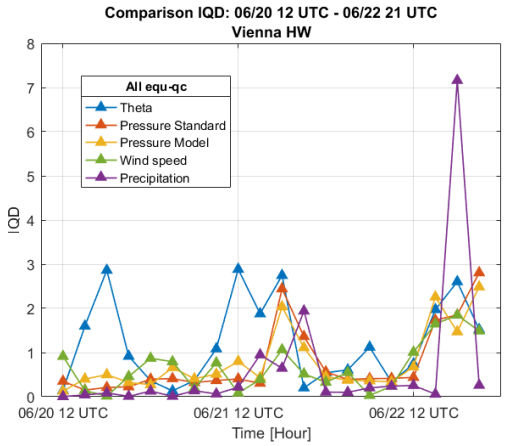
(b) Kernel and non kernel comaprison for Vienna over the verification period for the wind speed. "Std" observation applied.

Figure 18: The verification done without (continuous line) and with (dashed line) the kernel density estimate. The left plot shows the "equ-qc" and the right the "std" observation ensemble. Note the different scales for the plots.

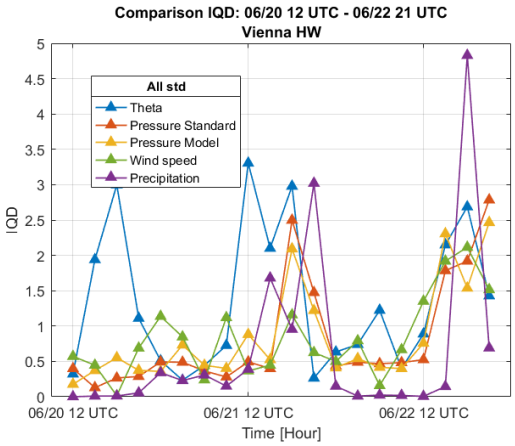
location between the observation and the forecast. In simple terms, the CLE is different from the observation and produced a bad forecast. The two different CLE pressure outputs are consistent, but in case of a big pressure gradient (frontal approach) the forecast ensemble is more likely to be wide (spread $\approx 4-6$ hPa). The pressure observation, on the other hand, is very good most of the time, which leads to a small spread ≈ 1 hPa and a bad verification outcome. This could also be an error source for the potential temperature. An explicit analysis for the temperature did not show that behavior all the time, but sometimes. Anyway, more frequently the bad location is the reason for high results. The already discussed wind speed has comparable ensemble spread and often a good location and is the "best" parameter for testing verification methods.

The bad results at the end of the comparison period could be explained with the longer lead time and the drop of predictability. For these conclusions, the "equ-qc" and "std" observation ensembles are used. The "std" set up offers poorer results, which was already observed previously with the wind speed. By comparing those two plots one has to keep in mind, that the y-axis are different scaled. The kernel approach was made, but is not shown, as no additional information was found.

Since we just discussed one grid point (Vienna) over time, a discussion about the behavior over space is also interesting. Therefor, the wind speed is the selected parameter and a time around the frontal passage is used, 06/21 12 UTC to be accurate. As measurement the IDQ is applied.
In Figure 20 the whole verification area is pictured. A bluish color suggests a good score. Whenever the points are turning into brighter colors it is worse. The biggest differences for the "equ-qc" set up are observed in the western Alps and Adriatic subsections (not pictured specifically). Here, the Alps and the Dinaric Alps are eye-catching. In the southern German part the foothills of the German Middle Mountains are visible. The deviations in the northwester part of Vienna could be caused by the approaching front. It is more or less the same with the "std" observation (see Figure 20b), despite a little bit more deviations are seen overall, but the prominent parts are equal. This coincides with the already found knowledge about that observation ensemble, because it has a more narrow spread and the scores are in general a bit higher.
This verification shows the already known difficulties in complex terrain, where the subsection Adri-



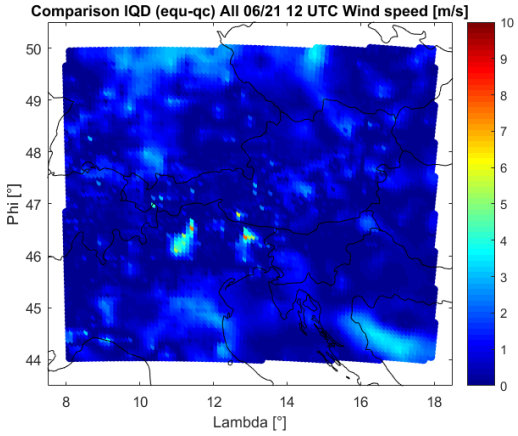
(a) Different parameters evolve over time. The pressure CLE forecast has two outputs. The “standard” uses a well known formula to reduce the pressure too the sea level. It is used for model comparisons. The “model” one, is specific and interesting for the forecaster.



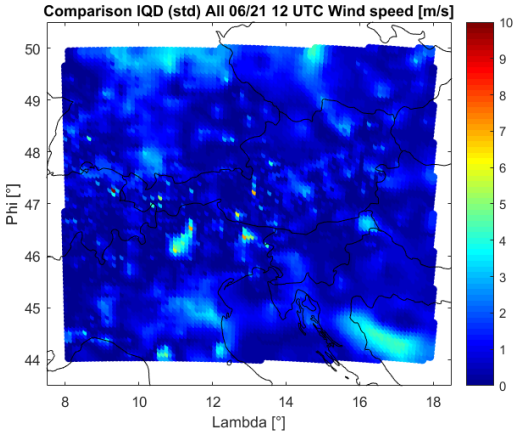
(b) Here, the “std” observation ensemble is compared with the forecast.

Figure 19: Time evaluation for different parameters for the grid point Vienna with the IQD. Potential temperature is also known as potential temperature. The left plot represents the “equ-qc” and the right one the “std” observation ensemble set up. Note the different scales for the plots.

atic and West Alps are the biggest anomalies. A different time step is not shown, but was investigated and the problem areas were similar.



(a) IQD space overview for the wind speed. The observation is the “equ-qc” ensemble.



(b) The IQD comparison like in (a) but for the “std” observation.

Figure 20: The area comparison for the IQD for 06/21 12 UTC and the wind speed. Bluish colors suggest a better agreement between the forecast and the observation ensembles. Brighter ones picture less agreement. The left plot shows the “equ-qc” and the right the “std” observation ensemble.

3.3 Verification discussion

Just to make it clear ones again, MesoVICT is a community project about verification methods inter-comparison and not a model inter-comparison. Despite the fact that by now no spatial verification is available, which is the main topic of MesoVICT, verification methods are investigated and compared.

The aim of the verification part of this thesis is to provide ensemble verification methods. Normally, a specific observation is used to verify an ensemble forecast. In this case, two ensembles are about to be compared. That topic is not discussed heavily so far in the meteorology, only because of the minor availability of observation ensembles. Now, Gorgas and Dorninger (Gorgas and Dorninger, 2012a) provide two of them, which makes an comparison very interesting. Those two are abbreviated with "equ-qc" and "std".

However, the "equ-qc" set up is wider and makes the forecast look better in comparison with the "std" observation ensemble, which is narrower. All the following results and conclusions are made with respect to both ensembles.
For the verification a distribution approach is applied. The ensemble entrances are used to specify parameters for the normal probability function (PDF), cumulative distribution function (CDF) and histograms. With those distance measurements information about the *accuracy* is collected. In order to avoid a restriction of the Gaussian distribution the kernel density estimate is applied for the CDF and PDF methods to by pass the distribution specification. The histograms itself did not use distribution information. Only the binwidth is influenced by the Gaussian assumption, which is fine for the verification, because a certain bin amount is needed. Additionally, nonparametric plots (Box and Q-Q plot) can be applied for the ensembles. They are good for a small number examples, but not feasible in a bigger amount of cases.
The example ensemble verification (see Section 3.2.1) is done for one grid point (Vienna) around the frontal approach (06/21 2007 15 UTC). The parameter is the wind speed. In addition to that, eight other grid points and four more parameters are investigated over the whole time period (see Section 3.2.2).
As a conclusion, the precipitation was unusable for that kind of verification. All the other parameters are feasible. The wind speed is the best one. The forecast and the observation distributions have a comparable range and a good location. The pressure evaluation is questionable due to the pressure gradient in the forecast. Here, a bigger spread in comparison with the observation occurs. A observation ensemble approach for the pressure may not be that useful, since the pressure measurement is very good. Sometimes a similar problem arises for the temperature. Here, it is not that crucial. A bad forecast is more often the source of a bad score.
High values for the scores are observed around the frontal approach (around 06/20 15-18 UTC) and at the end of the case study. Since the CLE is more inertial, due to the three-hourly data, a delay in comparison with the hourly VERA is likely. The errors at the end of the case study are due to increasing uncertainty after some forecast time.

The following methods and Table 5 are represented with input from the wind speed, pressure and the potential temperature. The table is subjective and only an indicator. Even if nine grid points and 54 hours are investigated in detail, a more precise study could expand the general validity.

The integrated quadratic distance (IQD) is an extension to the continuous ranked probability score (CRPS). Both are using the CDF to gather information. The CPRS gives the result without an observation ensemble and the IQD with an observation ensemble. A comparison between those two could reveal the influence of the observation uncertainty on that score. Here, it would be around 50%. A

more general method is introduced in Section 4.1. The IQD is more sensitive to the location than to the shape. The quadric structure penalizes bigger values more. The CDF smooths due to the integration of existing deviations. This results in less variation in comparison with the Kulllback-Leibler divergence (KLD) or the perfect receiver operating characteristic comparison (PROCC) when the kernel density estimation is applied.
The PROCC uses the PDF and is very sensitive for well matched PDF'S (location and shape). The representative score d_{ABP} is bounded. This results in a discrimination problem for bad forecasts, because the worst score is achieved at a specific point and whenever the forecast is getting even worse no effect on the score can be seen. The kernel density estimate seems to have not that much effect on the method, but in perspective to the less overall ABP value it is up to 20%.
The KLD is sensitive for location and shape and has a good discrimination. It uses probability vectors which are comparable with infinitive small PDF parts. The sum of the vectors is equal to one. Due to the sensitivity the kernel density estimate has a big effect on that method. The deviation is around 30%. The non-symmetric arithmetic of the KLD makes it important how the score is applied.
The kernel density estimate overall helps to minimize the distribution error. It decreases as well as increases the score, but it is not an universal remedy. It is sensitive to the bandwidth and also the kernels can have a different shape.
Three more measurements are found with histograms. Here, ten bins are assumed. This is the point of emphasis. A different amount of bins influences the score in a big fashion. This amount of bins is chosen to make a verification possible, it does not mean that this is the perfect or only approach. A future investigations would be useful. Especially with the background that in Cha's paper (Cha, 2007) 45 different approaches are provided. The good thing about a histogram is, that no assumption about the distribution is necessary. The IQD and the Euclidean histogram measurement have the same quadratic structure, where bigger differences are penalized to a higher degree. They evolve most of the time in the same way, like all the scores do, but deviations occur. This is particularly the case for narrow distributions. Here, one bin is dominating and the bin amount should be changed.

