Consortium

for

Small-Scale Modelling

Technical Report No. 16

COSMO Priority Project "INTERP": Final Report

December 2009

DOI: 10.5676/DWD_pub/nwv/cosmo-tr_16

Deutscher Wetterdienst

MeteoSwiss

ΕΘΝΙΚΗ ΜΕΤΕΩΡΟΛΟΓΙΚΗ ΥΠΗΡΕΣΙΑ

Ufficio Generale Spazio Aereo e Meteorologia

Instytucie Meteorogii i Gospodarki Wodnej

Administratia Nationala de Meteorologie

Agenzia Regionale per la Protezione Ambientale del Piemonte

Agenzia Regionale per la Protezione Ambientale dell' Emilia-Romagna: Servizio Idro-Meteo-Clima

Centro Italiano Ricerche Aerospaziali

Amt für GeoInformationswesen der Bundeswehr

www.cosmo-model.org

Editor: Massimo Milelli, ARPA Piemonte
Printed at Deutscher Wetterdienst, P.O. Box 100465, 63004 Offenbach am Main

COSMO Priority Project "INTERP": Final Report

Pierre Eckert

Pierre Eckert
MeteoSwiss
Geneva
Switzerland

Contents

Project report

1 Summary

The project had the goal to explore various ways to verify very high resolution models taking into account the fact that this kind of model are not to be looked at one gridpoint, at least not for precipitation, on which the project finally focussed on. We explored methods which allow building statistics out of the distribution of values surrounding a gridpoint. These methods are commonly called fuzzy verification. Applying the various methods onto idealized precipitation patterns, upscaling and the fraction skill score were kept for their adequate properties. This is shown in Section 3. Verifications have then been carried out on different regions and seasons with the Swiss COSMO-2 (2.2 km) and the German COSMO-DE (2.8 km). It turns out that the useful scales of these models are situated between 30 and 50 km. This value can even be higher depending on the considered precipitation threshold. The verification scores of very high resolution models are however better than the corresponding high resolution models COSMO-7 and COSMO-EU which have a grid spacing of $\approx$ 7 km. Sections 3 and 4 show details on these facts. Optimal verification methods should lead to corresponding products. Upscaling is quite commonly used by building averages of precipitation around gridpoints or over hydrological catchments. A product linked to the fraction skill score is provided by the neighborhood method in which probabilities are built out of the distribution of precipitation included in a spatial and temporal neighborhood of an observation point. Optimization of the method and examples are shown in Section 5. It is also interesting to compare the distribution functions of both the observed and the modelled precipitation contained in boxes of 100 km x 100 km. Global models are clearly unable to reproduce the tails of the distributions while high resolution models behave better although not perfectly. This leaves possibilities to use higher moments of the distribution function like for instance some quantiles, or to calibrate the distribution as done with EPS. This work is exposed in Section 6.

2 General introduction

The project had the goal to investigate advanced postprocessing methods, including hydrological applications, but it soon appeared that the most challenging postprocessing was the one consisting of extracting pertinent information from the very variable fields provided by very high resolution models (1-3 km), in particular for precipitation. This report will thus target on the scale dependency of the precipitation verification. In very high resolution models (1-3 km), the precipitation fields show usually a lot of variability. Although not positioned at the right position at the right time, the structures are often more realistic than those given by high resolution models (5-10 km). Typical images are shown in Fig. 1.

Clearly, the forecasts that can be generated by these two forecasts will differ, the probability of exceeding some small threshold being smaller in the 2 km case. Even if one can argue that the 2 km forecast is more useful than the 7 km forecast, the classical verification scores will be smaller for the very high resolution model when looking at the grid scale (this is often referred to as the double penalty effect). Therefore, methods have been developed for extracting the information out of boxes of growing sizes. These so called fuzzy verification methods are described in various papers (non exhaustive list): Contiguous Rain Area (Ebert and McBride, 2000), Neighborhood method (Theis et al. 2005), Fraction skill score (Roberts 2005), Intensity scale technique (Casati et al. 2004, Mittermaier 2006). Other references can

