Comparison of spatial verification methods

Bachelor Thesis

Ludwig-Maximilians-University Munich
Meteorological Institute Munich

Submitted by **Stefan Geiß**
Supervisor **Dr. Christian Keil**

August 11, 2015

Vergleich von räumlichen Verifikationsmethoden

Bachelorarbeit

Ludwig-Maximilians-Universität München
Meteorologisches Institut München

Eingereicht von **Stefan Geiß**
Betreuer **Dr. Christian Keil**

11. August 2015

Erklärung

Hiermit versichere ich, dass ich diese Bachelorarbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe.

München, den 11.08.2015

.....
Stefan Geiß

Abstract

The Mesoscale Verification Inter-Comparison over Complex Terrain (MesoVICT) with a set of six cases tries to explore new verification methods for more realistic meteorological scenarios. In this thesis the first core was chosen for the 20 - 22 June 2007 in and around the Alps region to investigate the comparability, quality and consistency of new spatial verification methods such as fractions skill score (FSS), structure amplitude length (SAL) and displacement and amplitude score (DAS) with a focus on the location components. The methods were applied to the VERA analysis and compared to COSMO-2 and GEM-LAM, all with a resolution of 8 km. The verified parameters are precipitation (1h accumulated) and wind strength. High bias percentiles are used instead of fixed thresholds. They also allow for an investigation of the spatial distribution of phenomena. It will be shown that to a high degree all methods in use lead to similar location errors.
All methods in use assess the 20 June as the day with the highest values in location errors, caused by low synoptic forcing, in contrast to 21 June with high synoptic forcing. Additionally there exists a correlation between location components for precipitation on 21 June.

Contents

1 Introduction

The continously increasing resolution of operational numerical weather prediction models (NWP), mainly due to greater computing power, leads to improved predictions of local weather, e.g. distribution of precipitation with more realistic spatial structure. Yet, mesoscale phenomena like squall lines are routinely forecasted. At small spatial scales forecast errors grow more rapidly (Lorenz 1969) and so the predictability has a natural limit.

For meteorological features with small errors in displacement or in the timing, traditional categorical verification scores such as Gilbert skill score (GSS; or equitable threat score) and threat score (critical success index, CSI) result in false alarms and missed events which get worse for smaller grid spacing (Wilks 2011). The feature can be penalized twice, if the feature is displaced slightly in space (and/or time), once for missing the observations and again for giving a false alarm (Gilleland 2009).

But what is a good forecast? As an essential part of NWP, verification of numerical forecasts has to describe general characteristics of a good forecast. Murphy (1993) defined three types of goodness in terms of consistency, quality (or goodness) and value.

To get a more informative forecast evaluation, new spatial verification methods were developed. The Spatial Forecast Verification Methods Intercomparison Project (ICP), established in 2007 aimed at comparing, developing and getting a better understanding of these new methods and tried to answer some questions such as how each method informs about forecast performance overall and whether the methods inform about location errors. The first phase focused on quantitative precipitation forecasts across the central United States.

The second phase, the Mesoscale Verification Inter-Comparison over Complex Terrain (MesoVICT) tries to explore new methods for more realistic meteorological scenarios, with more variables in addition to precipitation and were applied to Europe including ensembles of forecasts and observations as well. A set of six cases was selected to cover a wide range of interesting meteorological phenomena that developed over time.

In this thesis the first core was chosen for the 20 - 22 June 2007 in and around the Alps region.
The aim of this paper is to investigate three different spatial verification methods on the selected area for deterministic forecasts of precipitation (1h accumulation) and wind strength. Furthermore their comparability to each other was examined as well as how they yield identical information such as location errors but from different perspectives.

The high spatiotemporal variability of precipitation and wind strength poses challenges for accurate predictions. However a good forecast of these two meteorological phenomena is very important, particularly because of extreme events and their far-reaching consequences such as floods and wind storms, big economic damage and social effects.
A second objective is to compare the results for two NWP-models (COSMO-2 & GEM-LAM). The used observation field is provided by VERA (Vienna Enhanced Resolution Analysis) having the advantage that it provides a regular grid in mountainous terrain by interpolation of sparsely and irregularly distributed observations, explicitly for precipitation and wind in order to get the observation field and the forecast field on the same grid.

For this endeavour a set of three verification methods was chosen: Fractions Skill Score (FSS) (Roberts and Lean 2008), Structure Amplitude and Location (SAL) (Wernli et al. 2008) and Displacement and Amplitude Score (DAS) (Keil and Craig 2007&2009) which are described in Chapter 3. Afterwards the results will be shown in Chapter 4 with the verification of precipitation and of wind. A discussion and conclusions are to follow.

2 Data and models

The data collection for MesoVict contains observations, VERA analyses and deterministic and ensemble model forecasts of the World Weather Research Programme (WWRP) Forecast Demonstration Projects (FDP): Mesoscale Alpine Programme (MAP) D-Phase (Rotach et al., 2009) and Convective and Orographically-Induced Precipitation Study (COPS) (Wulfmeyer et al., 2008).

2.1 Observation data

The so-called JDC data set (Dorninger et al., 2009; Gorgas et al., 2009) consists of reports from more than 12,000 stations all over Central Europe with a mean station distance of approximately 16 km (Figure 1), which have been provided by the GTS as well as other networks during the whole year 2007. It has been established as a unified data set of surface observations in a joint activity of MAP D-Phase and COPS. The data set is used to compute the VERA analysis.

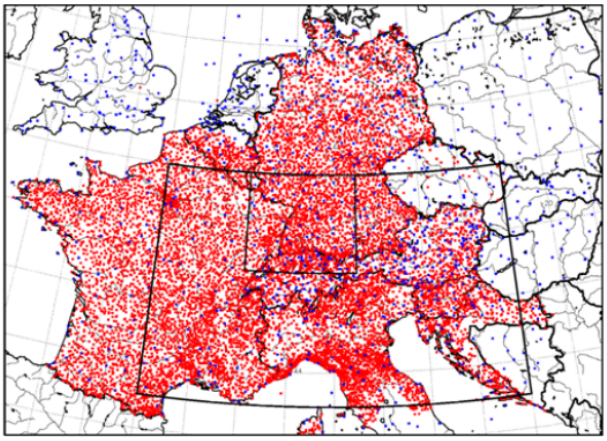


Figure 1: Station locations as used in the JDC data set. Blue: GTS stations; red: non-GTS stations. Frames indicate COPS regions (smaller frame) and D-PHASE region (larger frame), respectively from (Dorninger et al., 2013).

2.2 VERA

The Vienna Enhanced Resolution Analysis (VERA) scheme (Steinacker et al., 2000) contains an analysis algorithm which is based on a thin-plate spline approach and focuses on the interpolation of sparsely and irregularly distributed observations to a constant grid in mountainous terrain (complex topography); in our case, the resolution is 8 km and covers the larger D-PHASE domain (Dorninger et al., 2013)
The big advantage is that no first guess field of a NWP-model is needed as background information, so the interpolation between the grid points is independent of the model. A quality control scheme, named VERA-QC (Steinacker et al., 2011), is used to pre-process the observation input data (of the GTS stations). The VERA-QC avoid artificial and unintentional patterns.
The output parameters include but are not limited to mean sea level pressure, surface potential and equivalent potential temperature (2m), near surface wind (10 m) and accumulated precipitation. The quality of the analysis is good as long as there is an adequate coverage of observation stations (Dorninger et al., 2008). This is ensured by GTS stations.

2.3 NWP-models

For MESOVict, datasets of two NWP-models were interpolated on the VERA-grid with a horizontal resolution of 8 km to apply verification methods. The output parameters are the same as for VERA.

COSMO-2

COSMO-2 is the high-resolution version of the non-hydrostatic meso-scale numerical model with full physical parametrisations weather forecasting model of the COSMO (**Co**nsortium for **S**mall-scale **Mo**deling) community (Steppler et al. 2003) and the operational MeteoSwiss forecasting tool. It covers the Alps region with a horizontal resolution of 2.2 km and 60 vertical levels. COSMO-2 is nested in the regional COSMO-7 model with 6.6 km mesh size, covering Central Europe, which obtains the boundary and initial conditions of the global

IFS model from ECMWF (Rossa et al., 2009; Baldauf et al., 2011). A data assimilation system based on a nudging technique (Schraff, 1997) is used for conventional observations. COSMO-2 has a forecast range of 24 h, starts at 00 UTC (Weusthoff et al., 2010).

GEM-LAM

The local Canadian high resolution **L**imited-**A**rea **M**odel (LAM) is nested in the non hydrostatic version of the **G**lobal **E**nvironmental **M**ultiscale (GEM) with a horizontal resolution of 2.5 km (Rombough et al., 2010) and 58 vertical levels (Erfani, 2005). The forecast was computed and provided over Europe. GEM-LAM has a forecast range of 24 hrs as well, but starts at 06 UTC.

3 Methods

The traditional grid-point-by-grid-point verification methods do not provide essential information about forecast performance. The skill scores used in this paper are calculated by using percentiles instead of fixed thresholds as the aim was to investigate the spatial distribution of phenomena. Smaller percentile thresholds are sensitive for larger-scale, flat features and higher percentile thresholds indicate small and peaked features (Roberts 2008). In recent years, a great variety of spatial verification methods has been developed.
To avoid that spatial errors are penalized twice (double penalty problem) for being a near miss, and again for being a false positive, it is necessary to choose a suitable verification method that also considers multiple scales. (Dey et al. 2014) For this purpose, three approaches are applied in this work.

3.1 Fractions Skill Score (FSS)

In ROBERTS and LEAN (2008) the FSS is described. First, compare two fields of fraction by the Mean squared error (MSE) (from a model and observations, denoted O and M). To calculate this, a threshold is selected as a fixed value (e.g. $1mmh^{-1}$) or as a percentile (e.g. the top 5 % of the precipitation field). This means, that the fields are converted to binary form with grid points set to 0 for values below the threshold and 1 for values above. Then the spatial window is selected and, for each neighborhood centred in a grid point, the fraction of grid points with the value '1' within this square is computed.

$$MSE_{(n)} = \frac{1}{N_x N_y} \sum_{i=1}^{N_x} \sum_{j=1}^{N_y} [O_{(n)i,j} - M_{(n)i,j}]^2. \tag{1}$$

The FSS is defined in terms of the ratio of MSE_n and $MSE_{(n)ref}$:

$$FSS_{(n)} = 1 - \frac{MSE_{(n)}}{MSE_{(n)ref}}, \tag{2}$$

where $MSE_{(n)ref}$ is the maximal MSE-value that can be obtained from the forecast and observed fractions.

$$MSE_{(n)ref} = \frac{1}{N_x N_y} \sum_{i=1}^{N_x} \sum_{j=1}^{N_y} [{O_{(n)i,j}}^2 + {M_{(n)i,j}}^2]. \quad (3)$$

The FSS varies between 0 (complete mismatch between observation- and forecast field) and 1 (perfect forecast). If there are no events forecast and some occur, or some occur and none are forecast the score is always 0. The believable skill is given by

$$FSS_{believable} \geq 0.5 + \frac{f_0}{2}, \quad (4)$$

where f_0 is the domain average observed fraction (e.g. for the 99th percentile: $f_0 = 0.01$). If f_0 is small, FSS can be approximated as $FSS_{believable} \geq 0.5$. For higher f_0 the approximation is not valid. The score is most sensitive to rare events (e.g. convective events).
As an example, Figure 2 illustrates how the algorithm works:

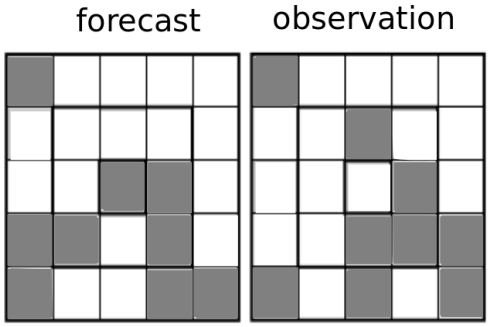


Figure 2: A schematic comparison between forecast and observation.