Again, overall all the scores evolve most of the time in the same way for the wind speed, the potential temperature and the pressure (both). Also the experimental PROCC approach is similar. Differences are seen with the histogram due to the bin number. Especially for data with a small ensemble spread the scores are not that diverse. For the precipitation a useful addition could be the Brier score (BS) and the (adjusted) reliability diagram. Here, thresholds are applied and the verification of a specific rain event could be done. However, the best way to verify field measurements are spatial verification methods. This is the main part of the MesoVICT project, but so far no useful methods are discovered.

Table 5: Method classification for the wind speed, potential temperature and pressure. This values are subjective and should be used only as lead

Scores/Interp.	Perfect	Good	OK	Bad	Worst
IQD	0	<0-1	<1-2.5	<2.5-4	<4
ABP	0	<0-0.15	<0.15-0.35	<0.35-0.45	<0.45-0.5
KLD	0	<0-2	<2-4	<4-8	<8
Euclidean	0	<0-0.50	<0.50-0.75	<0.75-1	<1
Soergel	0	<0-0.6	<0.6-0.8	<0.8-0.95	<0.95-1
Lorentzian	0	<0-1	<1-1.5	<1.5-1.8	<1.8

Table 5 suggests that the forecast is less good if the "std" observation is used for the verification. It is may not be proper to refer the results to an observation product but since not enough knowledge

about observation uncertainty is available, it is done. No matter which VERA ensemble is used, problems in the complex terrain are observed. This indicates a problem in the forecast ensemble, since the VERA uses, except for the wind speed and precipitation, specific fingerprints in this regions. Even for parameters without those fingerprints the analysis is more convincing than the forecast.

4. Uncertainty estimates

No matter what forecast or observation is used the verification result is one value without any uncertainty estimates or error bars. Sure, whenever an ensemble approach is valid there is already an uncertainty in the input data; however, it is not shown in the score itself. Section 4.1 provides information (error bars) about that issue.
It is difficult to define a proper observation ensemble. There are numerous questions which have to be answered. Is it better to have a narrow or a wide ensemble? For which parameter dose it makes sense? How can they be used in the verification process? With exception of the last question (see Section 3) non of those could be answered here in an reasonable manner. What can be delivered, is a strategy to compare different observation ensemble results in Section 4.2.
This comparisons led also to error bars. To see if the *score uncertainty* or a *different use of the observation ensemble* is more prominent a discussion is done in Section 4.3.

4.1 Score specified uncertainty

4.1.1 Absolute error uncertainty

The absolute error (AE) is a verification method were a deterministic forecast is compared to one observation (see Equation 3.2). Here, the forecast is the CLE deterministic (CLE mean) and the observation is the VERA reference (see Figure 21 green line). However, for the uncertainty estimate in the verification results the observation ensemble is used. Therefore, the absolute error evaluation is adjusted. See Equation 4.1 for an example with the "equ-qc" ensemble. The same is done with the "std" observations.

$$\begin{aligned} AE_{1,equ-qc} &= |p - o_{1,equ-qc}|, \\ AE_{2,equ-qc} &= |p - o_{2,equ-qc}|, \\ &\vdots \\ AE_{50,equ-qc} &= |p - o_{50,equ-qc}|. \end{aligned} \tag{4.1}$$

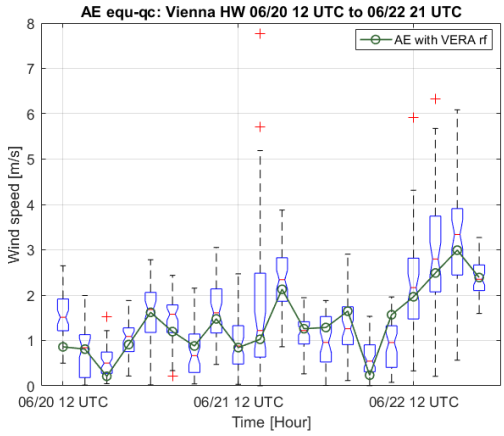
where:

$AE_{xx,equ-qc}$ = Absolute error for one (xx) "equ-qc" ensemble member
p = Deterministic forecast
$o_{xx,equ-qc}$ = Observation xx "equ-qc" ensemble member

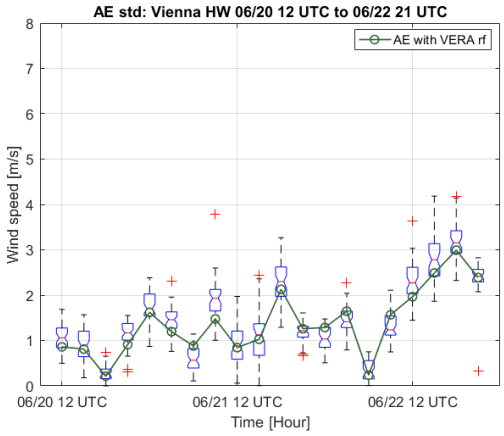
The absolute error is calculated multiple times. Each of the 50 ensemble members are used as one new observation, like o in Equation 3.2. Consequently, we get 50 different absolute errors. The forecast remains the same. This set of new solutions is represented as a box plot with notches (see Section 3.1.5).

In Figure 21 the wind speed is used. The grid point is Vienna and the whole verification period is shown. The VERA reference calculation is pictured with the green line and the circles are located at a particular time. The "equ-qc" set up is represented in Figure 21a and "std" one is in Figure 21b.
The AE shows similar properties as the ensemble verification methods (see Section 3.2.1). During the frontal passage (06/21 15-18 UTC) the score gets worse. The same occurs at the end of the period of

review. It is noticeable that the error bars for the “equ-qc” output are much wider. This correlates with the wider spread for this set up. The “std” ensemble, on the other hand, has a narrower spread and also narrower error bars.



(a) The AE uncertainty shown for the grid point Vienna. The observation is the “equ-qc” wind speed ensemble.



(b) The AE uncertainty shown for the grid point Vienna. The observation is the “std” wind speed ensemble.

Figure 21: The uncertainty estimation for the AE is done for the grid point Vienna over the whole verification period. The parameter is the wind speed. The green line shows the absolute error calculated with the VERA reference. The uncertainty due to the different observation ensembles is pictured with a box plot with notches.

It is also visible that those bars are about to get bigger in the frontal passing and for a long forecast period. Some exception are there, for example, the last time step. Altogether, it is a clear enough evidence that, especially in the areas of higher and therefore worse verification scores, this values are more uncertain.

As a conclusion, the absolute error for the “equ-qc” can be doubled as well as reduced by a factor of two. This is the extent of the IQR in the box plots. The lower and upper whisker as well as outliers are not taken into account, because the IQR is more representative. The notches also show the 95% confidence interval of the median. For the “std” ensemble less deviation is evident. The AE can still be higher and lower by a factor of one third, but definitely not as big as for the “equ-qc” ensemble. An observation ensemble is the absolute necessity in order to get this uncertainty estimates. If those ensembles are available this approach is valid for all verifications where one observation and not an ensemble is used (CRPS, reliability diagram). It is also possible to replace the box plots with confidence intervals.

Here, only the observation ensemble is applied to quantify the uncertainty estimate. But it is also possible to use the forecast and even both (forecast and observation) ensembles to calculate the same box plots as in Figure 21. Therefore, Equation 4.1 is adapted. For the forecast measure the difference is computed with a changing p and a constant o. If both ensembles are used for an uncertainty estimate p and o are changing. In Figure 22 the box plots, due to different possibilities, are shown. For the explanation the absolute error is calculated with the “equ-qc” observation. The grid point is Vienna and the parameter is the wind speed.

It is seen, that the deviations have a much bigger extent where the forecast is included. This is understandable, because the forecast ensemble has a wider spread than the observation ones. Nevertheless, the observation IQR covers around one third of the FC and FC-OBS IQR’s. Which legitimate an

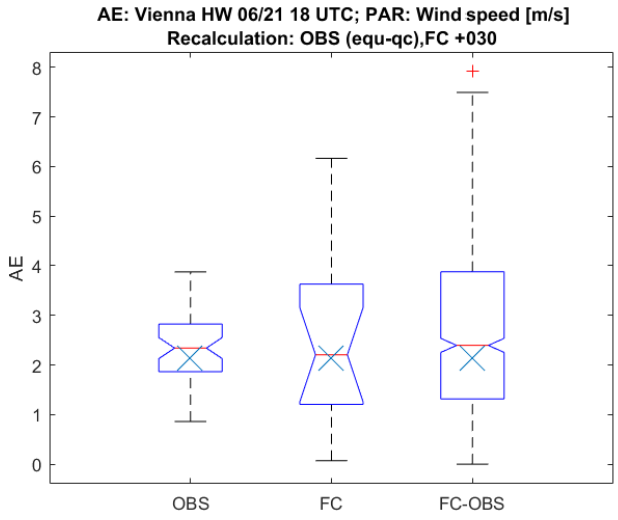


Figure 22: The absolute error is calculated as in Equation 4.1(x-axis OBS). The FC and FC-OBS (both x-axis) uses an adapted method. For the FC box plot the observation remains the same and the forecast is changed and for the FC-OBS one both inputs are variating.The grid point is Vienna and the parameter the wind speed. The time is 06/21 18 UTC and the observation is "equ-qc".

uncertainty estimate for the "equ-qc" observation. The "std" spread and the corresponding box plots are much narrower (no comparison shown). Therefore, an estimate is possible but may not be that useful.

Due to the limited computational power we only use one example, which is not that representative. But at least an overview is provided.

4.1.2 Integrated quadratic distance uncertainty

The integrated quadratic distance (IQD) uses two ensembles and compares them against each other. To get the same error bars (or confidence intervals) as in Section 4.1.1 a re-sampling approach is necessary (see next paragraph bootstrapping). Otherwise no deviation without using a different ensemble is achieved. The IQD serves as an example for the ensemble forecast verification. This approach is valid for all of the methods.