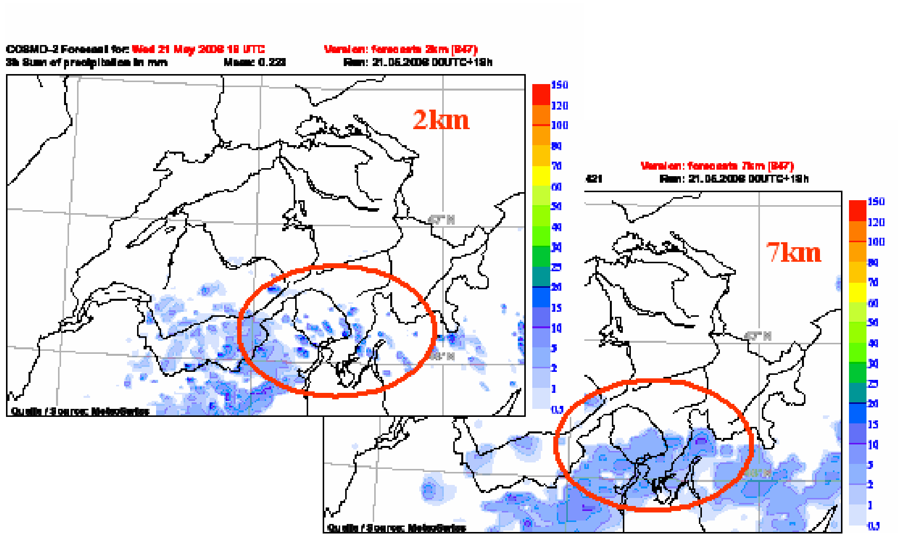

Figure 1: Example of precipitation field.

be found in Section 9. The generic behavior of such verifications has been sketched by N. Roberts (Fig. 2).

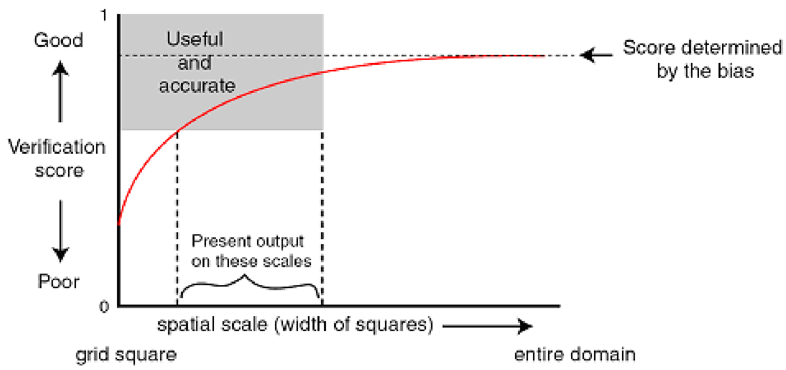

Figure 2: Generic behavior of verification.

The goal of the present project was to find out the highlighted "useful and accurate" scale. Of course, the other interesting question is to know if at this scale the very high resolution models show better scores than high resolution models. The scale dependency has been looked at from four different points of view which are reproduced in the following chapters:

- Fuzzy verification: theoretical aspects and application during the MAP D-PHASE DOP, by T. Bähler, F. Ament, T. Weusthoff
- Verification of DWD-models against high resolution precipitation and SYNOP measurements, by U. Damrath
- Post-processing of COSMO-7 and COSMO-2 Precipitation with the Neighborhood Method, by P. Kaufmann
- Statistical properties of precipitation in boxes of various sizes, by M. S. Tesini, C. Cacciamani, T. Paccagnella.

3 Fuzzy verification: theoretical aspects and application during the MAP D-PHASE DOP (T. Weusthoff, T. Bähler, F. Ament, MeteoSwiss)

3.1 Introduction

High resolution numerical weather forecast models are often punished when evaluated by traditional scores, which are based on a point wise comparison of observation and forecast. Although the structure of the observation may seem well forecasted by a high resolution model, a standard score may reveal a bad score due to e.g. spatial discrepancies. The coarser resolution model which cannot resolve the small-scale structure may in such a case get better scores although the subjective impression tells us different. However, so-called fuzzy verification methods try to avoid that problem by evaluating the forecast within a window around the region of interest. The forecast is hereby evaluated on increasing spatial scales by moving window domains of various sizes over the verification domain. The result is then presented as a function of spatial scale and threshold (see Fig. 3).

A variety of methods proposed by different authors can be allocated to "fuzzy verification", like those listed in Tab. 1. They are compiled by E.E. Ebert in form of a fuzzy verification toolbox - the IDL code can be downloaded from her webpage:
http://www.bom.gov.au/bmrc/wefor/staff/eee/beth_ebert.htm.