The grid squares which set to '1' are shaded (threshold has been exceeded) and those which are set to '0' are coloured white (threshold has not been reached). For the central-grid square the forecast fraction is $1/1 = 1$ and the observation fraction is $0/1 = 0$. For the 3 x 3 square the forecast fraction and observation fraction is $4/9 = 0.44$ and for the 5 x 5 square the fractions are equal too and have a value of $9/25 = 0.36$, so the FSS = 1 and forecast is correct for the larger domains.

3.2 Structure Amplitude Length Score (SAL)

WERNLI et al. (2008) formulated the object-oriented verification method SAL based on three components of forecast errors: Structure errors 'S', amplitude errors 'A' and location errors 'L'.
In WERNLI et al. (2008) a threshold depending on the amount of precipitation, like $R^* = R^{max}/15$ is used. The comparison of the weighted sums of forecast precipitation and observation precipitation objects yields the structure component (Eq. 5), with R_n as the precipitation sums of the objects weighted by the object maxima, $V_n = R_n / R_n^{max}$. The values of the S-component lie in the interval $[-2, 2]$.

$$S = \frac{V(R_f) - V(R_a)}{0.5[V(R_f) + V(R_a)]} \tag{5}$$

$$V(R) = \frac{\sum_{n=1}^{M} R_n V_n}{\sum_{n=1}^{M} R_n} \tag{6}$$

The amplitude component (Eq. 7) of SAL resembles a normalised bias with $\overline{R}_f$ and $\overline{R}_a$ as the mean forecast and observed precipitation amounts. It provides a measure of the quantitative accuracy of the total amount of precipitation for the whole domain. A ranges in $[-2, 2]$.

$$A = \frac{\overline{R}_f - \overline{R}_a}{0.5[\overline{R}_f + \overline{R}_a]} \tag{7}$$

Finally, the location component (Eq. 8) consists of two components. The first one (Eq. 9) is the normalized distance between the centres of mass of the modelled and observed precipitation field with $x(R)$ as the overall mass centres and d_{max} as the maximal distance in the fields. The second part (Eq. 10) describes the resemblance of distribution of the objects within the fields. The mass weighted distance between the overall mass centres and the individual objects is given by L. L takes values in the interval $[0, 2]$.

$$L = L1 + L2 \tag{8}$$

$$L1 = \frac{|x(R_f) - x(R_a)|}{d_{max}} \tag{9}$$

$$L2 = 2\frac{|r(R_f) - r(R_a)|}{d_{max}} \tag{10}$$

$$r = \frac{\sum_{n=1}^{M} R_n |x - x_n|}{\sum_{n=1}^{M} R_n} \tag{11}$$

In terms of all components, 0 denotes perfect forecast. If $S > 0$, the structure of the field is too large and/or too flat in contrast to $S < 0$, which implies that the precipitation field is to too small and/or too peaked. A negative (positive) value of A implies an underestimation (overestimation).

3.3 Displacement and Amplitude Score (DAS)

The DAS method is an optical flow-based technique intended to answer the question how good a forecast is in terms of amplitude and location is described in KEIL and CRAIG (2007 & 2009) and summarized here for ease of reading.
By using an optical flow technique, a vector field will be computed that deforms, or 'morphs', one image into a replica of another and all features will be displaced simultaneously in the image. It is based on a pyramidal matching algorithm and is described in detail by Zinner et al (2008) and Keil and Craig (2007). Therefore different grain sizes can be defined as 2^F, with the subsampling factor F.
To calculate DAS, two fields are constructed: A displacement error field $DIS_{obs}(x, y)$ in the obeservation space to show the magnitude of the displacement vector, and an amplitude error field $AMP_{obs}(x, y)$ by using the root-mean-square (RMS) difference between the observation field and the morphed forecast field. Wherever the observation field is zero, both fields are set to zero, so errors are only defined where an observed feature is present.
Whenever $DIS_{obs}(x, y) = 0$ either no feature was forecast within the maximum search distance D_{max} or it is a perfect location forecast. A feature being missed implies that the amplitude error will be large. A non-zero value means that there was a forecast feature within the maximum search distance D_{max}.
In forecast space $DIS_{fct}(x, y)$ and $AMP_{fct}(x, y)$ are computed similarly, but now a large amplitude error for a feature and zero value for the displacement error indicate a false alarm.

The algorithm to calculate DAS can be described as follows: First, the average of RMS of the amplitude errors over the verification area A gives a scalar amplitude score in observation space and is defined as

$$\overline{AMP}_{obs} = \frac{1}{n_{obs}} \left[\sum_{A} AMP_{obs}(x, y)^2 \right]^{\frac{1}{2}}. \tag{12}$$

The mean displacement error in the observation space is computed as

$$\overline{DIS}_{obs} = \frac{1}{n_{obs}} \sum_{A} DIS_{obs}(x, y). \tag{13}$$

$\overline{AMP}_{fct}$ and $\overline{DIS}_{fct}$ are defined analogously. n_{obs} and n_{fct} are the numbers of non-zero points in the observation- and forecast field, respectively. The weighted averages define the amplitude error (AMP)

$$AMP = \frac{1}{(n_{obs} + n_{fct})}(n_{obs}\overline{AMP}_{obs} + n_{fct}\overline{AMP}_{fct}) \tag{14}$$

and displacement error (DIS):

$$DIS = \frac{1}{(n_{obs} + n_{fct})}(n_{obs}\overline{DIS}_{obs} + n_{fct}\overline{DIS}_{fct}). \tag{15}$$

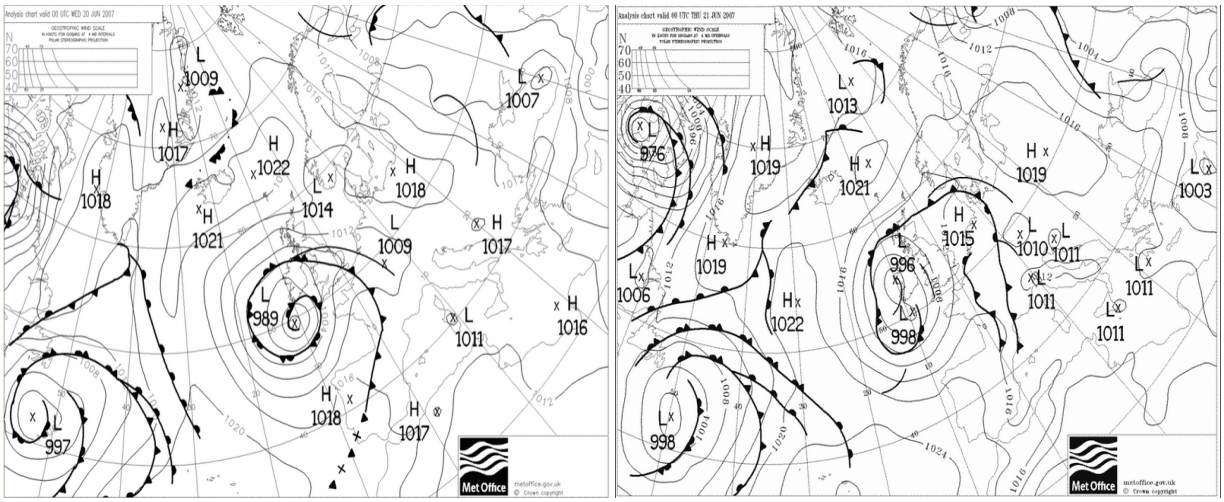
Finally, the displacement-amplitude score, DAS, is defined as the average of the two normalized components. DIS will be weighted by the maximum search distance D_{max} and AMP will be normalized by a characteristic intensity I_0 whose value depends on the application. Keil and Craig (2008) suggest the RMS amplitude of the observed field or for large datasets, I_0 could be specified by a climatological rain rate.

$$DAS = \frac{DIS}{D_{max}} + \frac{AMP}{I_0}. \tag{16}$$

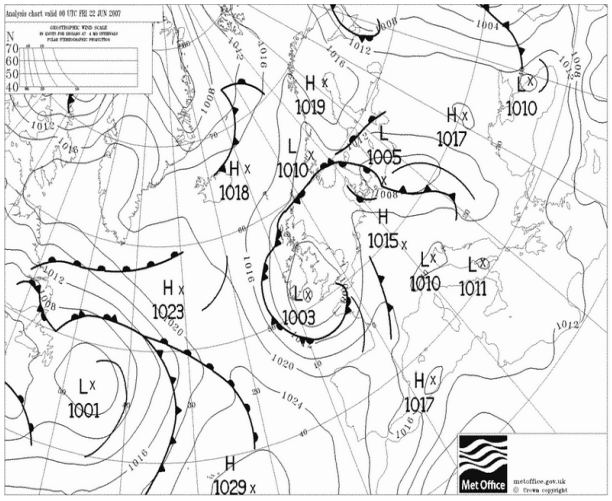
The DAS values are in range $[0, +\infty]$, but it is normally of order one.

4 Verification of precipitation and wind

First, we consider the synoptic situation and in addition, the rationale for using percentiles. Afterwards the results will be analysed and important outcomes will be shown for the used verification methods by using the VERA analysis and the NWP model COSMO-2. GEM-LAM results will be discussed in the next chapter to focus on COSMO-2 in this chapter only. The latter's run starts at 00 UTC in contrast to GEM-LAM which starts at 6 UTC. So the consideration of the entire day leaves more time for verification.
On 20 June a trough was located over the British Isles and ahead of that, warm moist air was advected towards the Alpine region (Figure 3 (a)). As a result, strong convective events began in the evening of 20 June 2007, in the region north of the Alps. To illustrate this, Figure 4 (a - d) show the xy-plots on this particular day at 19 UTC for precipitation and an accumulation period of 1h Fig. 4 (a, b). The wind field, represented by the strength and direction (represented by vectors), is shown in Fig. 4 (c, d). The precipitation field caused by convection seems spotty and heterogeneous for the model forecast and the analysis in contrast to the wind field, which seems relatively homogeneous.
On 21 June, a cold front moved rather quickly eastwards and reached the Alps from the west at circa 5 UTC (Figure 3 (b) & (c)) and again, convective events are observed ahead of the front and strong westerly winds occurred. The cold air mass is too shallow and cannot spill over the mountains.
Figure 5 (a - d) show the frontal structure for the 21 June at 16 UTC. Now, the rain area is more widespread for both, the model and the analysis, and the structure of the front is stretched from south-west to north-east. The wind-vectors rotate through 90° behind the front what is well computed ("Eyeball" verification). In the area of the front, the wind strength is lower in comparison to the near surrounding area. It is remarkable that the 10 m wind magnitude is relatively low in the Alps region all over the time, what could depend on an error in the analysis and in the model forecast, caused by the orographic influence on the model forecast and VERA analysis.
The last day, 22 June, shows a situation like on 20 June with convective events by sun heating, because the cold front left the domain.



(a) (b)



(c)

Figure 3: Met Office surface analysis valid at 00 UTC on 20 (a), 21 (b) and 22 (c) June 2007.