The bootstrap is a re-sampling *with replacement* method (Wilks, 2011, pp. 172). The idea is also known as *plug-in principle*. A distribution rearranges itself by replacing original values with others from the same distribution. The easiest way to explain it is to use an example. Let us say the sample is $\{4,5,1,3,2,6\}$. Now the re-sampling is applied. The new sample is randomly ordered with $\{1,4,5,6,5,3\}$. Some of the values are drawn multiple times (here 5), others are not drawn at all (here 2). This is possible because each entry is available everytime a new member of the new sample got drawn. Let us think about a hat with numbered paper stripes in it. After a number is drawn, it is put back in the hat where it can be drawn again. This is done numerous times until a new sample with the same size n is gathered.
This approach can be repeated, for example $n_B = 100$ times. The result is a data set with 100 samples of the same size n. The new data set is used to estimate the precision of a sample or distribution (deviation from the mean, variance, ect.).

The sample size for the observation ensemble is $n = 50$. Since it is computational wise expensive to apply the bootstrapping, a comparison was made for this thesis. In the literature $n_B = 1000$ is suggested (Ferro, 2007). The limitation of computational power makes $n_B = 1000$ impossible to perform. Here, $n_B = 50$ is used. This circumstance limits the approach to a guide for how it has to be done, but with less reputation and therefore, less representativeness.

The IQD is calculated with the same idea like Equation 4.1. With each bootstrap information for one new observation CDF ($O_{n_B xx}(x)$), which is applied to get a new IQD (see Equation 4.2), is provided.

$$\begin{aligned} d_{IDQ,n_B1,equ-qc} &= \int_{-\infty}^{\infty}[P(x) - O_{n_B1,,equ-qc}(x)]^2 dx, \\ d_{IDQ,n_B2,equ-qc} &= \int_{-\infty}^{\infty}[P(x) - O_{n_B2,,equ-qc}(x)]^2 dx, \\ &\vdots \\ d_{IDQ,n_B50,equ-qc} &= \int_{-\infty}^{\infty}[P(x) - O_{n_B50,,equ-qc}(x)]^2 dx. \end{aligned} \tag{4.2}$$

where:

$d_{IQD,n_B xx,equ-qc}$ = Distance measurement for one bootstrapped “equ-qc” observation
$P(x)$ = CDF for the forecast
$O_{n_B xx,equ-qc}(x)$ = CDF for one bootstrapped “equ-qc” observation

This is done with every bootstrapped sample $n_B = 50$. The so found 50 new IQD’s are shown as a box plot. Both observation ensembles are computed separately. In Equation 4.2 the “equ-qc” is pictured.

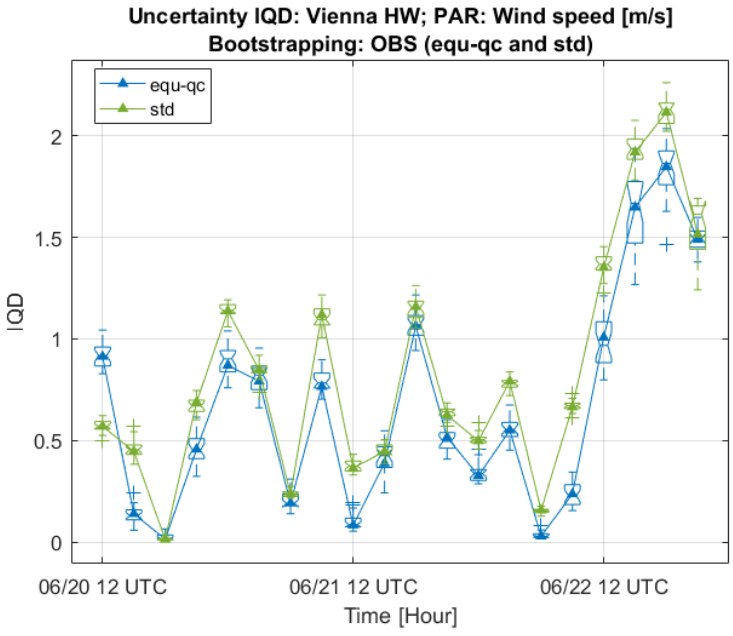


Figure 23: The integrated quadratic distance is estimated multiple times to provide an uncertainty estimate. Therefore, the observation ensemble is bootstrapped 50 time. The estimate is shown as a box plot. The “equ-qc” ensemble is the blue curve and the “std” the green one. The parameter is the wind speed for Vienna.

The result is plotted in Figure 23. The IQD is shown for the wind speed through the whole verification period. The “equ-qc” ensemble is the blue line and the “std” one is green.
Altogether, the score deviation is around 10% and can influence the score in a positive and negative way. This is much less than the uncertainty in the previous section. One reason could be the less

amount of bootstraps, because more repetitions could provide more deviation. However, as long as the ensembles have a narrow spread the bootstrapping can not stretch it too much. This is seen for the “std” set up. The “equ-qc” is wider, but also not that wide as seen in Figure 21. Note that the ensemble approach itself is an estimate for the uncertainty. Whenever the IQD is getting worse the error bars are also becoming a bit wider, again not that much, but this property is comparable with the absolute error uncertainty. A verification discussion is done in Section 3.3.

As in Section 4.1.1 only the observation for the uncertainty classification is changed. To get knowledge about the magnitude of the observation estimate in comparison with the forecast and forecast-observation one, the same procedure is done with those ensembles. Equation 4.2 is adjusted to keep the observation CDF constant and modify the forecast CDF in the same way as done with the observation previously ($n_B = 50$). The result is shown in Figure 24. One time step is chosen and the “equ-qc” is displayed.

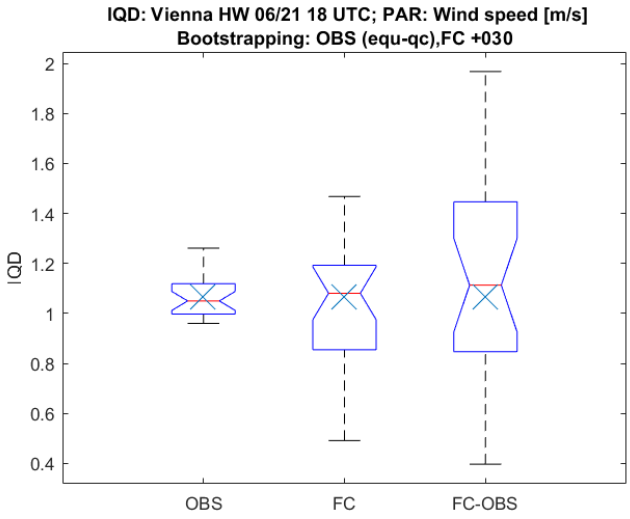


Figure 24: The resampling approach bootstrapping is used for the uncertainty estimation in this plot. The deviation due to a changing observation ensemble is shown as the left plot (OBS). The middle one is due to the calculation with the forecast (FC) and the right one is calculated with both.(FC-OBS). The grid point is Vienna and the parameter the wind speed. As a time 06/21 18 UTC is valid and the observation is “equ-qc”.50 bootstraps are performed.

The forecast is plotted as the middle box plot. The right plot is done with a deviation of the forecast and observation CDFs. This IQR extent is the most prominent. The FC and the OBS IQC are less, where the OBS is one quarter of the FC-OBS one. In the previous section it was one third. For the “equ-qc” set up an uncertainty estimate based on the observation is reasonable, but the deviation due to the forecast is higher.
The “std” set up (not shown) would be to narrow in comparison with the forecast and the bootstrap approach is not that necessary. Note, that the FC-OBS plots are also effected by a smaller observation ensemble.
However, this is just a random sample and the deviation at another spot or time could be different. Since the overall properties from the observation and forecast ensemble (see Figure 5) are not changing over time, the random sample is representative for this grid point.

4.2 Uncertainty estimates between different observations

The previous discussed uncertainty estimates (see Section 4.1.2) are due to the observed ensemble itself. Since especially the observation ensembles are experimental a comparison between them is necessary.
Here, two observation ensembles are available ("equ-qc" and "std"). With each of them a score and their uncertainty is calculated, for this example the absolute error (AE) and the integrated quadratic distance (IQD) are used.

The AE uncertainty is computed like in Equation 4.1. One for the "equ-qc" set up ($AE_{xx,equ-qc}$) and one for the "std" ($AE_{xx,std}$). All AE are now used to calculate the difference between every $AE_{xx,equ-qc}$ and every $AE_{xx,std}$. Mathematically it is called "outer product difference" (see Equation 4.3).

$$\Delta_{AE} = \begin{bmatrix} AE_{1,equ-qc} - AE_{1,std} & \dots & AE_{1,equ-qc} - AE_{50,std} \\ AE_{2,equ-qc} - AE_{1,std} & \dots & AE_{2,equ-qc} - AE_{50,std} \\ \vdots & \ddots & \dots \\ AE_{50,equ-qc} - AE_{1,std} & \dots & AE_{50,equ-qc} - AE_{50,std} \end{bmatrix}. \tag{4.3}$$

where:

Δ_{AE} = AE deviation due to the different observation ensemble
$AE_{xx,equ-qc}$ = AE calculated with the "equ-qc" ensemble set up, xx is the ensemble number
$AE_{xx,std}$ = AE calculated with the "std" ensemble set up, xx is the ensemble number

Since it is a 50×50 matrix Δ_{AE} has 2500 entries, which are used to show the deviation via box plot. Again notches are included in order to see the 95% confidence interval from the median. The uncertainty estimate is calculated for Vienna and over the whole verification period. The parameter is the wind speed. The result is seen in Figure 25.

The dark green line for the AE, calculated with the VERA reference, and the blue ("equ-qc") and green ("std") box plots for the different observation ensembles are a combination of the elements in Figure 21. On the plot below in Figure 25 the deviation or Δ_{AE} is shown.
The magnitude and values of the box plots are more important than the sign. The sign works as an indicator and shows which observation ensemble has a higher score. The value is effected by the space in between the two different results collected by both observation ensembles and the magnitude is sensitive to the box plot extent. The space between and the magnitude interact with each other. That means the space between the scores can be big, resulting in a higher Δ_{AE}, but whenever the box plot extents are big as well, spread from the Δ_{AE} is more prominent.
Here, a negative value means that the "std" set up is higher overall. We have to be careful in consideration of a closer discussion. Because if we take the Δ_{AE} median as the *uncertainty estimate due to different observation ensembles* it would be around 10%. However, the deviation, shown with the box plot and here representative with the IQR, suggests an uncertainty around 30%. A solution could be to name the IQR as deviation. Therefore, no specific value has to be classified but an uncertainty interval is delivered. It can be named percentage or numerical wise.