The different verification methods are in a first instance evaluated by means of idealized test cases. This way, three methods are identified to be well suited for the purpose of high resolution precipitation forecast. Those methods, namely the Fraction Skill Score, Upscaling and Intensity Scale, are then employed for the verification of precipitation forecasts of the Swiss COSMO models during the D-PHASE operations period.

Table 1: List of methods from the fuzzy verification toolbox (Ebert, 2008) which were evaluated by means of the test bed approach.

Acronym	**Method**	**Score**	**Decision model for useful forecast**
UP	Upscaling (Zepeda-Arce et al., 2000; Weygandt et al. 2004)	ETS	Resembles observation when averaged to coarser scales.
YN, MC	Anywhere in Window, 50% Coverage (Damrath, 2004)	ETS	Predicts event over minimum fraction of region.
FZ	Fuzzy logic (Damrath, 2004)	ETS	More correct than incorrect.
JP	Joint probability (Ebert, 2002)	ETS	More correct than incorrect.
ME	Multi-event contingency table (Atger, 2001)	HK	Predicts at least one event close to observed event.
IS	Intensity scale (Casati et al., 2004)	SS	Lower error than random arrangement of observations.
FB	Fractions skill score (Roberts and Lean, 2007)	FSS	Similar frequency of forecast and observed events.
PG	Pragmatic approach (Theis et al., 2005)	BS	Resembles forecast based on perfect knowledge of observations.
PP	Practically perfect hindcast (Brooks et al., 1998)	ETS ratio	Can distinguish events and non-events.
CS	Conditional square root of RPS (Germann and Zawadzki, 2004)	CSRR	High probability of matching observed value.
RM	Area-related RMSE (Rezacova et al., 2007)	RMSE	Similar intensity distribution as observed.

3.2 Test bed

3.2.1 Set-Up

The fuzzy verification package of E.E. Ebert contains a large number of methods and scores which are applicable for different purposes. In order to determine which of the available verification methods are suitable for verification of high resolution precipitation forecasts, the methods were tested in an idealized case study. A test bed was set up as follows (Fig. 3): a synthetically generated field that reflects a precipitation field containing simplified convective rain cells is modified by a perturbation generator which induces several known and assessable errors on that field. For each perturbation the resulting field was verified against the initial synthetic field by means of the fuzzy verification toolbox. In addition to that, the initial field was verified against itself, which represents a perfect forecast and should result in perfect scores. The observed response to the perturbations is then compared to the expected behavior. The sensitivity of the different methods and their cross-redundancy was determined. Working with a synthetic field - instead of radar data - provides the advantage of a simplified field with easier appreciation of the induced errors. However, the whole procedure was repeated with real radar data as will be presented in an upcoming publication. The results for the idealized cases are presented in the next section.

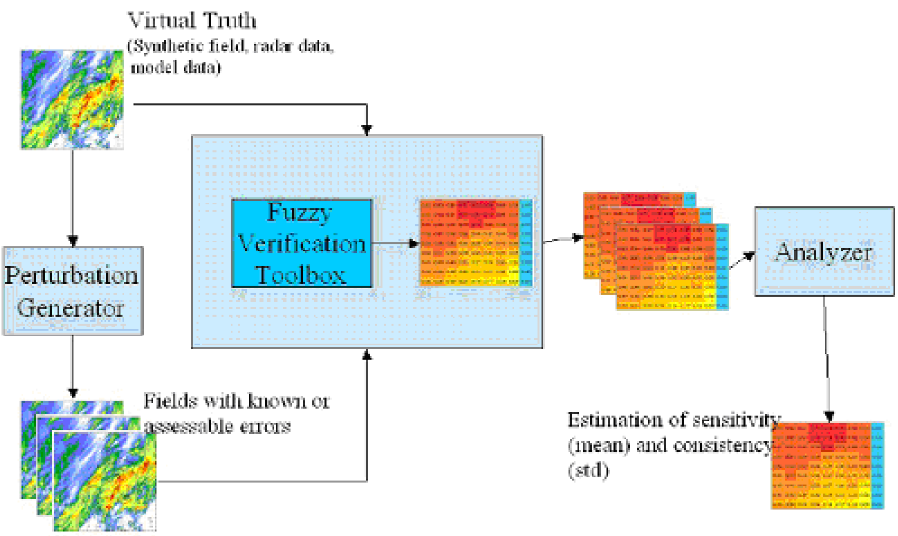

Figure 3: Test bed setup for the evaluation of fuzzy verification methods.