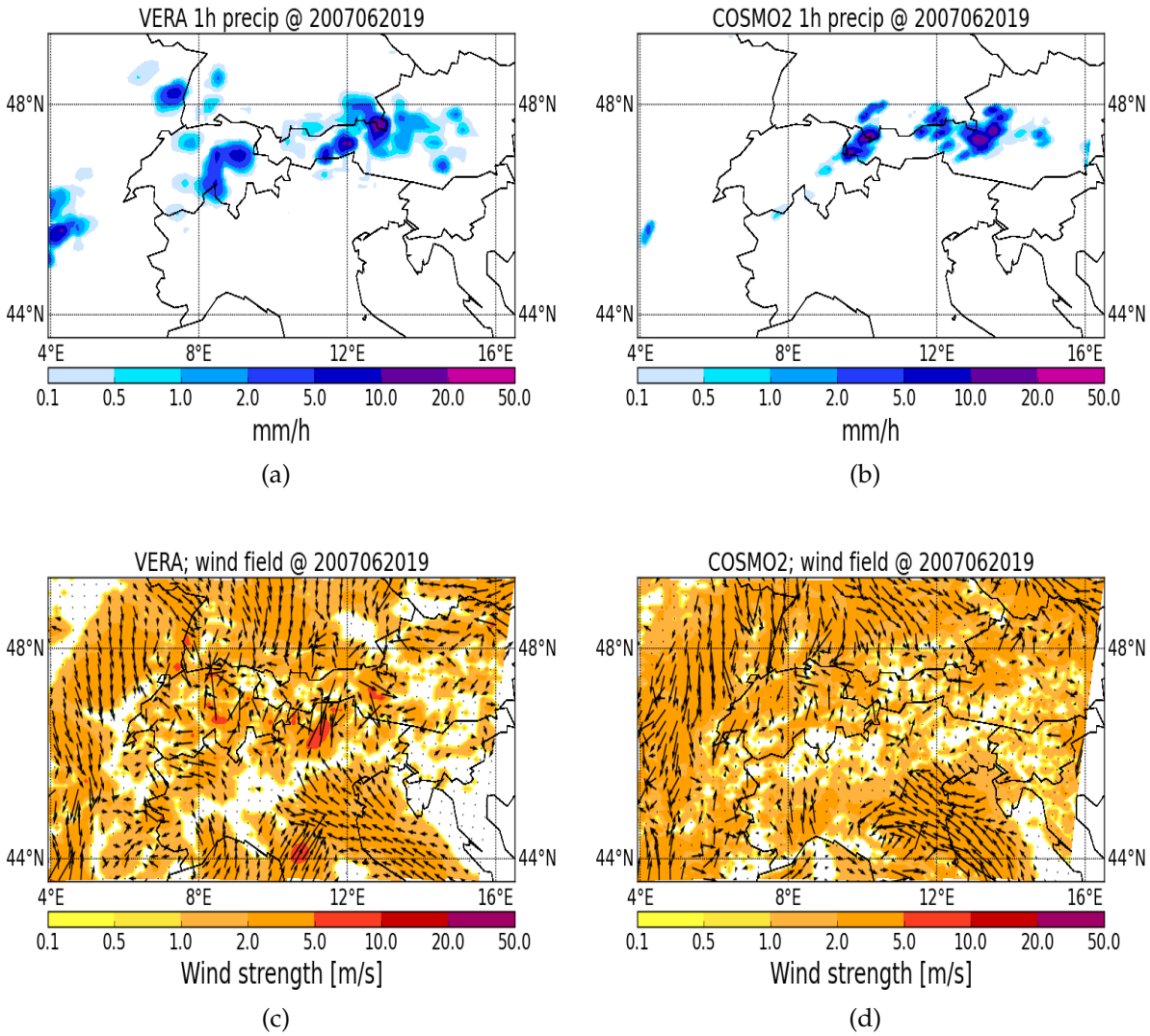


Figure 4: (a)&(b) Precipitation field in [mm/h] and (c)&(d) wind field including magnitude in [m/s] and direction for VERA analysis (a)&(c) and COSMO-2 (b)&(d) on 20 June at 19 UTC.

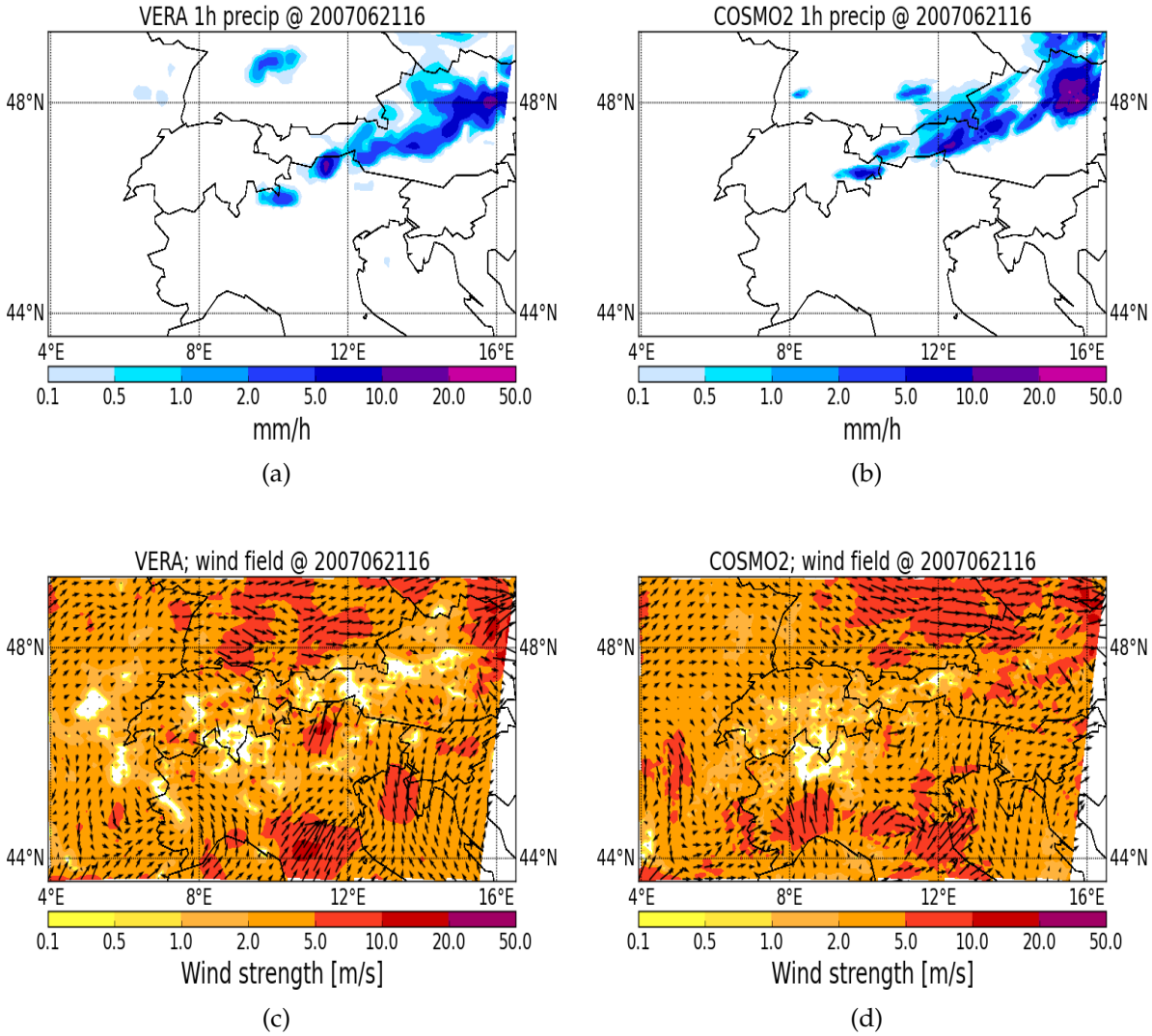
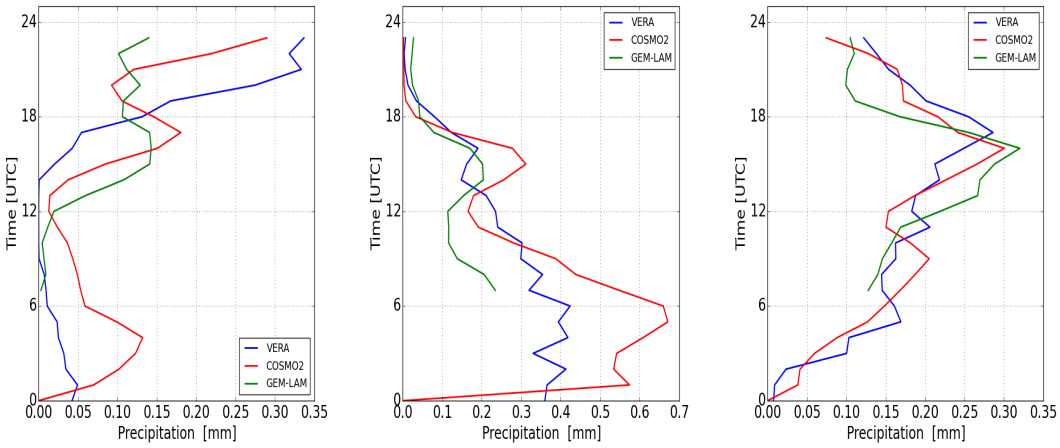


Figure 5: (a)&(b) Precipitation field in [mm/h] and (c)&(d) wind field including magnitude in [m/s] and direction for VERA analysis (a)&(c) and COSMO-2 (b)&(d) on 21 June at 16 UTC.

4.1 Precipitation and wind percentiles

The VERA analysis (blue) and the NWP models COSMO-2 (red) and GEM-LAM (green) are shown in the time series for the 1h accumulation average of precipitation (Figure 6). This plot shows a relatively high bias, especially for the 21 June, but also for the 20 June. Therefore it is advisable to use percentiles.

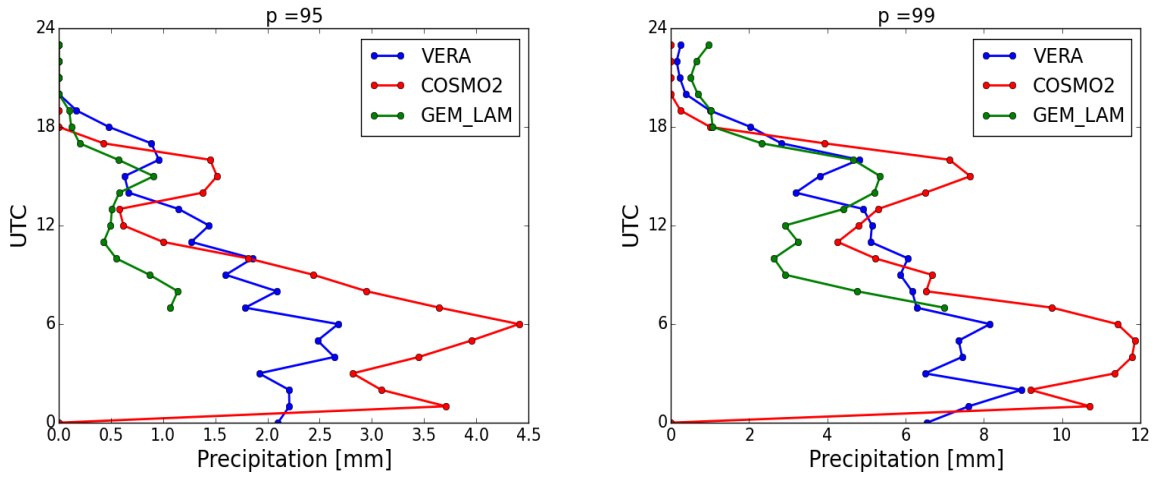


(a) (b) (c)

Figure 6: Time series of mean precipitation [mm] per grid-point from 20 - 22 June 2007 (a) - (c).

On 22 June, the bias is nearly negligible (for a particular time), but it is remarkable that between two time steps a bias exists. So for VERA, the peak of the precipitation amount is at 17 UTC and for COSMO-2 and GEM-LAM at 16 UTC. The usage of percentiles instead of fixed thresholds overcomes model bias and allows the spatial distribution of phenomena to be investigated. Again, this supports the importance of using percentiles instead of fixed thresholds. On all three days it is noticeable that the precipitation time series is smoother for the models than for the VERA analysis. The bias in the models compared to VERA analysis and the variation over time are an important consideration for model evaluation.