The AE is calculated with one observation and forecast. The ensembles are used for the error bars (box plot) in the results. The ensemble verification methods have also uncertainty estimates, but

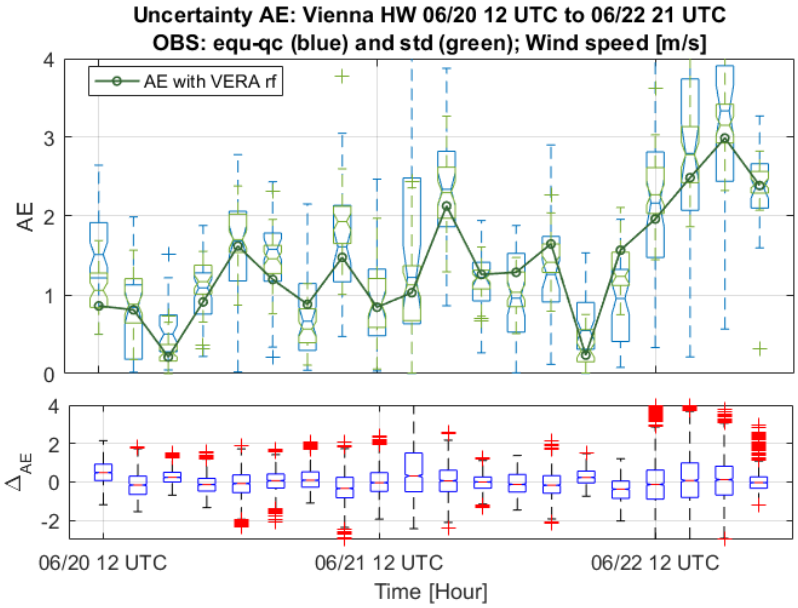


Figure 25: The AE uncertainty due to different observation ensembles is discussed for the grid point Vienna over the whole verification period. The parameter is the wind speed. The above plot shows the AE calculated with the VERA reference with the dark green color. The blue box plots show the AE for the "equ-qc" observation and the green one for the "std" one. Below the deviation or Δ_{AE} are shown. Again a box plots with notches are used.

they are computed via bootstrapping. The *uncertainty due to different observation ensembles* is done like in Equation 4.3, but with the IQD (see Equation 4.4).

$$\Delta_{IQD} = \begin{bmatrix} IQD_{1,equ-qc} - IQD_{1,std} & \dots & IQD_{1,equ-qc} - IQD_{50,std} \\ IQD_{2,equ-qc} - IQD_{1,std} & \dots & IQD_{2,equ-qc} - IQD_{50,std} \\ \vdots & \ddots & \dots \\ IQD_{50,equ-qc} - IQD_{1,std} & \dots & IQD_{50,equ-qc} - IQD_{50,std} \end{bmatrix}. \tag{4.4}$$

where:

Δ_{IQD} = IQD deviation due to the different observation ensemble
$IQD_{xx,equ-qc}$ = IQD calculated with the "equ-qc" ensemble set up, xx is the bootstrapping reputation
$IQD_{xx,std}$ = IQD calculated with the "std" ensemble set up, xx is the bootstrapping reputation

As a conclusion the result is Δ_{IQD} and computed with different bootstrapping reputations (xx). This result is pictured in Figure 26. The "equ-qc" observation ensemble is blue and the "std" one is green. Below Δ_{IQD} is pictured. The grid point is Vienna and the parameter is the wind speed.

We have to keep in mind that two properties (extent and location) can effect the *uncertainty between different observations*. Since the box plots did not have that much extent the value of each observation ensemble (location) is more important. Again, a negative value means that, overall, the "std" set up is higher.
The deviation shown with the median of Δ_{IQD} is more representative than the previous Δ_{AE} estimate. This is sometimes caused by the smaller box plots and less deviation. Overall, the uncertainty is around ±50%. If the IQR is included the deviation is not changed that much. However, to make a

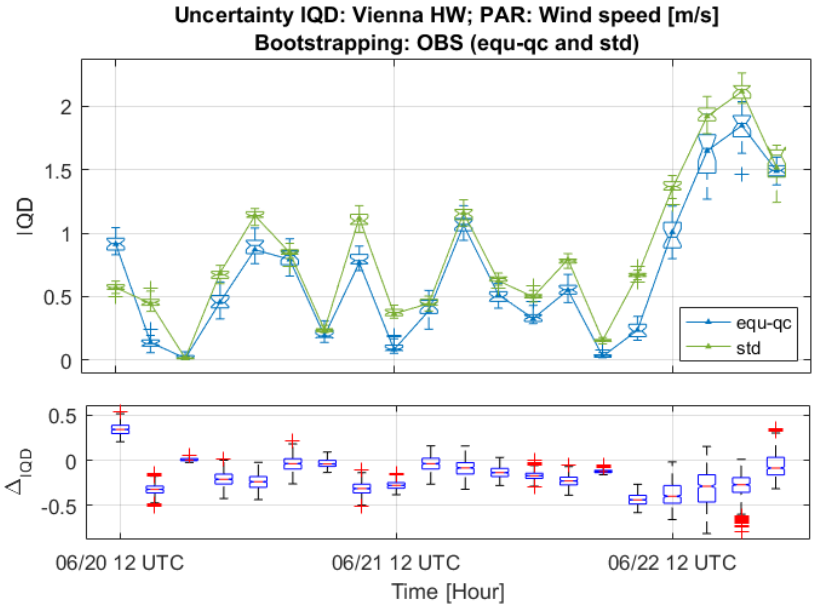


Figure 26: The IQD is used to show the uncertainty due to different observation ensembles for ensemble verification methods done with the wind speed for the grid point Vienna over the whole verification period. In the above plot the "std" observation is pictured green and the "equ-qc" one blue. The below plot shows Δ_{IQD}.

consistent *uncertainty estimate due to different observation ensembles* in comparison with the Δ_{AE}, the deviation is named as an interval and percentage or numerical wise.

4.3 Comparison between score uncertainty and uncertainty due to different observation

There are different sources for uncertainty in an observation and further in the verification result. Some of them could be due to the interpolation method, the grid resolution or maybe the observation density. However, Gorgas and Dorninger (2012b) came to the conclusion that the observation density and grid resolution is more important than the interpolation method.
To make these error sources countable not one observation, but an ensemble is produced. In this thesis two of them are available (see Section 1.3.1, VERA "equ-qc" and "std" observation ensemble). Both are used to calculate an uncertainty estimate.

Here, two different approaches depending on the verification purpose (see Section 4.1) are investigated.
The absolute error (AE) uses one observation and compares it with a deterministic forecast. To get some deviation in the verification result an observation ensemble is used by taking every single ensemble member as a new observation.
Another way of verification is to compare two ensembles (forecast and observation) against each other (Integrated quadratic distance). As the observation ensemble is already used for the comparison, a different approach is necessary to quantify the uncertainty. Here, the bootstrapping is applied.

Those are two different methods for getting an uncertainty estimate due to one observation ensemble. It is also interesting to know what is the *uncertainty corresponding to two different observations* (see Section 4.2). This calculation is done with the "outer product difference".

Now we do have two different uncertainty estimates computed with two observation ensembles, but which one is more important?

The AE with the "equ-qc" observation ensemble is doubled or reduced by the factor of two in comparison with the $AE_{VERA-ref}$. The "std" one is much less, around one third. Within the comparison of different observation ensembles against each other the uncertainty is similar to the "equ-qc" one. $AE_{VERA-ref}$ can be doubled or reduced by a factor of two.
To see how good those overall estimates are, an example with a numerical output is done. The Δ_{AE} at 06/21 15 UTC in Vienna and the wind speed, for the comparison between two ensembles, is from −0.2 to +1.5. The same is done with the uncertainty estimates due to one ensemble itself. The $AE_{xx,equ-qc}$ is between −0.3 and +1.3 and the $AE_{xx,std}$ is from −0.3 to 0.2. All those three results are a possible uncertainty estimate for the $AE_{VERA-ref} = 1.1$ (see Table 6).

Table 6: Uncertainty estimates for the absolute error. The time is 06/21 15 UTC for Vienna and the parameter is the wind speed.

Ensemble [V]	VERA "equ-qc"	VERA "std"	VERA comparison (both)
$AE_{VERA-ref} = 1.1$	−0.3 - +1.3	−0.3 - +0.2	−0.2 - +1.5
$AE_{VERA-ref} = 1.1$ (100%)	−27% - +118%	−27% - +18%	−18% - +136%

The IQD has not one reference value like it is the $AE_{VERA-ref}$. Here, two references are defined. IQD_{equ-qc} and, for the other ensemble, IQD_{std}. For every reference value an uncertainty estimate via bootstrapping and one with the "outer product difference" is available. The uncertainty for the IQD_{equ-qc} due to the bootstrapping approach is around 10%. The "outer product difference" suggests a bigger uncertainty with up to ±50%. For the IQD_{std} one, the bootstrapping estimate is a bit smaller than 10%, but not that much. The ensemble comparison is also ±50%.
Here, also an example is shown, but not for 06/21 15 UTC, because this time step has nearly no deviation. Therefore, 06/20 12 UTC for the grid point Vienna is used (see Table 7).

Table 7: Uncertainty estimates for the integrated quadratic distance. The time is 06/21 15 UTC for Vienna and the parameter is the wind speed.

Ensemble [V]	VERA "equ-qc"	VERA "std"	VERA comparison (both)
$IQD_{equ-qc} = 0.9$	−0.1 - +0.1	—	+0.3 - +0.4
$IQD_{equ-qc} = 0.9$ (100%)	−11% +11%	—	+33% - +44%
$IQD_{std} = 0.6$	—	−0.05 - +0.05	+0.3 - +0.4
$IQD_{equ-qc} = 0.6$ (100%)	—	−8% - +8%	+50% - +66%

First of all, those two examples for the AE and the IQD clearly are not enough to be used as any prove for the overall estimate. But they are doing what an example does. They are showing how the uncertainty deviations look like in a real case. Therefore, both measurements fit the overall found estimate quite well.
We have to keep in mind that the *score uncertainty* effects the *uncertainty due to different observations*. In that case of the absolute error the box plot extent from the "equ-qc" observation exceeds the other one. That is why the estimate due to both ensembles is nearly the same as the VERA "equ-qc" one. For

the IQD, on the other hand, the location from the two box plots is more important. This is because of nearly the same box plot extent. The uncertainty with both ensembles is much bigger than the bootstrapping one.

What should be retained from this section is, that an uncertainty can be produced via ensemble. The way how an ensemble is used depends on the verification purpose. If numerous ensembles are available a comparison can or should be done. The main focus here, is to see which ensemble has the most prominent extent and how far those intervals are apart. This information can be used within the process of decision making, to investigate whether a different ensemble provides more or another uncertainty information. This is important since observation ensembles are highly experimental and, therefore, a comparison is useful.