3.2.2 Perturbations and expected response

The synthetic field was perturbed by different perturbations for which a specific response was expected and additionally verified against itself ("perfect forecast"). Five types of perturbations are applied:

- XSHIFT - horizontal translation (10 gridpoints)
- BROWNIAN - no small scale

- LS_NOISE - wrong large scale forcing
- SMOOTH - high horizontal diffusion
- DRIZZLE - overestimation of low intensity precipitation

The appearance of the perturbations to the idealized precipitation fields is displayed in Fig. 4, while the expected response to each perturbation is shown in Fig. 5. Here, red boxes indicate scales where any sensitivity to the respective perturbation is expected, while blue boxes denote the scales for which no response is assumed.

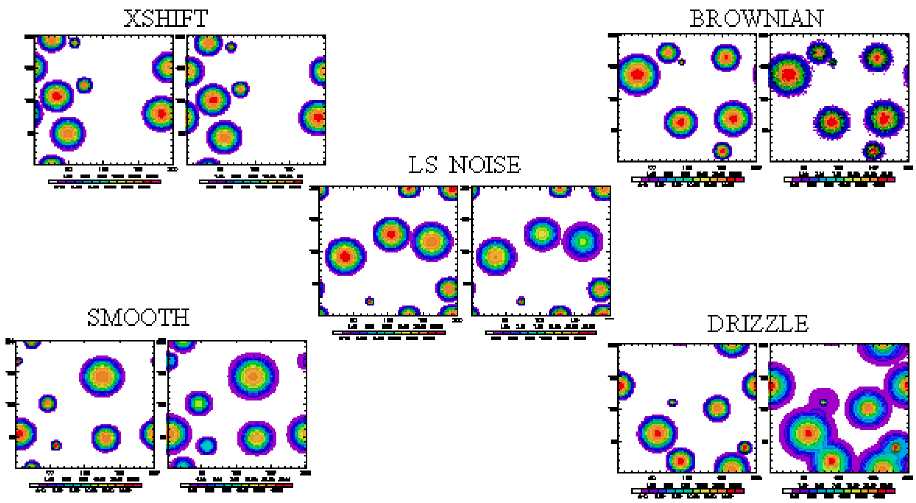

Figure 4: Perturbations on the synthetically precipitation field.

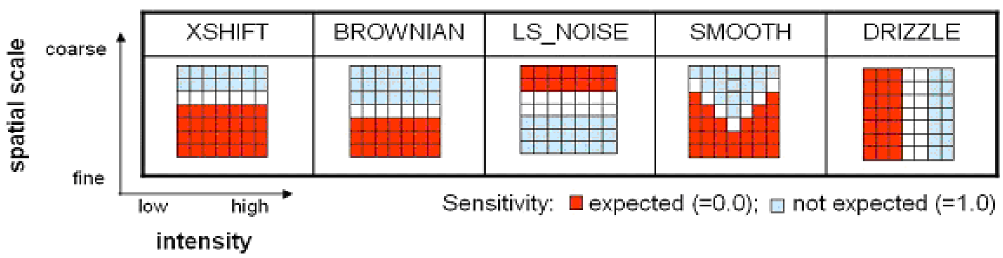

Figure 5: Expected response to the perturbations. Red boxes indicate that a response is expected on these scales why blue boxes indicate scales where no sensitivity to the perturbation is expected. When applying XSHIFT, for example, the verification results are expected to be sensitive for small scales and all thresholds.

3.2.3 Results of Sensitivity Studies

The sensitivity of the results is expressed by means of the difference between the expected behavior and the actual response to the perturbations. As measure for this difference, the contrast is calculated. It is defined as the mean value of the fuzzy boxes where no response was expected minus the mean value over all boxes where a response was expected:

contrast = mean(not_expected) - mean(expected)

The more the actual response equals the expected one, the higher the value of the contrast. The results of the idealized cases in terms of contrast are given in Fig. 6. Some methods were found to miss a perfect forecast, i.e. when the initial field was verified against itself, the scores did not give the perfect value, which is usually 1. Those "leaking scores" were not further considered (FZ, JP, ME, PG, CS). For the five perturbations, the scores revealed very different responses. Some are fairly good for all perturbations (e.g. UP and PP), while others are especially good for a special kind of perturbation but fail to detect others (e.g. YN and JP).