Through this thesis, the 95th and 99th percentile thresholds are used which are associated with smaller, more extreme forecast features.



(a) (b)

Figure 7: Precipitation [mm] of VERA, COSMO-2 and GEM-LAM for the 95th (a) and 99th percentile threshold (b) on 21 June 2007.

In Figure 7, VERA (red), COSMO-2 (blue) and GEM-LAM (green), differ by up to 2mm (4mm) in accumulation values for the 95th (99th respectively) percentile threshold on the front passing day (21 June). However the percentiles' thresholds follow the same overall trend by decreasing throughout the day. This suggests the assumption that the NWP models produce elongated precipitation features. The first peak of COSMO-2 for the 95th percentile at 5 UTC is because of strong predicted convective events ahead of the front. The xy-plot in Figure 8 visualises the precipitation field (a, b) and the corresponding 95th percentile thresholded field (c,d) at 5 UTC (front reach the Alps). It shows clearly that the threshold considers only areas of convective precipitation when the precipitation is strongest leading to a corresponding value with a very high peak for COSMO-2 in comparison to VERA. Have a further look at the 95th percentile thresholded fields ((c)&(d)), VERA compute a precipitation area over south Switzerland, whereas COSMO-2 does not. The structure of the latter field is at one stretch in contrast to the VERA analysis which has three separate rain areas.

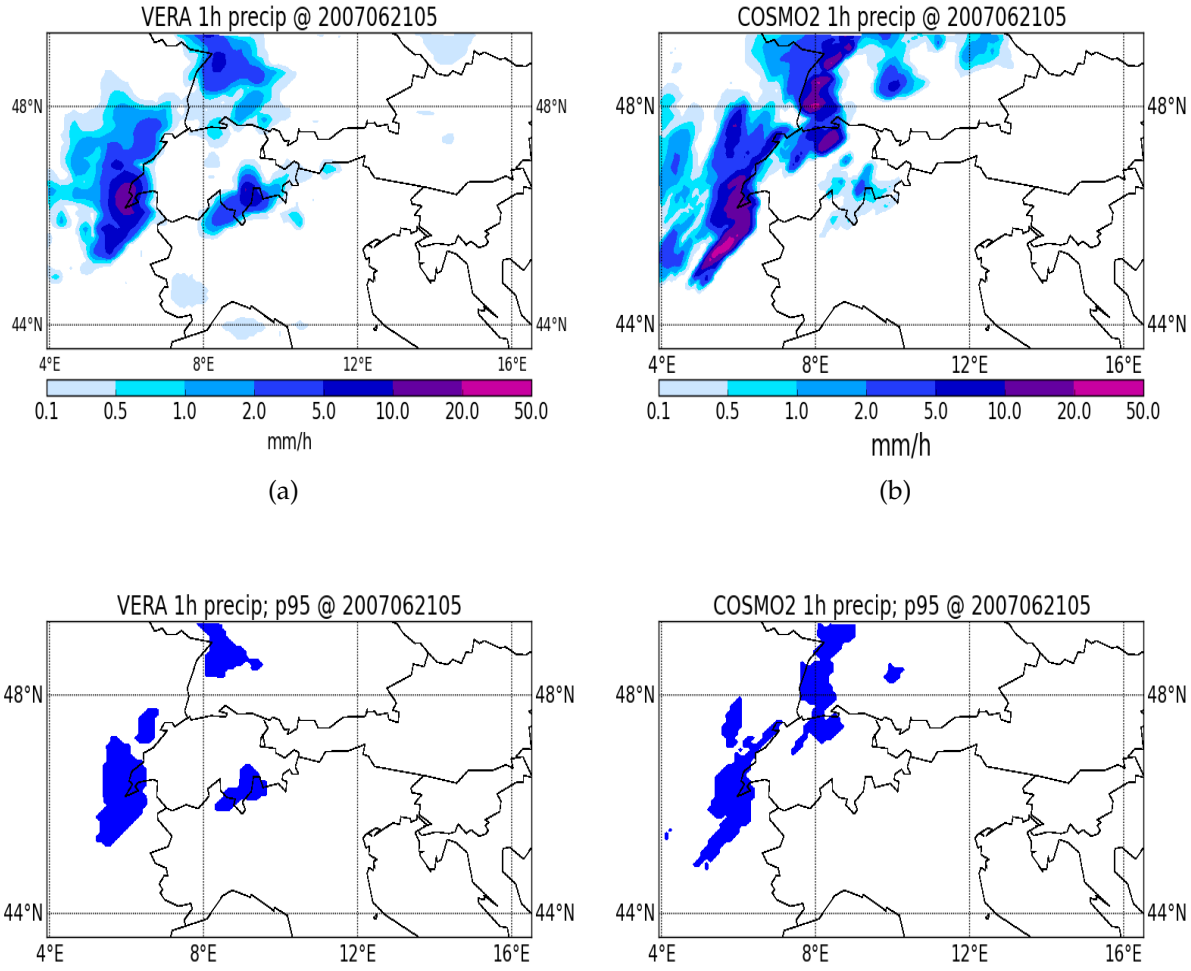


Figure 8: (a)&(b) Precipitation field in [mm/h] and (c)&(d) the 95th percentile threshold of VERA analysis (left) and COSMO-2 (right) on 21 June at 05 UTC.

The overall wind field is also well computed by the NWP for the 95th percentile threshold of wind strength as can be seen in Figure 9, which shows the analysis and model values of wind strength against time. The time series has a peak at 15 UTC with lower values before and afterwards and jibes with the peak of the 95th percentile of precipitation. The increase until 15 UTC is because of strong pre-frontal wind strength, after 5 UTC, when the front is inside the domain, and additionally an increase in wind strength behind the front. The front is leaving the domain after reaching the peak and that leads to a decrease of the time series. So there is indeed a correspondence, caused by the cold front, between the time series of the 95th percentile thresholded precipitation and wind values.

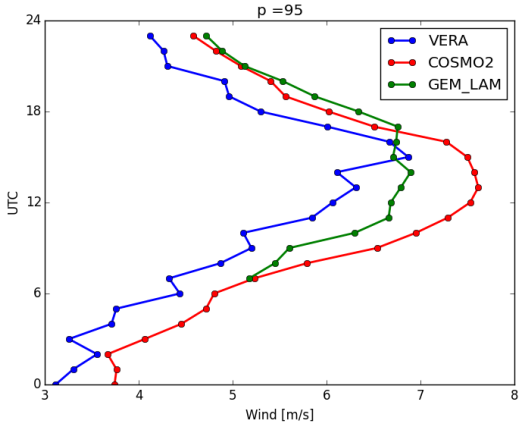


Figure 9: Wind strength [m/s] of VERA, COSMO-2 and GEM-LAM for 95th percentile threshold on 21 June 2007.

4.2 Spatial verification

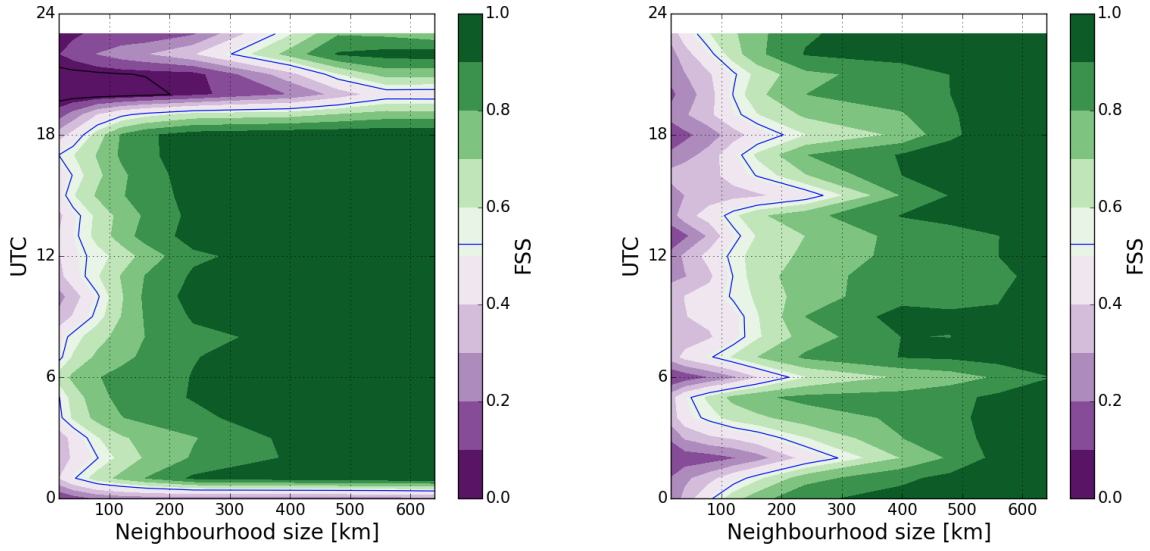
Now we consider the results of the different scores for the front passing day (21 June), because on this day, convective events occurred ahead of the front which transformed during the day from a scattered and heterogeneous rain area to a stretched and elongated structure (compare Figure 8 (a), (b) and Figure 5 (a), (b)). The comparison of the verification methods and discussion for all three days and the two NWP models follow in the next chapter.

4.2.1 Fractions Skill Score - FSS

The FSS result on 21 June for the 95th percentile threshold of precipitation (a) and wind strength (b) is shown in Figure 10. The x-axis corresponds to the neighbourhood size in km and the y-axis corresponds to time in UTC. For each time step (1h from 01 UTC to 23 UTC) and for selected space steps (1 grid point to 80 gridpoints; 1gp represents 8 km) the FSS was calculated. The blue line represents the believable scale at which a forecast is believable and the black line represents a random forecast.
Such plots are similar for all meteorological phenomena. At small scales it results in low values of FSS and increases for larger scales. The relatively high temporal variability of the believable scale is because of the uncertainty in the locations of the showers (and more uncertainty for wind strength maxima).
At 5 UTC the precipitation field of the 95th percentile of VERA and COSMO-2 (Figure 8 (c, d)) have slightly different structures but with the same local maxima in comparison to the original fields (Figure 8 (a, b)). East of Switzerland, the precipitation fields of the 95th percentile have a good match in space, but the structure is totally different. COSMO-2 shows also rain areas in south-western Germany stretched to northern Switzerland which are not present in the VERA analysis but results all in all into a relatively low believable scale (s_{fss}) of 16 km.
At 10 UTC, Figure 11 (a), (b) show the thresholded precipitation fields of VERA and COSMO-2 with relatively widespread rain areas but at different places with nearly no match. The s_{fss} increases up to 77 km, whereas s_{fss} is out of bounds at 21 UTC and this is because of nearly no forecast and observed pre-