5. Outlook

As already mentioned multiple times observation ensembles, which are the main part of this thesis, are not that well discussed yet. Hence, different approaches are provided in Section 5.1.
Those ensembles should be applied in the verification. Here, another important and, in this case, well discussed part, is the high impact weather (severe weather). That is way, some of the previously found ensemble verification methods can be adjusted (see Section 5.2).
However, a verification with different forecast models should also be done. Therefore, the MesoVICT project is doing some model re-runs, which should be available soon. An outlook for the upcoming forecast models is done in Section 5.3.

5.1 Different observation uncertainty estimates

The uncertainty estimate discussed in this thesis uses random Gaussian perturbations of station observations to generate ensembles (Gorgas and Dorninger, 2012a). This is not the first attempt of creating observation uncertainty estimates. For example, Ahrens and Jaun (2007) generated random perturbed ensembles via modeling spatial interpolation uncertainty (see Sequential Gaussian Simulation). Germann et al. (2009), on the other hand, tried to use the LU-factorization to get uncertainties from radar data.
Those are just two of multiple examples. A few computational cheap, new or promising approaches are listed underneath. The list does not claim to be complete. The shown points where discovered within the research for this study and since this field is not discussed that widely, different attempts have to be mentioned.

5.1.1 Error Climatology

Australia is a country with multiple severe weather threats. Next to tropical cyclones, thunderstorms, flash floods and droughts, also wildfires are important. Therefore, the Australian Bureau of Meteorology and partners tries to predict the behavior of wildfires. Chris Bridge from the Australian Bureau of Meteorology was gave a talk on the 10/18/2016 at the National Center for Atmospheric Research (NCAR) (Bridge, Chris, 2016).

The fire spread simulators produce 2D information about flame height, fire rate a the burnt area. The required input is the topography, the vegetation and the weather. Especially the weather has a big uncertainty. This fact limits fire behavior models as some of those prediction systems are very sensitive to the initial conditions. To get some information about the uncertainties of weather inputs, something like an “error climatology” is derived. Here, an observation (station or on a grid) is compared to a corresponding forecast. The deviation between the observation and the forecast is the error. Those have been compared multiple times over the last few years. The result is a big data set with errors. Now, the standard deviation or a similar data property can be derived and used as an uncertainty estimate in the fire spread simulators.

Interestingly, those fire bodies are threated like precipitation fields in the (spatial) verification. This topic is feasible for future collaborations.

5.1.2 Spatially correlated random fields

Interpolation and a sparse observation network are already discussed as uncertainty reasons by Gorgas and Dorninger (2012b). Now, Newman et al. (2015) names the measurement representativeness and the measurement errors itself as uncertainty sources. In Newman's paper, an observation ensemble is wanted to make advanced data assimilation systems and other tools in the land-surface and hydrological modeling feasible. This should be applied for a gridded precipitation and temperature field.
Rain measurements are very problematic. Gauge-based estimates have a problem with resolving the full spatial variability of precipitation due to a too low station density or other sampling errors. This is noticeable in complex terrain. Radar measurements capture the spatial variability better, even vertical intersections are possible. However, the ground based radar estimates have uncertainties as well. One of them is the terrain blocking, the curvature of the earth or the uncertainty in the rain rate-reflectivity relationship. The air temperature is easier to measure, but also here similar problems in comparison with the precipitation measure occur. The biggest issue is the observation density. Less observations means more uncertainty especially in complex terrain.

To make some of those uncertainties count the probability approach is necessary. Newman et al. (2015) provide a few examples for already applied observation uncertainty estimates. Most of them use radar or satellite data to create these estimates. They are often made in an ad hoc manner. In the hydrology a robust uncertainty estimate is essential. Therefore, a daily gridded but station-based, ensemble data set of precipitation and temperature is derived. This work is done for North America for the period 1980-2012. Hence, spatially correlated random fields are used. This method provides information about the spatial field by generating random numbers (standard normal distributed) step by step for points on the grid. The new random numbers are conditioned to the previously generated values. 100 ensemble members are produced. This is not the whole story about spatially correlated random fields, but an overview. One advantage of interpolation-based schemes is that the maximum value is not limited to the highest observed measurement. So, extreme values are better represented. The so found ensembles have a good reliability and discrimination, calculated with the corresponding diagrams. Additionally, precipitation occurs more realistic.

5.1.3 Re-sampling approach for daily data

Uncertainty in the observation is also interesting in long-term data. Aalto et al. (2016) use several climate variables for Finland from 1961-2010 and calculates daily gridded time series with permutation-based uncertainty estimates. Furthermore, the climate data is investigated for temporal trends.
To fit the station observations on a regular 10 km × 10 km grid, an interpolation scheme is necessary. Here, the Kriging method is applied. It minimizes the mean squared error between the interpolated and observed values to produce a 2D grid. Additional information (topography and water bodies) are added to increase the interpolation quality.
The so found gridded data set is required to have an uncertainty estimate with as less computational effort as possible. The Kriging interpolation scheme provides a variance, which should not be applied as an uncertainty estimate. Therefore, the daily data is used by taking one specific day and choosing 75% randomly of the sample. The mean is calculated to have one ensemble member. This is done 50 times, every time with a different random 75% sample. The so computed ensemble provides the standard deviation for one grid point at a certain data. This method is robust over the number of interpolation runs and the sample size.

It can be seen that the uncertainty increases towards the north. This regions have a less dense station network, which is responsible for that. Since the estimates are different from day to day a seasonal change is observed. The temperature and the pressure have less uncertainty in the summer. The precipitation and the relative humidity in winter. This could be due to the convective season in the summer. The snow coverage is also produced. The maximum deviation is in the winter, which is not surprising because as there is less or no occurrence of snow during summer. However, the gathered information can be valuable for the verification and network planning.
The trend analysis indicates that several variables have a significant one. As a conclusion, a clear but spatial-wise changing signal of climate change during the observation period is located.

5.1.4 Sequential Gaussian Simulation

So far the grid resolution and the observation density are more important than the interpolation method (Gorgas and Dorninger, 2012b). Whenever a very dense network is available the interpolation method matters again. If this is the case, the Sequential Gaussian Simulation (SGS) is feasible to calculate the spatial interpolation uncertainty (Gorgas and Dorninger, 2012b).
Therefore, several random realizations of interpolation fields are created. However, for the SGS a measure of uncertainty at every grid point is needed. This can be achieved, for example, by using the Ordinary Kriging (OK), which is related to the basic Kriging (see Section 5.1.3). The sequential Gaussian simulation output has an advantage for the OK variance. It provides a gridded covariance structure derived from the observation distribution and the observation values themselves. If an interpolation approach like the VERA (error estimates at observation locations and not grid points) is done, the SGS has to be adjusted for a different uncertainty measure.
Even though the interpolation error is small in comparison with uncertainty sources, for specific application could be useful.

5.1.5 Ensemble of Data Assimilation

Observations are normally collected via ground based observation systems or satellites. At some places the network density is still very low, even with satellites. Whenever an information about such a place is wanted reanalysis has to be performed. Here, a dynamic model and the spare observations are connected to a new observation, the so called reanalysis. This is also done for observation information in the past, because as further back we go as less information is collected.
Patrick Laloyaux from the European Centre for Medium-Range Weather Forecasts (ECMWF) was gave a talk on the 11/29/2016 at the NCAR (Laloyaux, Patrick, 2016) and discussed new features of reanalysis products including uncertainty estimates.

The ECMWF uses a data assimilation scheme (4D-VAR) and an adapted version of a NWP model, where appropriate atmospheric forcings are represented, to gather a reanalysis. A normal NWP model often applies constant forcings, due to the limited time period and the constant forcings during that span. Specifications of errors and biases (see Section 3) are necessary. They often occur whenever a new observation method is established (e.g. radiosondes). Those circumstances are responsible for proper changes in the observing system or in the model physic.
The computational set up changed throughout the previous years. Now, a coupled earth model is used. This means the atmosphere, land, composition, ocean, wave and sea ice are calculated separately. Dependent on the studied time scales (medium-range, monthly, seasonal) different modules are brought together. The latest version is called CERA-20C and is the first Coupled Atmosphere-

Ocean Assimilation System (CERA) from the ECMWF for the 20th century (20C).
To quantify the uncertainty an ensemble of data assimilation is provided. The system initiates perturbations for the observations, sea surface temperature and stochastic physics. The initial condition gets perturbed and afterwards the normal data assimilation system is performed. This is done ten times, with the result of a 10 member data assimilation ensemble.

Reanalysis takes information from observations using the model to spread those in time and space and also across variables. The ECMWF global reanalysis has a big resolution (e.g. land-surface 125 km horizontal grid), which makes the study of regional events insufficient. The solution could be a nested approach, where the global reanalysis is used for the boundary condition. The uncertainties gathered from the ensemble can be applied in the verification, but they are controversial. Since a reanalysis is using an prediction system in the background, the independence is questionable. Observations are wanted without any dynamic model assumptions (e.g. VERA).

5.1.6 Poor man's ensemble

A poor man's ensemble is typically applied for a forecast purpose (Arribas et al., 2005). Here, different NWP's are used to get a probability forecast, where each deterministic model imitates an ensemble member. Since all NWP's are usually calculated independently from each other the computational effort is less compared to the calculation of a whole ensemble prediction system by itself. Additionally, different models and independent analyses are valid. However, the initial condition and model evolution errors are evolving randomly. This is important for short-range weather forecasts, if high impact weather could occur and the probability for an event is wanted. Poor man's ensembles are initially designed for and often applied in the medium-range forecasting.
Arribas et al. (2005) showed that, poor man's ensembles have a competitive performance for short-range prediction in comparison with an ECMWF ensemble prediction system, even with a smaller member size. However, the best performance was achieved for a hybrid configuration. This set up combines output from a small subset of the ECMWF ensemble prediction systems with other different NWP models.

The poor man's ensemble approach can also be used in the purpose of creating an observation ensemble (Gorgas and Dorninger, 2012b). There are different ways to get the necessary ensemble members. A list is provided below. Additionally different possibilities for uncertainty intervals are discussed.