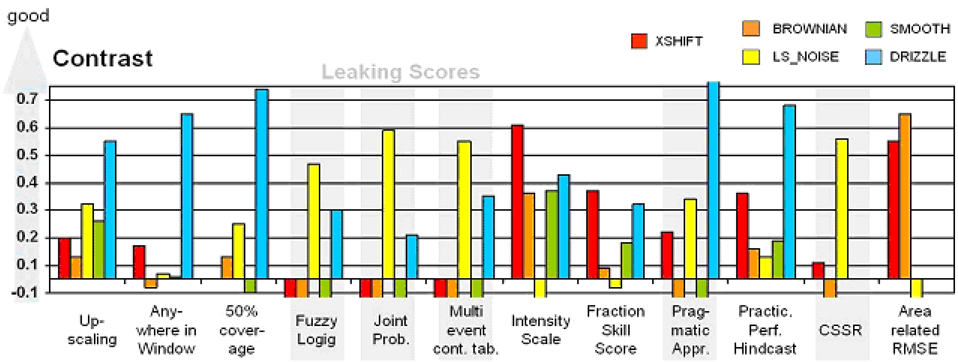

Figure 6: Results of the test bed experiment for the five perturbations on the idealized precipitation field.

Finally, the redundancy of the different methods was determined by calculating the correlation of the scores resulting from the test bed experiments. The scores for all window sizes and thresholds are therefore averaged over all perturbations and the correlation between the different scores is determined (Fig. 7). Three main groups of redundant methods have been identified this way: the first contains the methods UP, YN, MC, FB and PP, another group includes FZ, JP and a third one FB, PP. In addition it should be mentioned that the reliability is good for all scores - they show low standard deviations. Best performance in this respect was found for RM, which has, however, no intensity component and requires a lot of computational time. The Fraction Skill Score (FB) and Practically Perfect Hindcast (PP) generally show very good results. As both scores were found to be redundant, the more popular Fraction Skill Score was chosen as an appropriate method for the verification of quantitative precipitation of high resolution numerical models. To account for large scale error patterns, the Upscaling technique (UP) is added as a second method. Finally, Intensity Scale (IS) seems to be a promising technique, which is very fast and able to detect a specific scale of a spatial error.

3.2.4 Conclusions

Out of the huge set of methods which are available in the Fuzzy verification tool box, three methods and corresponding skill scores were identified to be suitable for a verification of high resolution quantitative precipitation forecast. The "Upscaling" method (UP) considers the mean value within each window to define an event (Zepeda-Arce et al. 2000; Weygandt et al. 2004), while for the "Fractions Skill Score" (FB) fractions of observed and forecast gridpoints exceeding a certain threshold are determined and a score is calculated based on these fractions (Roberts and Lean 2007). "Intensity Scale" (IS) uses a slightly different

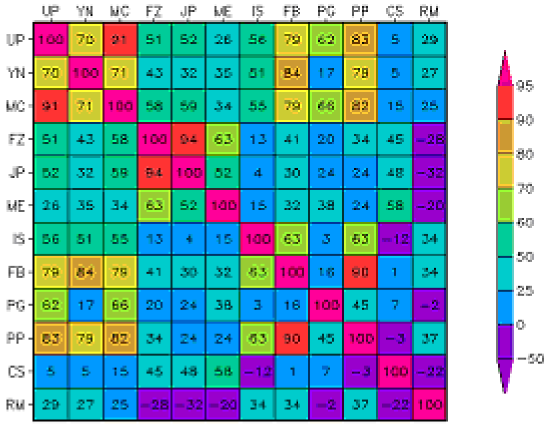

Figure 7: Correlation of the tested scores (see Tab. 1) derived from the averaged scores for each method.

approach: observation and forecast are transformed into binary images (rain larger then threshold or not) and a wavelet analysis is performed on the binary error image (Casati et al. 2004). For more information on the methods, the reader is referred to the given references.