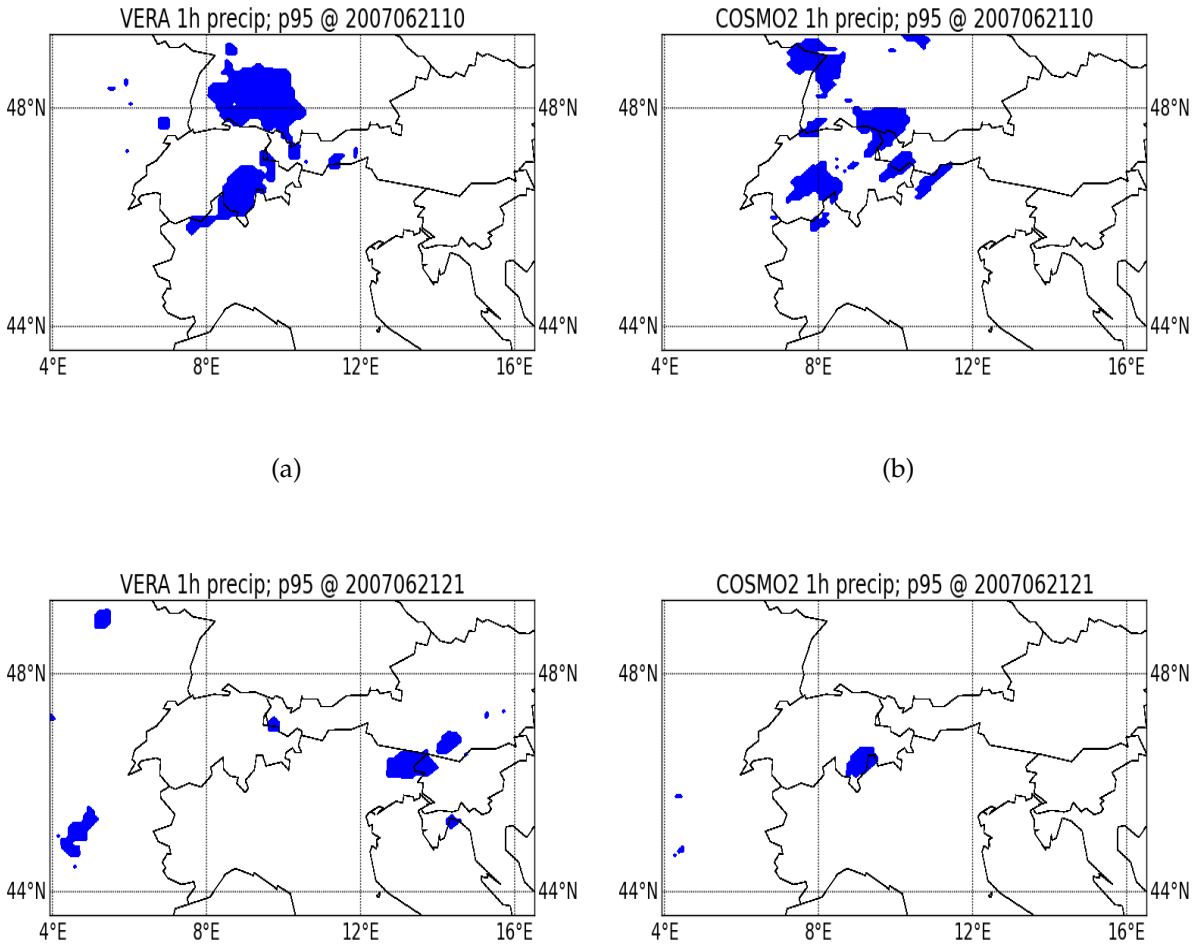
cipitation features and which are additionally distributed in space (Figure 11 (c), (d)).

So it is clear that bigger neighborhood sizes are needed to obtain a better alignment between the calculated fractions, what in turn results in higher values of FSS. The believable scale varies for the precipitation verification between 16 km and 90 km until 18 UTC. Than, the $FSS_{believable}$ increase out of 80 grid-points and this is because of just a few and localized areas, with low precipitation amounts, immediately after the front left the domain.



(a) (b)

Figure 10: The FSS for the 95th percentile threshold on 21 June 2007 for COSMO-2. Left the verification of precipitation and right of wind. The blue line represent the believable scale ($FSS_{believable} = 0.525$) and the black line at $f_0 = 0.05$ represent the scale for a random forecast.



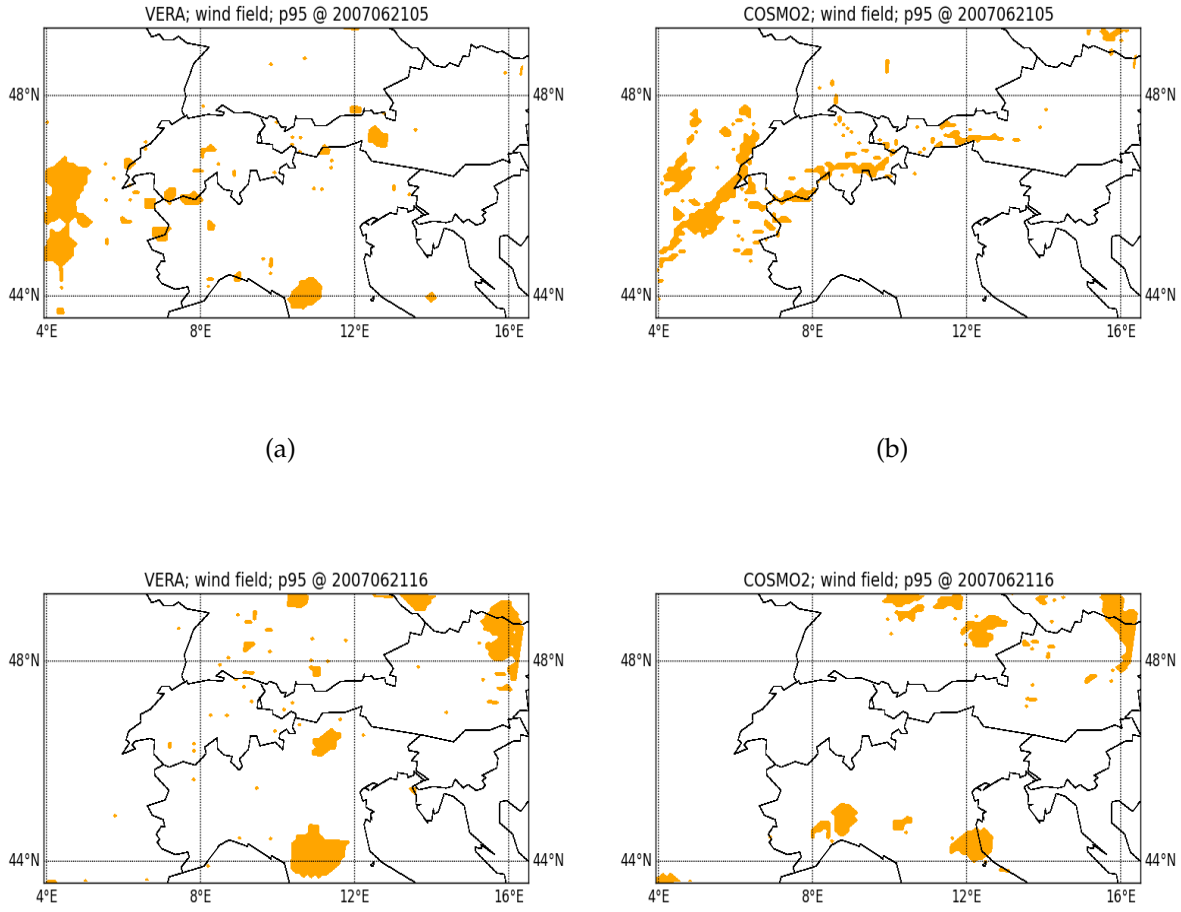
(a) (b)

(c) (d)

Figure 11: 95th percentile thresholded precipitation field of VERA (left) and COSMO-2 (right) at 10 UTC (a)&(b) and at 21 UTC (c)&(d) on 21 June 2007.

The believable scale of FSS for wind strength show a higher variation over time. This is because of the heterogeneous and spotty wind field of VERA and COSMO-2. The local maxima of wind strength are distributed over the whole domain and lead therefore all in all to higher values of s_{fss}. Of course, the wind field has always areas with values in such a big domain and so it is not

possible that the believable scale can get out of bounds (> 80 grid points).
Figure 12 illustrates this by the xy-plots of VERA (a&c) and COSMO-2 (b&d) wind field for the 95th percentile threshold at 05 UTC (upper panels) and 16 UTC (lower panels). The VERA analysis shows a high maximum behind the front and other maxima scattered all over the domain. The COSMO-2 maximum is stretched from west to east and is more patterned the one VERA yields. The s_{fss} has a minimum of 42 km at 5 UTC.
The 95th percentile thresholded wind fields at 16 UTC show a better agreement in the structure of the maxima but with relatively high local errors. For example, VERA and COSMO-2 show a maximum in the south of the domain over Italy but shifted. Additionally, VERA shows a widespread feature in northern Italy which was not forecast. This leads to a believable scale of 146 km.



(a) (b)

(c) (d)

Figure 12: 95th percentile thresholded wind field in [m/s] for VERA (left) and COSMO-2 (right) at 5 UTC (a)&(b) and 16 UTC (c)&(d).

4.2.2 Structure Amplitude Length - SAL

Wernli et al. (2008) show all three components of the verification of QPF by the SAL in one figure. This is interesting to evaluate longterm forecasts, but to verify a single forecast run it is better to show the variability of single components against forecast leadtime to analyse the quality over forecast lead time. Figure 13 shows this exemplarily for the 21 June component-by-component for precipitation. The verification of wind by SAL and DAS is shown and discussed in the next chapter.

Shown are the components for the 95th percentile threshold, enabling a fair comparison between the different verification methods and a threshold, suggested by Wernli et al., of $R^* = \frac{1}{15} R_{max}$. R_{max} is the maximum value of the VERA analysis of all grid points at the specific time step. To avoid confusion it should be mentioned that for computing SAL, the field is not transformed into a binary field to get a statement about structure, amplitude and location error. The single components in the three panels, computed for both thresholds, have a high variability over time but with the same overall trend throughout the day. The structure components (Figure 13 (a)) for the two different thresholds show deviations, especially until 12 UTC.

So the S-component is an indicator about the structure. As mentioned, the trend is the same, but there is also a relatively high spread until 12 UTC. This is explicable with the two different threshold values: At 12 UTC, $R^* = 1.07mm$ and the 95th percentile threshold is 2.48 mm and until that time step, R^* is always lower than the 95th percentile threshold. By using the latter threshold, the precipitation field of the forecast looks more peaked in comparison to the observation field than for R^* (compare 8 (c)&(d) with Figure 14 (a)&(b)). After 12 UTC the divergence between the two lines of the structure component drop down to zero. This is because of nearly equal values of the threshold and therefore similar precipitation and observation fields.

The A-component in Figure 13(b) shows a good match by comparing the amount of the observation- and forecast field for the thresholds R^* and the 95th percentile.