- Observations and subsequently analyses depending on the *interpolation method*. Since NWP's are calculated on a regular grid observations are wanted to be gridded as well. Therefore, the station data is interpolated. In this thesis the VERA is the used method (see Section 1.3.1). In Section 5.1.3 and 5.1.4 the Kriging or ordinary Kriging is introduced. However, the Barnes interpolation is important too. Here, the first produced field has a low resolution and is used a the first guess field. The so provided background information is necessary to continue with a more detailed analysis. Anyway, all of the discussed schemes can be characterized as model independent. That means, they do not need any additional information than the station data for the interpolation.
- Another property which can be changed is the *grid spacing*. A smaller scale is useful, if the station data density is good. Otherwise, a smaller grid resolution, did not result in a better analysis. Whenever the "best" grid space is found a wider one delivers a different field and consequently, uncertainty, if a comparison is done.

- As already mentioned multiple times the *station data density* is the main factor for a good analysis. A originally well covered network can be changed to a different one, if chosen stations are neglected on purpose. The corresponding analysis is different as well and new output is created. In some studies stations are left out to see if the skipped one is representative or a new one would improve the network.
- All numerical weather prediction systems need to know the state of the atmosphere before computing in the future. This is often done with *model-fitted observation* data. A verification with these estimate would not be model independent and not representative. Since, this is done for every NWP an analyses ensemble from each NWP analysis stage is possible. Certainly they are still not model independent, but with the use of multiple models a comparison is more representative. One restriction is, that not for every hour a new analysis is calculated and, therefore, the verification would be incomplete.

A poor men's ensemble can be gathered by performing one option from the above discussed points itself, or by combining multiple approaches. The ensemble spread is already a good estimate for the uncertainty. Additionally, resampling techniques and the sequential Gaussian simulation (see Section 5.1.4) are useful. Note, that the use of different resampling methods (e.g. Bootstrapping, Jackknife) can also be a source for some kind of a poor men's ensemble.

5.2 Extreme value approach

Lerch (2012) commits his diploma thesis to a general verification framework of rare events. As a result, some of his conclusions and methods are interesting for this thesis as well.
Extreme events or high impact weather is often underrepresented in weather models. Therefore, the verification can be tricky. To not underestimate rare events in the ensemble comparison processes (see Section 3) specific parts of the fitted distributions have to be modified. With the kernel density estimate (see Section 2.3) an adjustment is already applied, but the following discussed points have there emphasis on weighting parts of the used distribution and compare areas of interest. Additionally, the generation of extreme value distributions are introduced. However, this is only an overview and no examples are calculated.

5.2.1 Threshold- and quantile-weighted CRPS and weighted IQD

The continuous ranked probability score (see Section 3.1.4) uses an ensemble forecast (via CDF) and verifies it against a given observation. The now provided approach applies a weight. Equation 5.1 uses a threshold-weight. The notation is taken over from Lerch (2012), in case a closer investigation is wanted.

$$CRPS^t(f,y) = \int_{-\infty}^{\infty} (F(z) - \mathbb{1}\{y \leq z\})^2 u(z)\,dz. \tag{5.1}$$

where:

$F(z)$ = CDF for the forecast
y = Observed value
$\mathbb{1}$ = Indicator function
$u(z)$ = Weight function

The weight can also be established quantile-wise. Like Equation 5.1 Equation 5.2 keeps the original notation from Lerch's thesis.

$$CRPS^q(f,y) = 2\int_0^1 (\mathbb{1}\{y \leq F^{-1}(\alpha)\} - \alpha)(F^{-1}(\alpha) - y)\,v(\alpha)\,d\alpha. \tag{5.2}$$

where:

$F^{-1}(\alpha)$ = Quantile forecast for α corresponding to the forecast CDF ($\alpha \in [0,1]$)
y = Observed value
$\mathbb{1}$ = Indicator function
$v(z)$ = Weight function

Both versions of the CRPS are not effected by the weight if $u = 1$ and $v = 1$. The threshold or quantile for the weight in use can be in the center, tails, right or left tail of the distribution. Weight in the tails is important for weather forecasting, because that is the place where high impact weather is represented.

After some adjustments the weight-approach could be applied for the integrated quadratic distance (see Section 3.1.6). This is investigated in Thorarinsdottir et al. (2013).

$$d_{WIQ}(F,G) = \int_{-\infty}^{\infty} (F(t) - G(t))^2\, w(t)\,dt. \tag{5.3}$$

where:

d_{WIQ} = Weighted integrated quadratic distance
$F(t)$ = CDF for the forecast
$G(t)$ = Observed CDF
$w(t)$ = Weight function

Equation 5.3 is shown with the notation from Thorarinsdottir et al. (2013). It is a gernal weight and can be adapted to specific regions.

5.2.2 Conditional and censored likelihood

Like the CRPS and the IQD the Kullback-Leibler divergence (see Section 3.1.8) can be changed to an estimate where a region of interest is highlighted. Here, the weight-approach can not be used. A likelihood-based method is needed in order to compare a specific area of interest. The full likelihood gets replaced with the conditional likelihood (see Equation 5.4).

$$CL(f,y) = -w(y)\log\left(\frac{f(y)}{\int w(s)f(s)ds}\right). \tag{5.4}$$

where:

CL = Generalized conditional likelihood scoring rule
f = Density forecast
y = Observed value
$w(y)$ = Weight function

For Equation 5.5 the full likelihood is exchanged with the censored likelihood.

$$CSL(f,y) = -\left[w(y)\log f(y) + (1-w(y))\log\left(1 - \int w(s)f(s)ds\right)\right]. \tag{5.5}$$

where:

CLS = Generalized censored likelihood scoring rule
$f(y)$ = Density forecast
y = Observed value
$w(y)$ = Weight function

All notations are maintained from Lerch (2012). For a deeper investigation papers refereed in Lerch's thesis are recommended.
The threshold- and quantile-weighted versions of the CRPS as well as the generalized conditional likelihood - and the generalized censored likelihood scoring rule are feasible for computations, if discretized versions are applied.

5.2.3 Extreme value theory

In order to get the probability of an event that is more extreme than any previous one the extreme value theory is necessary (Lerch, 2012). With a set of training data parameters for Equation 5.6 are found. Here, the generalized extreme value distribution is applied.

The aim is to combine three types of limits, known as the Gumbel (Type I), Frechet (Type II) and Weibull (Type III) family. The solution is Equation 5.6, where these families are combined into a single distribution function. Here, for the cumulative distribution function.

$$F_{GEV}(z) = \exp\left\{-\left[1+\xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\right\}. \tag{5.6}$$

where:

F_{GEV} = CDF of the Generalized extreme value distribution
μ = Location parameter
σ = Scale parameter
ξ = Shape parameter

These modeled extreme values can be obtained as values exceeding a high threshold or as block maxima.
Another method to generate extreme values is the generalized Pareto distribution (see Equation 5.7). Therefore, again a sample of training data is needed. The generalized Pareto distribution is used to shape the tails of another distribution.

$$F_{GP}(y) = \begin{cases} 1-\left(1+\frac{\xi(y-\mu)}{\sigma}\right)^{-\frac{1}{\xi}} & \text{for} \xi \neq 0, \\ 1-\exp\left(-\frac{y-\mu}{\sigma}\right) & \text{for} \xi = 0. \end{cases} \tag{5.7}$$

where:

F_{GP} = Generalized censored likelihood scoring rule
μ = Location parameter
σ = Scale parameter
ξ = Shape parameter

The generalized Pareto distribution and generalized extreme value distribution have a close relationship with each other. They deviate in their description of extreme events, but the resulting output is comparable.
Lerch (2012) used the extreme value theory (generalized extreme value and generalized Pareto distributions) to get extreme events for the probabilistic wind speed forecasts of gusts. The result has to exceed a high threshold.

5.3 MesoVICT numerical weather prediction systems

For the MesoVICT project different observations and forecasts are obtainable. The available observations are already discussed in Section 1.3.1 and different approaches for getting new ones, in Section 5.1. Since this thesis is about verification methods and their uncertainty, the forecast is not noticed that much. A summary and outlook for the deterministic (see Section 5.3.1) and ensemble (see Section 5.3.2) forecasts are done.

5.3.1 Deterministic forecast

In the verification the deterministic part was taken by the mean of the ensemble forecast. This was done due to the fact that, no control (main) run or deterministic model from the same provider (ARPA-SIMC) was available. A different model was not used, because consistency should not be jeopardized. However, for the MesoVICT project deterministic forecasts are present (Dorninger et al., 2013).

- The Canadian high-resolution Limited Area Model (GEM-LAM) from environment Canada is a nested version of the Global Environmental Multiscale (GEM≈30 km grid) also from the environment Canada. The grid size is 2,5 km and the forecast time 18 hours.
- MeteoSwiss is also member of the COSMO group, like ARPA-SIMC and uses the ECMWF (≈10 km grid) data as boundary condition. For the MesoVICT cases MeteoSwiss does the operational COSMO-2 model. Here a 2,2 km mesh size is applied. The prediction is 24 hours long.
- Additionally, MeteoSWiss planes to recalculate the MesoVICT cases with the experimental COSMO-1 and the COSMO-7 models. COSMO-1 has a 1,1 km and COSMO-7 a 7,7 km gird size. Both set ups have the ECMWF (≈10 km grid) as boundary condition.

All models have the alpine region as prime domain.

5.3.2 Ensemble prediction system

The main part of the thesis is the ensemble verification. Therefore, one model is introduced (see Section 1.3.2). Another model from MeteoSwiss is ready to use, but has to be released for the MesoVICT community.

- The mentioned MeteoSwiss ensemble is a probabilistic forecast called COSMO-E. It is done for the Alpine region. 21 ensemble members are calculated with a mesh size of 2,2 km. The global framework for the ensemble is an ECMWF product with grid size ≈20 km.
- It would also be possible to produce a poor men's ensemble (see Section 5.1.6), if multiple deterministic forecasts are on-hand. It is imaginable that a hydride set up for the poor men's ensemble is feasible.

All forecasts using the JDC data set as observational input. However, note that all forecast no matter if deterministic or ensemble, are interpolated on the VERA grid (8 km) in the MesoVICT project.

Two projects are maybe important to look at, due to the supply of weather models. The Short-Range Numerical Weather Prediction Programme (`http://srnwp.met.hu/`[03/20/2017]) has members all over Europe and gives information about deterministic and ensemble forecasts. The TIGGE-LAM (`https://software.ecmwf.int/wiki/display/TIGL/TIGGE-LAM+archive`[03/20/2017]) group is specified on ensembles and may overlap with the previous mentioned program. Both could provide useful information, if a verification and no inter-comparison for verification methods is needed.