3.3 Application during the D-PHASE operations period (DOP)

3.3.1 D-PHASE - Settings

The Forecast Demonstration Project (FDP) of the WWRP (World Weather Research Programme of WMO) D-PHASE (Demonstration of Probabilistic Hydrological and Atmospheric Simulation of flood Events in the Alpine region) was initiated to demonstrate the ability of forecasting heavy precipitation and related flooding events in the Alpine region. The D-PHASE operations period (DOP) covers the months from June to November 2007. During this time, several numerical and probabilistic models were run and alerts were issued based on those models. In this section, results of the verification of the quantitative precipitation forecast from two of those models, namely the Swiss COSMO models (COSMO-2 (2.2 km) and COSMO-7 (6.6 km)) are presented. The reference observation is the Swiss radar composite (NASS product). The verification is done on different spatial scales by varying the window sizes and for various thresholds defining an event. For the present analysis, window sizes of 1, 3, 5, 9 and 15 gridpoints were chosen for the coarser model and 1, 3, 9, 15, 27 and 45 for the high resolution model. The window sizes $whigh = 3{*}wcoarse$ of the high resolution model are comparable to window sizes $wcoarse$ of the coarse resolution model. The considered thresholds depend on the accumulation times that were evaluated. For 3 h accumulations, the following thresholds were chosen: 0.1, 0.2, 0.5, 1, 2, 5, 10, 20 mm/3h. For 24h accumulations, respective larger thresholds were chosen.

The scores are determined for 3h and 24h rain accumulations and are then aggregated on different time scales. Hereby, for the 3h accumulations always the latest model run is taken for both models beginning with a leadtime of 3h. This means that for COSMO-2 the leadtimes +3 to +6 are evaluated and for COSMO-7 (6.6 km) the leadtimes +3 to +15. It will be shown that the leadtime has no considerable influence on the difference in the skill of the two models. However, if all runs of the lower resolution model are available they should be used for a "fair comparison". The 24h accumulations are only evaluated for the 00UTC runs of both models without any cutoff, i.e. for leadtimes +0 to +24.

3.3.2 Results

The results concerning the verification of quantitative precipitation forecasts during the D-PHASE Operations Period (DOP, June to November 2007) are presented for the two Swiss COSMO models. In Fig. 8, the scores calculated for 3h accumulations using the two methods UP and FB are given - the actual values for the score as well as the differences in the score's value of the two models for the respective scale. Generally, the skill increase with increasing window size and decreasing threshold for both methods and both models. For the Fraction Skill Score, which is displayed in the bottom row of the plot, useful scales are plotted in bold numbers. Those are the scales at which an acceptable level of skill is reached, which is given by $FSS_{uniform}$ lying halfway between random forecast skill and perfect forecast skill (Roberts and Lean 2007). The difference plot (COSMO-2 - COSMO-7) reveals that COSMO-2 has better skill on all scales. Especially small spatial scales (6.6 km to 50 km) and thresholds between 0.4 and 5 mm/3h can be pointed out here. According to the Fraction Skill Score, useful scales are especially those with low thresholds and large window sizes. The COSMO-2 precipitation forecast is hereby also useful on grid scale for very low precipitation thresholds, while COSMO-7 reaches a useful scale only for window sizes larger or equal 3 gridpoints.