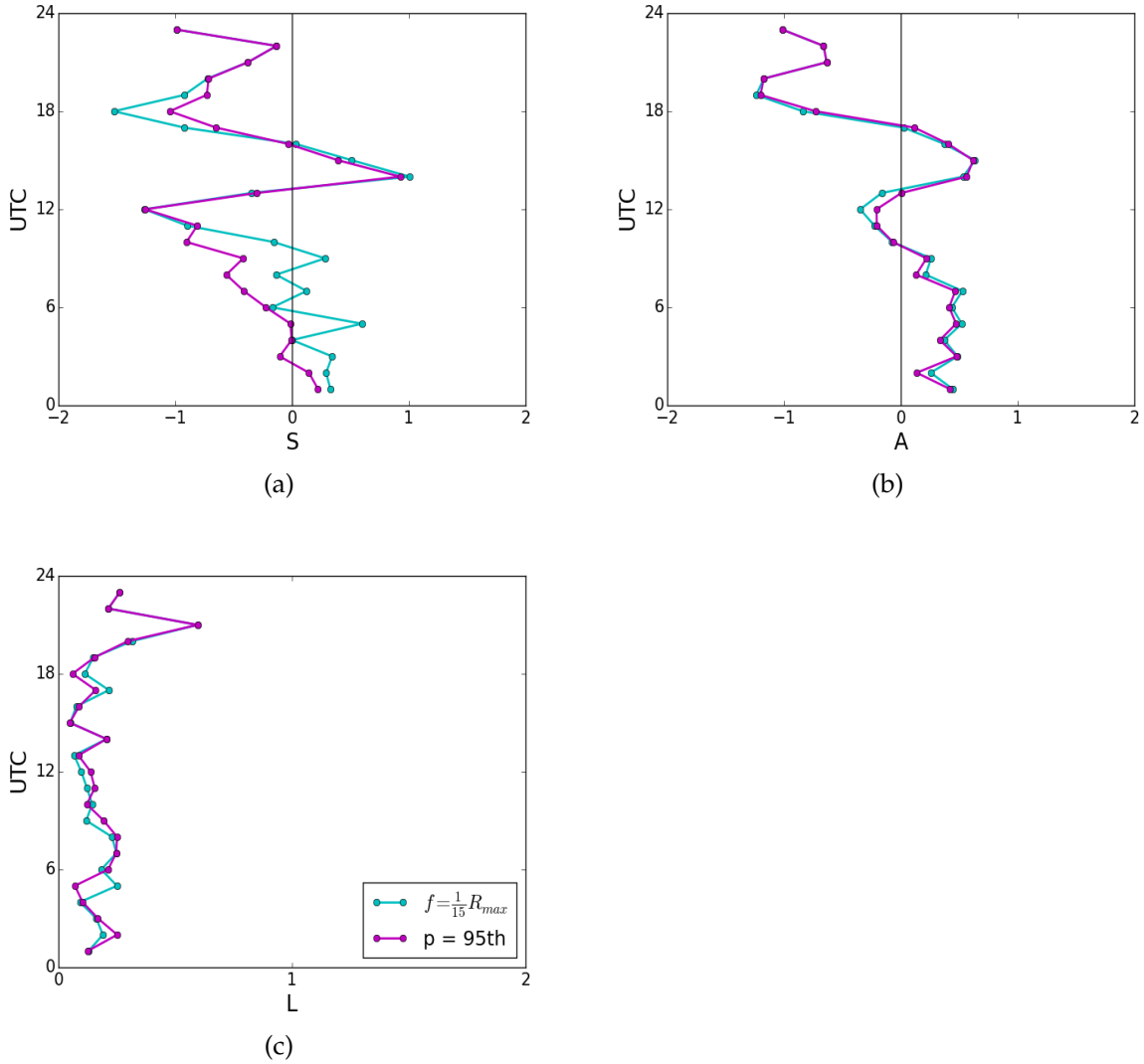


Figure 13: SAL components against time - (a) Structure, (b) Amplitude and (c) Length for 21 June.

The variation corresponds to the time series of the entire amount (Figure 6 (b)) and the 95th percentile threshold (Figure 7 (a)) and therefore gives no more information, but again underlines the importance to use no fixed thresholds. Until 9 UTC, COSMO-2 predicted an exaggerated pre-frontal precipitation amount in comparison to VERA and this leads to a positive A-component. The highest positive amplitude component is between 13 UTC and 17 UTC,

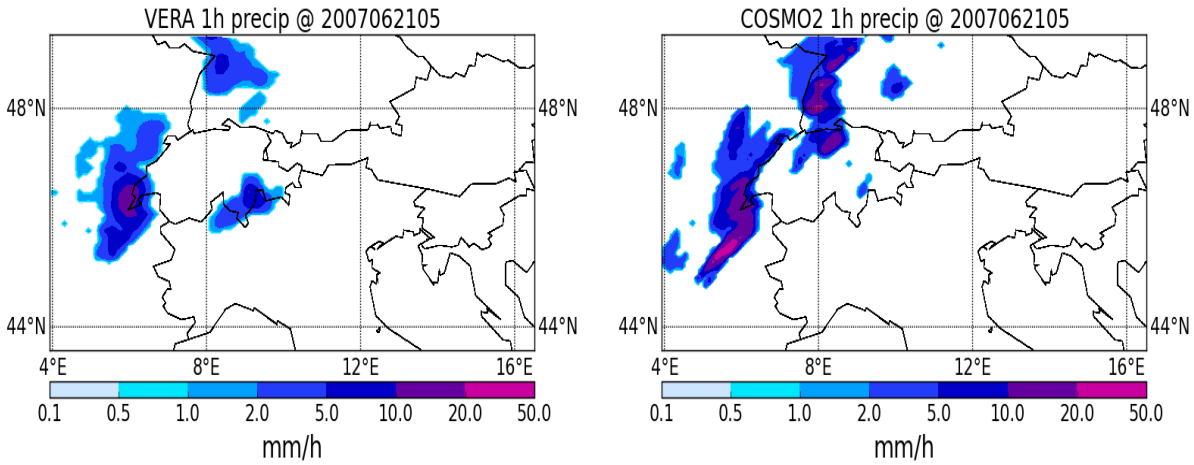


Figure 14: Observation field (VERA) (a) and forecast field (COSMO-2) (b) on 21 June after applying a threshold of R^* at 5 UTC.

because of the higher relative overestimation (about 3 times). From 17 UTC to 23 UTC, the A-component is lowest and reaches -1, caused by nearly no forecasted precipitation events and goes along with an underestimation by a factor of 3.

Figure 13(c) shows the L-component with relatively low values between 0.2 and 0.3 until 19 UTC and increase than constantly until 21 UTC. This is because of just a few rain areas with low rain rates and additionally high distributed in space. Call to mind that the L component consists of two parts. The first part is the difference between the computed center of mass of the forecasted and observed precipitation field, weighted with the largest distance between two boundary points. If rain areas are spread over the domain (with low rain rates) in both fields, a complete match of the computed center of mass is virtually impossible. As a consequence, the second part, which considers the averaged distance between the centre of mass and single precipitation features, can also get very high. This is the reason for the relatively high values after 19 UTC.

4.2.3 Displacement and Amplitude Score - DAS

Finally, this section gives the results of the DAS method (Figure 15), component-by-component like before for the SAL. To calculate DAS as the sum of $\frac{AMP}{I_0}$ and $\frac{DIS}{D_{max}}$, the 95th percentile was used as threshold and a subsampling factor of $F = 4$, so the coarse-grain pixel elements contain 16 by 16 grid points per pixel element ($2^4 = 16$). The subsampling factor of $F = 4$ was used, because then, the computed DAS demonstrate a good accordance to the other methods and has a high variability over time. The resulting value of the search distance D_{max} is 90.51 grid points (= 724 km). By using percentiles it is obvious that $n_{obs} = n_{fct}$, because the points with values smaller than the used percentile are set to zero and just the points with higher values (which have the same value as before thresholding; no binary field) are used to compute DAS. The number of (nonzero) points of the thresholded fields are equal and the total amplitude error is then defined as

$$AMP = \tfrac{1}{2}(\overline{AMP}_{obs} + \overline{AMP}_{fct}).$$

The characteristic intensities I_0 are calculated separately for each day as the mean RMS (root mean square per hour) amplitude of the 95th percentile of the observation field. Therefore the AMP component has been weighted by the characteristic $I_0 = 2.50$ mm on 21 June (1.75 mm on 20 June & 2.41 mm on 22 June). Anyhow, $\frac{AMP}{I_0}$ has up to 3 times higher values (due to the peaks) than the $\frac{DIS}{D_{max}}$ component and in consequence the amplitude component dominates DAS (Figure 15 (a)) in the overall trend.

For analysing it is necessary to have a look at both components separately.
The weighted amplitude error (Figure 15 (b)) shows relatively high values until 8 UTC, caused by relative high forecast and observed precipitation amount on the one hand and on the other hand by the spread in the time series of the 95th percentile (Figure 7 (a)). Then the component decline in a zigzag and drops nearly to zero at 22 UTC, not fully reaching it because of some observed and forecast precipitation features.
Figure 15 (c) shows the DIS component with a relatively high variability over time with two prominent peaks, first at 5 UTC and the second at 21 UTC. The

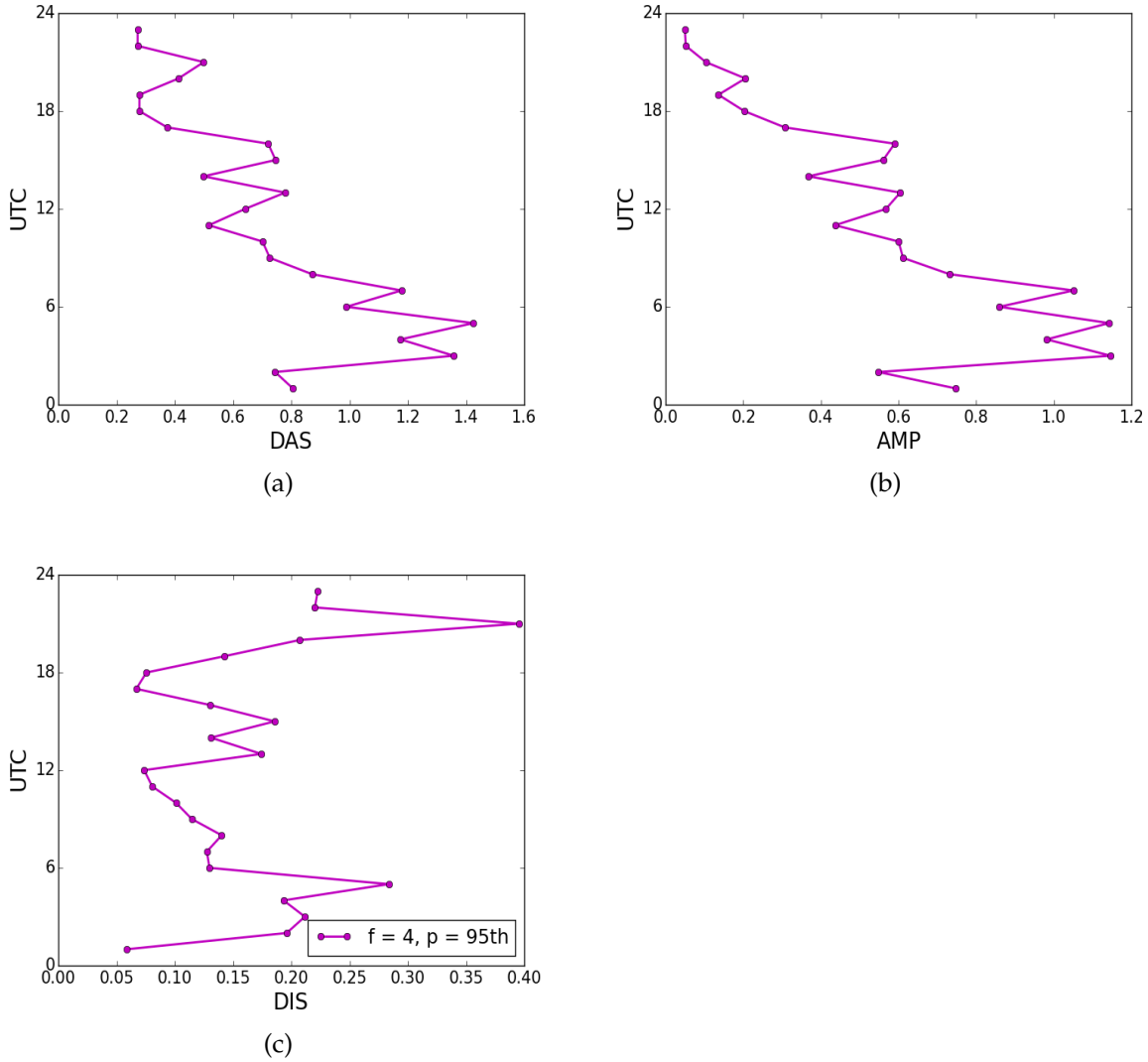


Figure 15: DAS components against time for precipitation- (a) DAS, (b) AMP and (c) DIS for 21 June.

reason for the latter peak at 21 UTC is explained already for FSS and SAL, but the peak at 5 UTC is surprising. So it is necessary to have a look on the sequence of different stages in the computation of DAS (Figure 16).