6. Summary and discussion

The Mesoscale Verification Inter-Comparison over Complex Terrain (MesoVICT) project has the purpose of comparing different spatial verification methods. No model inter-comparison is done (see Section 1.2).
Therefore, different models and observations are available. Including forecast and even observation ensembles. The evaluation of forecast ensembles against one certain observation is well known, but the verification from two ensembles against each other is not discussed that often.
Hence, this thesis provides information and method on how a verification is possible. Since the ensemble comparison approach is quite new and the spatial verification, in the spirit of MeosVICT, is challenging it will be postponed to later studies and the verification is done point-wise.

The numerical weather prediction system in use is the COSMO-LEPS (CLE). It is a gridded limited area prediction system and provided by the weather service of Emilia-Romagna. It has 16 members and uses the ECMWF as boundary condition. The mesh size is 10 km. If a control run or a deterministic model is needed, the mean of all ensemble members takes the place (see Section 1.3.2).
The observation product is calculated from the University of Vienna and uses a model independent real-time mesoscale interpolation method (see Section 1.3.1). This is necessary, because the station observations had to be interpolated on a regular grid. The Vienna Enhanced Resolution Analysis (VERA) has an 8 km grid resolution. The forecast model is interpolated on the same grid. The VERA follows three steps to produce a proper gridded observation. The station observations are controlled before used for the spline interpolation. Additionally, to resolve the small-scale phenomena fingerprints are made.
The observation ensemble is also computed from the University of Vienna. The VERA is the basis for the ensemble. A complex method where the station observation uncertainties and spatial patterns are gathered is used. Furthermore, a stochastic part is included (random Gaussian generator). The results are two observation ensembles called “std” and “equ-qc” with 50 members (see Section 1.3.1). The shortcuts explain how the spatial observation is integrated. The observation ensembles are like the forecast ones a measure for a specific uncertainty, with the addition that they are highly experimental.

Here, the station observations are supplied from two big measurement campaigns during 2007. The so found data set is called JDC (see Section 1.3.1) and is used for all forecast and observation products. Due to the measurement campaign in 2007 the MesoVICT project chose six cases for a verification inter-comparison in this period. Those are selected, because of prominent weather events. Since the VERA ensembles are only available for one case (core case), the time section of the specific case is picked. The core case took place from 20-22 June 2007, where a cold front approaches the Alpine region associated with prefrontal showers.
The areas of interest are the Alps including the Mediterranean Sea. This landscape fits the task of a complex terrain quite well. In this frame nine cities and six areas are under special notice. They are used to picture the properties of defined verification methods. The main focus is on Vienna, Austria.
As parameters the wind speed (V), the potential temperature (Θ), precipitation (RR) and the pressure (p,p_{sf} or p_{md}) are chosen. Here, p is the VERA sea level pressure, p_{sf} is the CLE sea level pressure calculated with the same formula as the VERA (standard formula) and p_{md} applies a model specific formula for the pressure at the lowest model layer. Those set ups are important if a model comparison is the purpose. In this thesis they are just checked for consistency.
The area, parameters and weather condition are debated in Section 1.4.

To determine how the ensembles are distributed certain tests have to be performed. The allocation is important since the verification should be done with distribution-depending measurements. The performed tests are the Lilliefors-, Jarque-Bera- and Anderson-Darling-(see Section 2). They are checking if the Gaussian (normal) distribution is valid. The conclusion is, that the data is not normal distributed all the time. Therefore, nonparametric statistics must be applied. If probability density functions (PDF's) or cumulative distribution functions (CDF's) are used for the evaluation process the kernel density estimate (see Section 2.3) fits the PDF and CDF to the data, which is equal to a nonparametric statistic. Note, that the kernel density estimate is not a panacea. It is sensitive to the bandwidth and different set ups can be used.

The verification in the meteorology can be done with a deterministic forecast against an observation. Furthermore, a forecast ensemble is verified towards one observation. In this thesis a fairly new approach is applied. The forecast ensemble is compared with an observation ensemble. This ensemble can also be used for an uncertainty estimate in the deterministic forecast verification, more on that later.
An eyeball-verification is possible with the box plot, but a quantitative score is wanted. Hence, six methods are selected. The integrated quadratic distance (IQD) uses two CDF's and measures the sum of squared differences. The perfect receiver operating characteristic comparison (PROCC) needs the observation and forecast ensemble represented as PDF's. This score is rather complex, but constructed on a comparison with a perfect result. Probability vectors are required for the Kullback-Leibler Divergence (KLD). It provides the difference of those. The Euclidean, Seorgel and Lorentzian distances are based on histograms and the related probabilities in each bin. All of theses measurements are explained in Section 3. The found methods can also be applied for a comparison of, for example, two forecast models.
The different observation ensembles are compared with each other and a worse result is achieved with the "std" set up. This is due to the fact that the "std" ensemble has a smaller spread. Which leads to the question what spread is wanted, a narrow or a wide one? No answers can be found here, but a intense discussion is suggested.
The ensemble verification with different parameters showed, that it is reasonable to use the wind speed and potential temperature. The pressure outputs evolve differently at the frontal approach. Here, the forecast gets wider while the observation remains narrow. This is because of the big gradient when a front is passing and the time-wise uncertainty. The observation is of good quality every time. The precipitation has to be verified in a different way, because of the occurrence and non-occurrence.
Table 5 provides an overview of the found verification methods and defines a judgmental interval for each score. The kernel density estimate improves as well as deteriorated the scores.

As already mentioned previously it is possible to compute uncertainty estimates via observation ensembles. For scores calculated with one observation each ensemble member can be used as "new" observation. This results in multiple solutions for one method. The so detected interval can be represented as confidence interval or, as it is done in this thesis, via box plot (see Section 4).
For the ensemble verification a different approach is necessary. Here, bootstrapping is applied. With this resampling technique "new" observation ensembles are created. Like before those are applied in order to calculate one score multiple times. Again, the produced interval is shown via box plot. The so found *score specific uncertainty* is compared with the *uncertainty between different observations*. This is essential since observation ensembles are experimental and no claim for correctness is possible. Therefore, the "outer product difference" is applied.
The *score specific uncertainty* interacts with the *uncertainty between different observations*. If the *score*

specific uncertainty is small, the location for the *uncertainty between different observations* is more important. On the other hand, a big extent of the boxplot, due to the *score specific uncertainty* suppresses the location influence in the *uncertainty between different observations* and the *score specific uncertainty* is more prominent. A details summary is given in Section 4.3.

Although two observations ensembles are provided, it is not clear yet, how the perfect observation ensemble should look like. That is why Section 5 provides an overview of different ensemble creation practices. Additionally, possible options for dealing with high impact weather are presented as well as an outlook for a possible model comparison. Since the MesoVICT project is an inter-comparison for methods a inter-comparison for models is interesting too.

The biggest challenge for the observation ensemble topic is a proper spread. Until that is found, the comparison methods are very important in order to keep a reference. Therefore, suitable scores are shown in the thesis. Those are not definitely, but for a topic that has not been in the main focus of research so far, a starting point is provided.

A. Abbreviations

A Adriatic

ABP Area between perfect

AD-Test Anderson-Darling-Test

AE Absolute Error

ARPA-SIMC Hydro-Meteo-Climate Service of the Environmental Agency of Emilia-Romagna

AUC Area under the cure

BS Brier Score

CDF Cumulative Distribution Function

CERA Coupled Atmosphere-Ocean Assimilation System

CLE Consortium for Small-scale Modeling (COSMO) Limited-Area Ensemble Prediction Systems (LEPS)

CLEPS Consortium for Small-scale Modeling (COSMO) Limited-Area Ensemble Prediction Systems (LEPS)

COPS Convective and Orographically-induced Precipitation Study

COSMO Consortium for Small-scale Modeling

CRPS Continuous Ranked Probability Score

CS-Test Chi-Square-Test

D-PHASE Demonstration of Probabilistic Hydrological and Atmospheric Simulation of flood Events in the Alpine region

DWT Discrete Wavelet Transform

EA East Alps

ECMWF European Centre for Medium-Range Weather Forecasts

equ Reference estimate of equal perturbation magnitudes for all stations

GEM Global Environmental Multiscale

GTS Global Telecommunication System

GV Great Vienna

ICP Inter-Comparison Project

IQD Integrated Quadratic Distance

IQR Interquartile Range

JB-Test Jarque-Bera-Test

JDC Joint D-PHASE and COPS

KDE Kernel Density Estimation

KLD Kullback-Leibler Divergence

KS-Test Kolmogorov-Smirnov-Test

LAM Limited Area Model

LEPS Limited-Area Ensemble Prediction System

L-Test Lillieford-Test

MAE Mean Absolute Error

MAP Mesoscale Alpine Programme

MesoVICT Mesoscale Verification Inter-Comparison over Complex Terrain

MSE Mean Squared Error

NCAR National Center for Atmospheric Research

NWP Numerical Weather Prediction

OK Ordinary Kriging

PCA Principal Component Analysis

PDF Probability Density Function

PROCC Perfect Receiver Operating Characteristic Comparison

PV Po Valley

-qc Quality Control applied

QC Quality Control

Q-Q Plot Quantile-Quantile Plot

RMSE Root Mean Square Error

ROC Receiver Operating Characteristic or Relative Operating Characteristic

SG South Germany

SGS Sequential Gaussian Simulation

std The standard deviation of the bias corrected VERA QC residuals of a one-month period

UTC Coordinated Universal Time

VERA Vienna Enhanced Resolution Analysis

WA West Alps

B. Nomenclature

p VERA Mean Sea Level (MSL) Pressure reduced with Standard Formula [hPa]

p_{md} Pressure on the lowest model formula reduced with a specific Model Formula [hPa]

p_{sf} CLE Mean Sea Level (MSL) Pressure reduced with Standard Formula [hPa]

RR Precipitation in three hour [mm/3h]

Θ Potential Temperature [K]

$\mathbf{V}$ Wind Speed [$\mathrm{m\,s^{-1}}$]

C. Bibliography

Aalto, J. et al. (2016). “New gridded daily climatology of Finland: Permutation-based uncertainty estimates and temporal trends in climate: New Gridded Daily Climatology of Finland”. In: *Journal of Geophysical Research: Atmospheres* 121.8, pp. 3807–3823.

Ahrens, B. and S. Jaun (2007). “On evaluation of ensemble precipitation forecasts with observation-based ensembles”. In: *Advances in Geosciences* 10, pp. 139–144.

Arribas, A. et al. (2005). “Test of a Poor Man’s Ensemble Prediction System for Short-Range Probability Forecasting”. In: *Monthly Weather Review* 133.7, pp. 1825–1839.

Bowler, N. E. (2006). “Explicitly Accounting for Observation Error in Categorical Verification of Forecasts”. In: *Monthly Weather Review* 134.6, pp. 1600–1606.