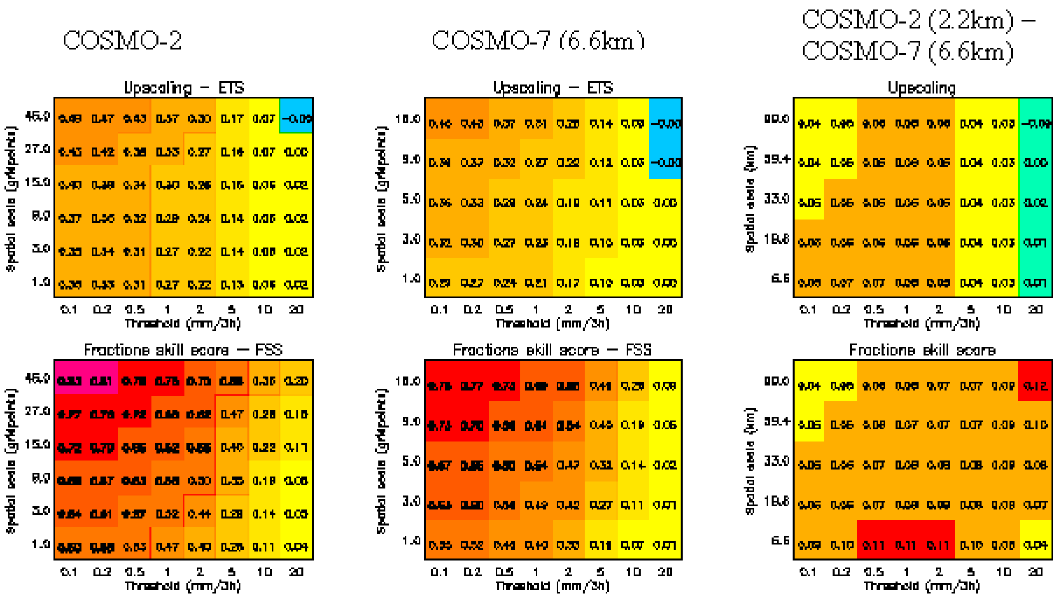

Figure 8: Verification results for the two methods UP (top row) and FB (bottom) and the period from June to November 2007 (3h accumulations). Displayed are the skill scores for the COSMO-2 (left), COSMO-7 (middle) and the differences between those two (right). For the latter, the spatial scale of the COSMO-7 in terms of gridpoints has to be multiplied by 3 to get the comparable spatial scale of COSMO-2.

The second example (Fig. 9) shows the results for 24h accumulations with only the 00UTC model runs used for verification. The skill is higher for both models on all scales and the useful scales are more numerous and almost identical for both models in terms of gridpoints. The difference plot looks slightly different from that for the 3h accumulations. The differences in the skill of the two models are smaller and for a very small number of scales the coarser resolution model has slightly better skill. However, for the larger number of evaluated scales the skill of COSMO-2 is slightly better. This is especially true for the scales where the most pronounced differences were found for 3h accumulations.

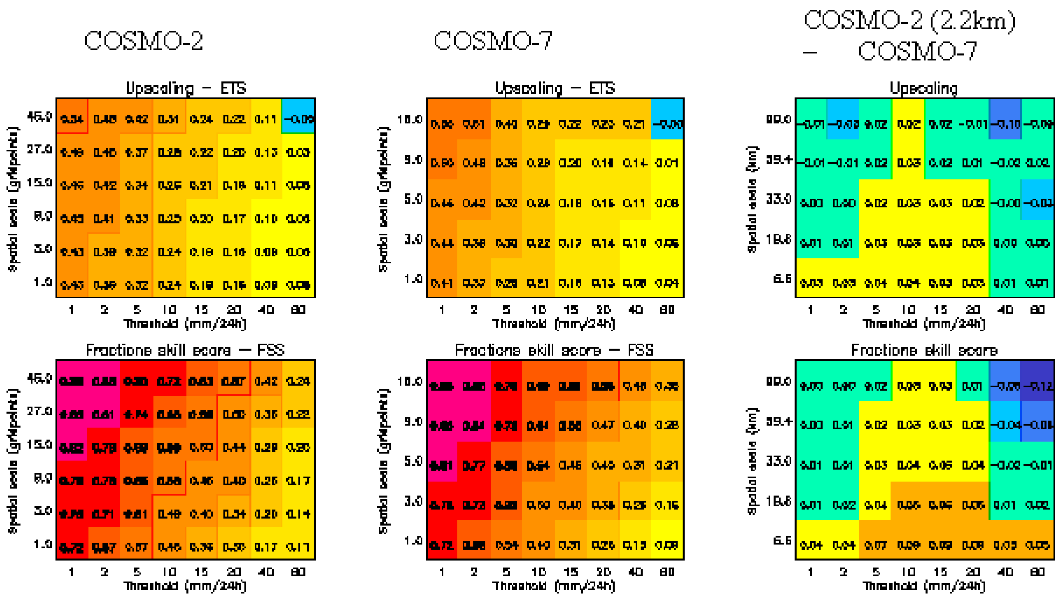

Figure 9: Verification results for the two methods UP (top row) and FB (bottom) and the period from June to November 2007 (24h accumulations). Displayed are the skill scores for the COSMO-2 (left), COSMO-7 (middle) and the differences between those two (right). For the latter, the spatial scale of the COSMO-7 in terms of gridpoints has to be multiplied by 3 to get the comparable spatial scale of COSMO-2.

In Fig. 10, the results of the verification by means of the "Intensity scale" method (IS) are shown. The advantage of this method is that each scale is evaluated separately, while for the other methods presented only the respective smaller scales are filtered. The results for the two models show no consistent behavior for the different scales. However, both accumulation times evaluated for each model reveal comparable skill. COSMO-2 has best skill for small windows (low thresholds) and for large window sizes. For a window size of 16 gridpoints, i.e. a spatial area of approx. 35 km x 35 km the score shows a relatively minimum. COSMO-7 behaves different: the skill is largest for high thresholds and large windows and a relative maximum in skill can be found for a spatial scale of 16 gridpoints, i.e. an area of approx. 105 km x 105 km. The test bed experiment revealed that the "Intensity scale" method is able to detect the specific scale of a spatial error. Imposing a perturbation in form of an horizontal shift, the response of the score was a minimum in skill for that specific scale. The results presented here are averaged over a period of 6 month. However, such a minimum in skill could be detected for the COSMO-2. To decide if this result is an artificial effect or originates from a shift in the precipitation forecast, more detailed investigations are needed.