The observation is shown in Figure 16 (a) and the forecast in Figure 16 (b) (rotated through 90°, caused by the algorithm to compute the fields). It is nec-

essary to minimise the difference between observation in forecast space and forecast in observation space. This is shown in Figure 16 (c) by the morphed observation superimposed with the displacement vector field. The image-matching algorithm stretch the upper observed precipitation feature in the forecast space by a divergent vector field. The result is a good accordance to the stretched forecast. The single feature in the middle of the plot is removed and contributes decisively to the magnitude of the displacement vector field of the observation (Figure 16 (e)).
The morphed forecast in observation space (Figure 16 (d)) has – in contrast – a convergent displacement vector field around the upper, stretched precipitation feature. The contribution to the magnitude of DIS is small. In contrast, the lower feature vanish completely, resulting in high values of DIS in forecast space (Figure 16 (f)).
In summary, one can say that although the precipitation features of the observation and forecast mainly agree in location, the contribution of the structure to the displacement error is relatively high. That is why the DIS component at 5 UTC is high in comparison to s_{fss} and the L-component of SAL.

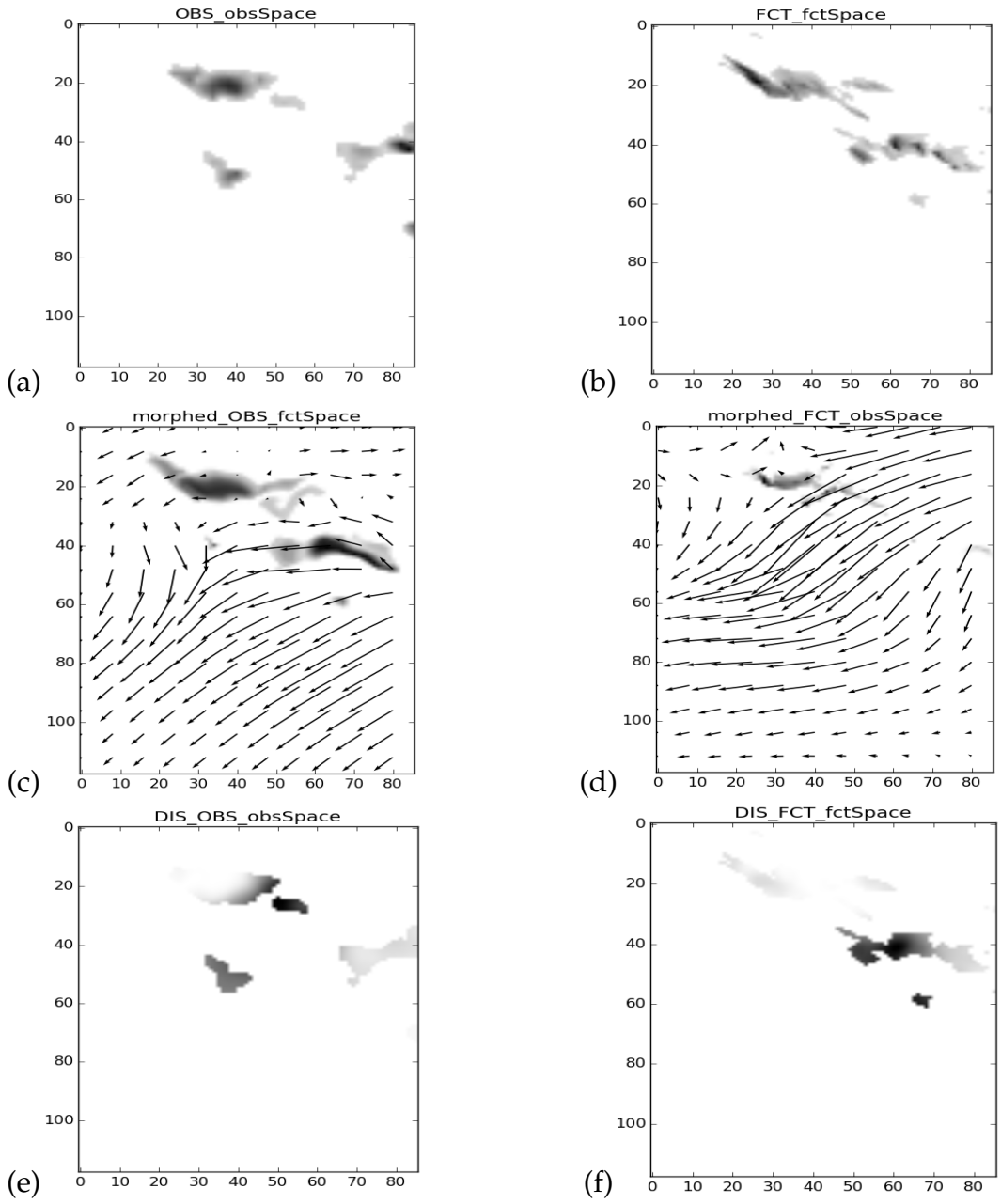


Figure 16: Sequence of different stages in the computation of DAS on 21 June at 5 UTC: observation in observation space (a), forecast in forecast space (b), morphed observation in forecast space, morphed forecast in observation space superimposed with displacement vector field (c,d), DIS in observation space (e) and DIS in forecast space (f).

5 Comparison and discussion

In the previous chapter, we have shown generic results obtained with the three verification methods in order to explain how the methods work in reality by applying them to precipitation and wind fields of the forecast and VERA analysis. The Fractions Skill Score (FSS) gives a good overview of the believable scale at which a forecast makes sense. The Structure Amplitude Length (SAL) provides important information about the structure of a precipitation (or wind) field compared to the forecast. In addition, the amplitude component yields whether a forecast is over- or underestimated. The L-component, as a last point, serves as a good indicator about the error in space. Amplitude (AMP/I_0) and displacement error (DIS/D_{max}) of DAS contain information about the quality of a forecast as well.
In general the three methods have in common that they make a statement about the location error of a forecast. The emphasis of this section is to investigate the goodness of the results by comparing location components of these three verification methods. This leads to the question of how the goodness can be determined.
It is established that the predictability increases with high synoptic forcing in contrast to low synoptic forcing, because then the small-scale processes dominate and convective precipitation becomes more unpredictable.
The investigated core case from 20 June 2007 to 22 June 2007 was dominated by a cold front moving eastward and caused elongated convective showers on 21 June (high synoptic forcing) leading to a homogeneous precipitation field. The results were shown previously for COSMO-2. On the day before/afterwards convective events were triggered by the uneven heating of the earth's surface caused by solar radiation. This led to a spotty and heterogeneous precipitation field. Over these three days the wind field was heterogeneous and spotty with local maxima distributed over the domain. This was tightened by the 95th percentile but the results are similar for lower thresholds (e.g. 70th percentile) as well and are not shown in this thesis.
The 95th percentile threshold was used (unless otherwise stated) to compute the following results.

5.1 Comparison of daily averages of COSMO-2 and GEM-LAM

GEM-LAM was the second NWP model investigated – after COSMO-2 – and the verification methods were also applied to the precipitation and wind fields measured during the three days in question. Comparing the daily means of the outcomes, namely the believable scale s_{fss}, L and DIS for the verification of precipitation of the two NWP models show a relatively consistent result (Table 1). Note that the value of s_{fss} for COSMO-2 and GEM-LAM is too low on 20 June, because between 9 UTC and 15 UTC, only some precipitation features were forecasted/observed and the FSS was set to zero. A comparison of the means is possible since just the outcomes before 9 UTC and after 15 UTC are used.

The mean values of the results of COSMO-2 are predominantly smaller in comparison to GEM-LAM on the 20 June with low synoptic forcing, as well as on 21 June with high synoptic forcing. The outcomes are approximately equal on 22 June, especially for <L>. <DIS> is slightly higher for COSMO-2 than for GEM-LAM and the believable scale for COSMO-2 is less by two grid points (16 km).

Table 1: Mean values of $FSS_{believable}$ scale(s_{fss}), L and DIS/D_{max} for the 95th percentile. Results are computed for COSMO-2 (C-2) and GEM-LAM (G-L) **precipitation field** for all three days.

	p95					
	$< s_{fss} >$ **/ km**		**<L> / km**		**<DIS> / km**	
Date	C-2	G-L	C-2	G-L	C-2	G-L
20 June	104	176	350	456	109	152
21 June	88	152	210	350	116	145
22 June	88	104	234	245	137	116

On 20 June, the values of <s_{fss}> (104 km, C-2; 176 km, G-L) and <L> (350 km, C-2; 456 km, G-L) are highest. This does not extend to the DIS component, which has values of 109 km for COSMO-2 and 152 km for GEM-LAM. The reason for the lower value of DIS, relative to the believable scale and the L-component, is the relatively low amount of observed precipitation features on that day. This leads to a low contribution to the magnitude of the displacement vector field of the observation and hence to a low displacement error. In addition, the values of the believable scale and the length component of SAL are approximately equal on 21 June and 22 June for COSMO-2. It is remarkable that it does not apply to GEM-LAM as well. In contrast the quality of GEM-LAM is worse on 21 June than on 22 June and the difference between the mean values on 21 June of the outcomes of COSMO-2 and GEM-LAM are highest. GEM-LAM is unable to capture the front accurately as can be seen in Figure 17. Shown is the x-y plot of precipitation for GEM-LAM at 16 UTC as an example. The comparing of Figure 5 (a & b) for VERA and COSMO-2 with Figure 17 shows clear that most of the cold front in the precipitation field of GEM-LAM left the domain.

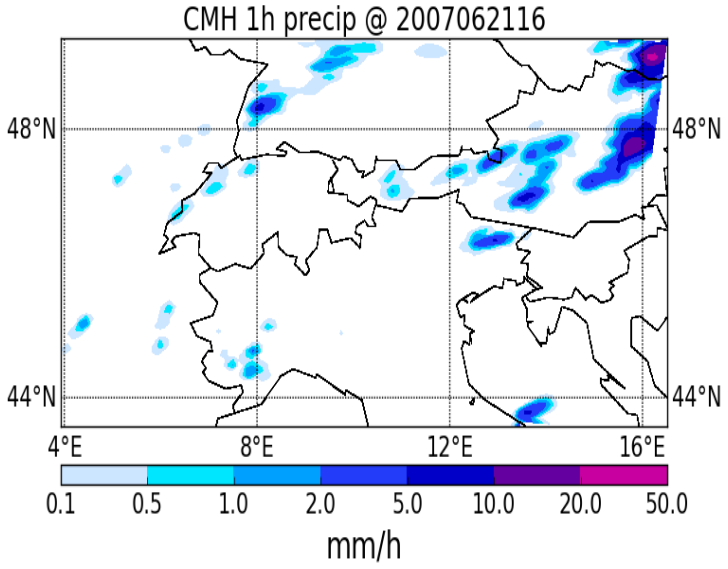


Figure 17: Precipitation field in [mm/h] of GEM-LAM on 21 June at 16 UTC.

The results for wind verification are not shown, but in summary the mean values are higher, except for DIS but are much more variable throughout the days and NWP models.
One reason for this behaviour is the heterogeneous wind field with maxima distributed over the domain. That implies that for calculating the FSS, the grid-box has to become bigger to reach the believable scale. The L-component also increases because of the great distance between the computed centre of mass of the wind field and the single features (wind maxima). The DIS component depends on the distance between the morphed forecast on observation space and the observation (and vice versa), but it could happen that the features are out of the search distance and then the DIS component is low. This is a good opportunity to verify a big domain and to conclude features indirectly. Nevertheless the comparability of the believable scale and the L-component with DIS is limited. Additionally, the magnitude of the displacement error field is lower, by morphing spotty features of forecast and observation fields (the length of the vectors is smaller).
Using the 99th percentile threshold leads to similar results for precipitation and wind, but also to higher averages. When using the 70th percentile threshold for wind verification the values decrease negligibly.