Bridge, Chris (2016). “Evaluation of Fire Spread Simulators”. National Center for Atmospheric Research, Boulder, Colorado, USA, RAL Seminar Series. Bridge Chris, Bushfire predictive services, Severe Weather Section, Bureau of Meteorology, Australia. Talk: 10/18/2016. URL: https://www2.ucar.edu/for-staff/daily/announcement-calendar-event/evaluation-fire-spread-simulators.

Candille, G. and O. Talagrand (2008). “Impact of observational error on the validation of ensemble prediction systems”. In: *Quarterly Journal of the Royal Meteorological Society* 134.633, pp. 959–971.

Candille, G. et al. (2007). “Verification of an Ensemble Prediction System against Observations”. In: *Monthly Weather Review* 135.7, pp. 2688–2699.

Carlos M. Jarque and Anil K. Bera (1987). “A Test For Normality of Observations and Regression Residuals”. In: *International Statistical Review/Revue Internationale de Statistique* Vol. 55 (No. 2), pp. 163–172.

Cha, S.-H. (2007). “Comprehensive Survey on Distance/Similarity Measures between Probability Density Functions”. In: *International Journal of Mathematical Models and Methods in Applied Sciences* 4.1.

Ciach, G. J. and W. F. Krajewski (1999). “On the estimation of radar rainfall error variance”. In: *Advances in Water Resources* 22.6, pp. 585–595.

Dorninger, M. et al. (2009). *Joint D-PHASE-COPS data set (JDC data set)*. Technical report. University of Vienna.

Dorninger, M. and T. Gorgas (2013). “Comparison of NWP-model chains by using novel verification methods”. In: *Meteorologische Zeitschrift* 22.4, pp. 373–393.

Dorninger, M. et al. (2013). *MesoVICT: Mesoscale Verification Inter-Comparison over Complex Terrain*. National Center for Atmospheric Research.

Fawcett, T. (2006). “An introduction to ROC analysis”. In: *Pattern Recognition Letters* 27.8, pp. 861–874.

Ferro, C. A. T. (2007). “A Probability Model for Verifying Deterministic Forecasts of Extreme Events”. In: *Weather and Forecasting* 22.5, pp. 1089–1100.

Germann, U. et al. (2009). “REAL-Ensemble radar precipitation estimation for hydrology in a mountainous region”. In: *Quarterly Journal of the Royal Meteorological Society* 135.639, pp. 445–456.

Gilleland, E. et al. (2009). “Intercomparison of Spatial Forecast Verification Methods”. In: *Weather and Forecasting* 24.5, pp. 1416–1430.

Gorgas, T. and M. Dorninger (2012a). “Concepts for a pattern-oriented analysis ensemble based on observational uncertainties”. In: *Quarterly Journal of the Royal Meteorological Society* 138.664, pp. 769–784.

Gorgas, T. and M. Dorninger (2012b). “Quantifying verification uncertainty by reference data variation”. In: *Meteorologische Zeitschrift* 21.3, pp. 259–277.

Kullback, S. (1978). *Information theory and statistics*. OCLC: 187308462. Gloucester, Mass: Smith. 399 pp.

Laloyaux, Patrick (2016). “Earth system climate reanalyses at ECMWF”. National Center for Atmospheric Research, Boulder, Colorado, USA, CGD Seminar Series. Laloyaux Patrick, ECMWF, Earth System Assimilation Section. Talk: 11/29/2016. URL: https://www2.ucar.edu/for-staff/daily/announcement-calendar-event/cgd-seminar-2.

Lerch, S. (2012). “Verification of probabilistic forecasts for rare and extreme events”. PhD thesis. Heidelberg University.

Marsigli, C. et al. (2005). “The COSMO-LEPS mesoscale ensemble system: validation of the methodology and verification”. In: *Nonlinear Processes in Geophysics* 12.4, pp. 527–536.

Marzban, C. (2004). “A Comment on the ROC Curve and the Area under it as Performance Measures”. In: *Weather and Forecasting* 19.6.

Mayer, D. et al. (2012). “Innovations and applications of the VERA quality control”. In: *Geoscientific Instrumentation, Methods and Data Systems* 1.2, pp. 135–149.

Metz, C. E. (1978). “Basic principles of ROC analysis”. In: *Seminars in Nuclear Medicine* 8.4, pp. 283–298.

Murphy, A. H. (1973). “A New Vector Partition of the Probability Score”. In: *Journal of Applied Meteorology* 12.4, pp. 595–600.

Newman, A. J. et al. (2015). “Gridded Ensemble Precipitation and Temperature Estimates for the Contiguous United States”. In: *Journal of Hydrometeorology* 16.6, pp. 2481–2500.

Raith, W. and S. J. Bauer, eds. (2001). *Erde und Planeten*. 2., aktualisierte Aufl. Lehrbuch der Experimentalphysik 7. Berlin: de Gruyter. 727 pp.

Sperka, S. and R. Steinacker (2011). “A Quality-Control and Bias-Correction Method Developed for Irregularly Spaced Time Series of Observational Pressure Data”. In: *Journal of Atmospheric and Oceanic Technology* 28.10, pp. 1317–1323.

Steinacker, R. et al. (2000). “A Transparent Method for the Analysis and Quality Evaluation of Irregularly Distributed and Noisy Observational Data”. In: *Monthly Weather Review* 128.7, pp. 2303–2316.

Steinacker, R. et al. (2006). “A Mesoscale Data Analysis and Downscaling Method over Complex Terrain”. In: *Monthly Weather Review* 134.10, pp. 2758–2771.

Steinacker, R. et al. (2011). “Data Quality Control Based on Self-Consistency”. In: *Monthly Weather Review* 139.12, pp. 3974–3991.

Stephens, M. A. (2006). “Goodness of Fit, Anderson-Darling Test of”. In: *Encyclopedia of Statistical Sciences*. Ed. by S. Kotz et al. DOI: 10.1002/0471667196.ess0041.pub2. Hoboken, NJ, USA: John Wiley & Sons, Inc.

Thorarinsdottir, T. L. et al. (2013). “Using Proper Divergence Functions to Evaluate Climate Models”. In: *SIAM/ASA Journal on Uncertainty Quantification* 1.1, pp. 522–534.

Wilks, D. S. (2011). *Statistical methods in the atmospheric sciences*. 3rd ed. International geophysics series v. 100. Amsterdam ; Boston: Elsevier/Academic Press. 676 pp.

D. Acknowledgment

A thesis is not just written by one person, numerous are needed. That is why I would like to thank:

- The team from the institute for meteorology and geodynamic at the university Vienna, for helping me with organization procedures and giving me a place to work.
- Ass. Prof. Mag. Dr. Manfred Dorninger, that he introduced me to the MesoVICT project and his members. Additionally, he was always available if help was needed and gave me useful information. Without him the work could not be done and my time at the NCAR would not be possible.
- All of the MesoVICT members for providing the data and a bibliography.
- Barbara Brown and Eric Gilleland, that both agreed to be my hosts at the NCAR and due to that, giving me a successful and unforgettable time in the Stats.
- Tara Jensen, Christopher Williams, Andrew Newman, Paul Kucera and Martyn Clark for the numerous talks during my time in Boulder.
- Christopher Ferro (University of Exeter), Chris Bridge (Bureau of Meteorology, Australia), Juha Aalto (Finnish Meteorological Insititute), Tilmann Gneiting (Heidelberger Institut für Theoretische Studien) and Patrick Laloyaux (ECMWF), that there was always time for a meeting and/or an immediate reply to every e-mail.
- Theresa Gorgas, for producing the observation ensembles, without them my work would not be possible, and for being available to discuss problems.
- My advisor o. Univ. Prof. Dr. Reinhold Steinacker, that he let me work on my own, but was there when help was needed. Furthermore, without his agreement the visit at the NCAR would not be possible.
- Christian Groinig, for correcting my thesis.
- Vinzent Klaus and Manuel Weber, for supporting me during our time of studying.
- Last but definitely not least a big thanks goes to my father and mother, as well as my relatives and friends. Without them my thesis and my study of meteorology would not be possible.

E. Danksagungen

Für eine Abschlussarbeit sind immer mehr Personen als der Verfasser selbst notwendigt. Daher möchte ich meinen Dank aussprechen:

- Dem Team des Institutes für Meteorologie und Geodynamik in Wien, für die Hilfe bei organisiatorischen Abläufen und der Bereitstellung eines Arbeitsplatzes.
- Ass. Prof. Mag. Dr. Manfred Dorninger, dass er mich mit dem MesoVICT Projekt und seinen Mitgliedern bekannt gemacht hat. Außerdem hatter er immer Zeit für Fragen und stehts hilfreiche Tipps zur Verfügung. Ohne ihn wäre die Arbeit nie fertig geworden und ich hätte auch nicht drei Monate am NCAR arbeiten dürfen.
- Den gesamten MesoVICT Mitgliedern für die Bereitstellung der Daten und der Literatursammlungen.
- Barbara Brown und Eric Gilleland, dass Sie sich als meine Gastgeber am NCAR zur Verfügung gestellt haben und mir dadurch eine lehrreiche und einzigartige Zeit in Amerika ermöglichten.
- Tara Jensen, Christopher Williams, Andrew Newman, Paul Kucera und Martyn Clark für die zahlreichen Unterhaltungen während meiner Zeit in Boulder.
- Christopher Ferro (University of Exeter), Chris Bridge (Bureau of Meteorology, Australia), Juha Aalto (Finnish Meteorological Insititute), Tilmann Gneiting (Heidelberger Institut für Theoretische Studien) und Patrick Laloyaux (ECMWF), dass jedem Treffen zugestimmt und/oder jeder E-Mail umgehend geantwortet wurde.
- Theresa Gorgas, für das erstellen der Beobachtungsensembles, ohne welche meine Arbeit gar nicht möglich gewesen wäre und für ihre Zeit, um Unklarheiten auszuräumen.
- Meinem Betreuer o. Univ. Prof. Dr. Reinhold Steinacker, dass er mir stehts freie Hand ließ, aber bei wichtigen Fragen immer weiter half. Außerdem wäre ohne sein Einverständins mein Aufenthalt am NCAR nicht möglich gewesen.
- Christian Groinig, für das Korrekturlesen der Arbeit.
- Vinzent Klaus und Manuel Weber, für ihre Unterschützung während unserer gemeinsamen Studienzeit.
- Abschließend möchte ich den größten Dank meinem Vater und meiner Mutter, sowie allen Verwandten und Freunden aussprechen. Ohne Sie wäre mein Studium, geschweige denn diese Arbeit, nicht möglich gewesen.