5.2 Comparison of hourly values of precipitation and wind strength

This section focuses on the precipitation verification of COSMO-2 on the 21 June 2007. The discussion of the daily averages has shown that in average COSMO-2 has a better quality than GEM-LAM. However this is not a statement about the variability during the course of the day: investigations have shown that comparability of the three methods is limited. This is because of relatively fluctuating results of the location components and possible reasons are already discussed in chapter 5.1.
The results for 95th percentile of precipitation is shown in Figure 18. The plot contains the time series for the believable scale s_{fss} (green), the L-component

of SAL (magenta) and DIS/D_{max} of DAS (blue). All these are weighted with their respective daily maximum value.

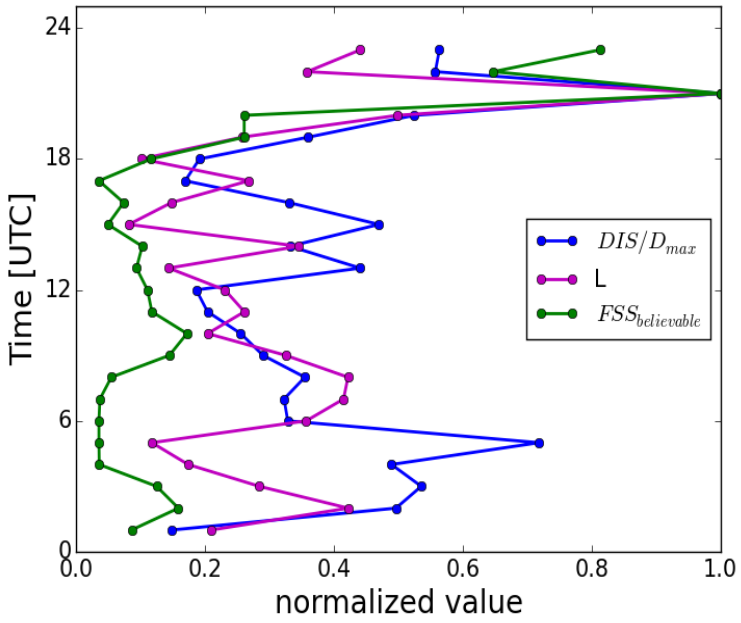


Figure 18: Results for COSMO-2 (normalized): s_{fss}, L- and DIS- component for the 95th percentile of precipitation on 21 June.

The L-component and DIS/D_{max} show a relative high variability over the day in contrast to s_{fss} which is smoother. All three time series have a peak and their maximum normalized value at 21 UTC, which is the time step after the front vanished completely out of the domain. COSMO-2 is obviously not able to resolve this process. After that time step, the values first decrease and then increase at the last time step. It is also remarkable that from 18 UTC on all three components follow the same trend by increasing until 21 UTC, because after 18 UTC the front is leaving the domain.
The DIS component also shows a peak at 5 UTC whereas the L-component has a local minimum which has previously been discussed. In addition, it can be said that the normalized values of DIS and L are higher than the normalized values of the believable scale. Thus the obvious question is the quality of the accordance between the components in their variation over the time.

5.3 Correlation between the components

The Pearson correlations between the values of the location components were calculated for each time step for the 95th and 99th percentile thresholded precipitation and wind fields of the two NWP models.
The results are shown in Table 2 for COSMO-2, compared to VERA and for the 95th percentile thresholded precipitation fields for the 21 June. Between all three components exists a positive correlation, best between the believable scale and the L-component with a coefficient of 0.72. So all three methods lead to similar statements about the quality of the location error for a deterministic forecast on day with high synoptic forcing. The Pearson correlation coefficient has similar, but lower values for the 99th percentile thresholded precipitation field.

Table 2: Pearson correlation coefficient between location components for the 21 June. Shown is the 95th percentile. For all values, $p < 0.05$. The matrix is symmetric with ones on the main diagonal. To enhance readability only values above the diagonal are given.

21 June

p95	s_{fss}	**DIS**	**L**
s_{fss}	1	**0.65**	**0.72**
DIS		1	**0.58**
L			1

In contrast GEM-LAM (not shown) has a positive correlation between the DIS component and the believable scale only on 21 June, using the 95th percentile of precipitation fields. Furthermore, the calculation of the correlation coefficient of the wind strength forecast by the NWP models and VERA show almost never a positive correlation between the components.
In summary, it can be said that in general, no correlation exists for precipitation between location errors of the verification methods in situations with low synoptic forcing on the one hand and wind verification for the investigated core

case on the other hand. But the positive result is that the location components can be compared on days with high synoptic forcing and they therefore yield identical information.

6 Summary and outlook

In this thesis, the comparability, quality and consistency of three verification methods were investigated, especially the believable scale s_{fss} of the Fractions Skill Score, the location component L of SAL and DIS of the DAS method. The methods were applied to the VERA analysis compared to COSMO-2 and GEM-LAM, all with a resolution of 8 km. The verified parameters are precipitation (1h accumulated) and wind strength. The 95th and 99th percentile thresholds were used, as a consequence of a high bias and to allow an investigation of the spatial distribution of phenomena. These high percentiles are associated with smaller, more extreme features.

The verification of precipitation shows that the NWP models compute the fields relatively accurately. The mean of the believable scale varies between 88 km and 104 km for COSMO-2 and between 104 km and 176 km for GEM-LAM for the 95th percentile. Interestingly, the results for 22 June are nearly as good as the front passing day (21 June). All methods in use assess the 20 June as the day with the highest values in location errors, except of <DIS>. The wind strength verification shows similar results, but the $<s_{fss}>$ is higher. Looking at the 99th percentile, the believable scale clearly increases for both NWP models and for all three days, in contrast to <L>, and $<DIS/D_{max}>$. The behaviour of L can be explained by a high weighting of the L_2 component by high precipitation/wind strength values.
The reasons for these discrepancies are that, on the one hand, the displacement error field is lower for spotty fields which get morphed as the displacement vectors have a lower magnitude. On the other hand, forecast and observation fields with nearly no precipitation barely contribute to the displacement error.

This leads to nearly no positive correlation between the components on days with low precipitation amounts. A further reason could be that the search distance is too small.

All in all, the FSS, SAL and DAS are good instruments to evaluate model errors, which clearly exist on days with low synoptic forcing, represented in this thesis by the high means of the location components on 20 June. The verification of wind (and precipitation) cause difficulties because of the heterogeneous field with high spatial distribution of the maxima. The direction of the wind field was not investigated. Hence it seems to be necessary to develop new verification methods, where both the strength and the direction matter.
Another method to improve the comparability of the verification methods is to split the domain into smaller areas of interest and to apply the methods in a second step. It could also be useful to develop a method which is spatially varying and following areas with high values of precipitation / wind strength (for example around a cold front).

As an outlook, it is of great interest to use ensemble forecasts and compare them with ensembles of observations to investigate both the error in observation and the error of the forecast. Therefore, datasets of COSMO-LEPS and VERA ensembles are available in MESOVict.

References

[1] Baldauf, M. A. Seifert, J. Förstner, D. Majewski, and M. Raschendorfer, 2011: Operational Convective-Scale Numerical Weather Prediction with the COSMO Model: Description and Sensitivities. *Mon. Wea. Rev., 139, 3887-3905.*

[2] Dey, S.R.A., G. Leoncini, N.M. Roberts, R.S. Plant, S. Migliorini, 2014: *A Spatial View of Ensemble Spread in Convection Permitting Ensembles.* Mon. Wea. Rev., 142, 4091–4107.

[3] Dorninger, M., M.P. Mittermaier, E. Gilleland, E.E. Ebert, B.G. Brown, and L.J. Wilson, 2013: *MesoVICT: Mesoscale Verification Inter-Comparison over Complex Terrain.* NCAR Technical Note NCAR/TN-505+STR, 23 pp, DOI: 10.5065/D6416V21.

[4] Dorninger, M., T., Gorgas, 2013: *Comparison of NWP-model chains by using novel verification methods.* Meteor. Zeitschr. Vol. 22, No. 4 (2013), p. 373 - 393, doi: 10.1127/0941-2948/2013/0488

[5] Erfani, A., J. Mailhot, S. Gravel, M. Desgagné, P. King, D. Sills, N. McLennan, and D. Jacob, 2005: The high resolution limited area version of the Global Environmental Multiscale model and its potential operational applications. *11th converence on mesoscale processes, American Meteorological Society, Albuquerque, NM, USA.*

[6] Gilleland, E., D. Ahijevych, B. G. Brown, B. Casati, and E. E. Ebert, 2009: *Intercomparison of Spatial Forecast Verification Methods.* Wea. Forecasting, 24, 1416–1430.

[7] Gorgas, T., M., Dorninger, 2012: *Concepts of pattern-oriented analysis ensemble based on observational uncertainties*, Q. J. R. Meteorol. Soc. Vol. 138, 769-784, April 2012 A, DOI:10.1002/qj.949

[8] Keil, C., G. C. Craig, 2007: *A displacement-based error measure applied in a regional ensemble forecasting system.* Mon. Wea. Rev., 135, 3248 - 3259

[9] Keil, C., G.C. Craig, 2009: *A displacement and amplitude score employing an optical flow technique.*, Spec. Coll. Spatial Forecast Verification Methods, 1297 - 1308

[10] Lorenz, E. N., 1969: *The predictability of a flow which possesses many scales of motion.* Tellus, 21A, 289-307, doi:10.1111/j.2153-3490.1969.tb00444.x.

[11] Murphy A.H., 1993: *What Is a Good Forecast? An Essay on the Nature of Goodness in Weather Forecasting.* Wea. Forecasting, 8, 281–293.

[12] Wernli, H., M. Paulat, M. Hagen, C. Frei, 2008: *SAL-A Novel Quality Measure for the Verification of Quantitative Precipitation Forecasts.* Mon. Wea. Rev., 136, 4470-4487.

[13] Roberts, N.M., H.W., Lean, 2008: *Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events.* Mon. Wea. Rev., 136, 78-97, doi:10.1175/2007MWR2123.1

[14] Roberts, N.M., 2008: *Assessing the spatial and temporal variation in the skill of precipitation forecasts from an NWP model.* Meteor. Appl., 15, 163-169, doi:10.1002/met.57.

[15] Rombough, H., H. Greene, B. Niska-Aro, B. Power, D. Schmidt, O. Stachowiak, C. Wielki, and A. Yun, 2010: GEM-LAM convective forecasts: How can they be used in an operational forecast environment?. *25th Converence on Severe Local Storms*

[16] Rossa, A. F. Del Guerra, and Daniel Leuenberger, 2009: An empirical radar data quality function. *Proceeding of the AMS Radar Converence. VA: Williamsburg.*

[17] Weisman, M.L., C.A. Davis, W. Wang, K.W. Manning, and J.B. Klemp, 2008: *Experiences with 0 - 36-h explicit convective forecasts with the WRF-ARW model.*, Weather and Forecasting, 23, 407-437, DOI: 10.1175/2007WAF2007005.1.

[18] Weusthoff, T., F. Ament, M. Arpagaus, and M. Rotach, 2010: Assessing the benefits of convection-permitting models by neighborhood verification: Examples from MAP D-PHASE. *Monthly Weather Review*, 138(9), 3418-3433.

[19] Wilks, D., 2011: *Statistical Methods in the Atmospheric Sciences*, 3rd ed. Academic Press, 676 pp.

Acknowledgment

Mein besonderer Dank gilt Dr. Christian Keil für die gute, lehrreiche und zugleich geduldige Betreuung.
Sehr herzlich bedanke ich mich auch bei Dr. Manfred Dorninger, der mir anfangs mit Rat und Tat zur Seite stand und mir die richtige Umgebung für das Thema geschaffen hat.
Außerdem bedanke ich mich bei Felix Liebrich für seine unermüdliche Ausdauer beim Korrekturlesen.
Mein letzter Dank gebührt natürlich meinen Bürogenoss*innen, die öfter mal für die nötige Ablenkung gesorgt haben.