3.3.3 Conclusions and Outlook

With fuzzy verification methods, forecasts are evaluated on different spatial scales and for varying thresholds. This procedure gives the chance to determine the scales on which a forecast model reveals the desired skill for a specific purpose. Furthermore, a better comparison

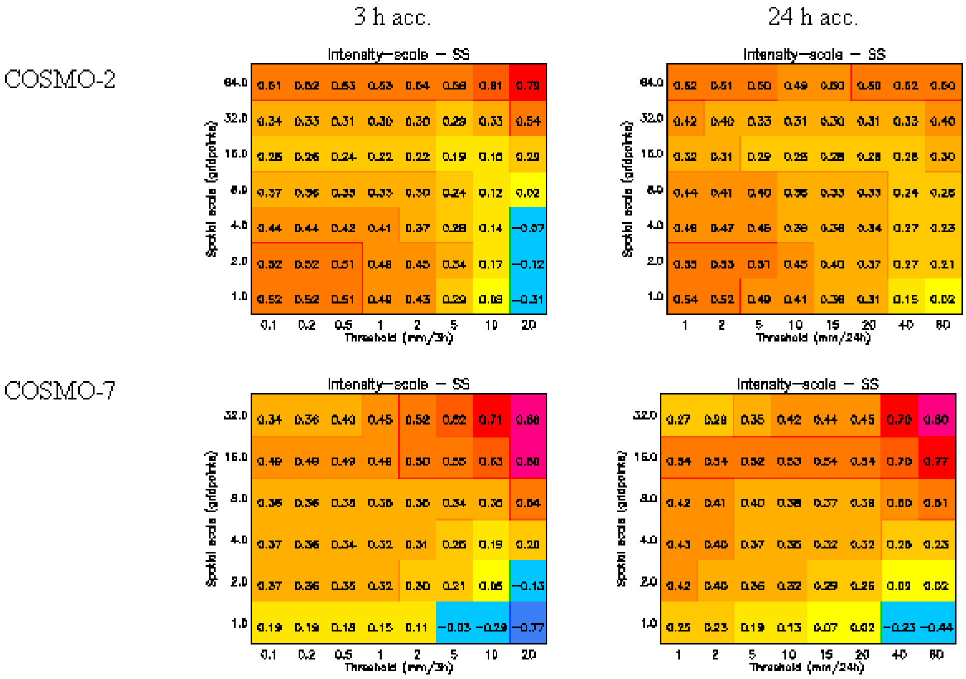

Figure 10: Verification results for the "Intensity scale" method (IS) for the period from June to November 2007 - 3h accumulations (left) and 24h accumulations (right). Displayed are the skill scores for the COSMO-2 (top) and COSMO-7 (bottom).

between models of different spatial resolution is possible. Three methods were chosen by means of a test bed experiment to be suitable for the purpose of precipitation verification and those methods were applied to the quantitative precipitation forecasts of the two Swiss COSMO models during the D-PHASE operations period. The three methods have different characteristics and thus give different results. The upscaling method dumps out high intensities by averaging over the respective window and spreads low intensities on the other hand. The resulting scores give an impression of the accuracy of the mean precipitation values. Fraction Skill Score in contrast considers the fraction of precipitation rates at grid-scale with values above a certain threshold and allows also a direct evaluation of high rain rates. However, due to the low sample of high rain rates, those scales are not classified as useful scales. The results of the two methods (UP and FB) reveal that the high resolution COSMO-2 (2.2 km) has better skill on all scales for 3h accumulations and on most scales for 24h accumulations. Useful scales, according to the Fraction Skill Score, are especially detected for low thresholds and large windows. The third method (IS) allows an evaluation for individual scales and reveals a more complex picture, which is harder to interpret. More detailed investigation is necessary to understand the results of this score. One of its main features is the ability to detect a horizontal shift of the forecast precipitation field. However, the three methods constitute an appropriate basis for the evaluation of quantitative precipitation forecasts of numerical models with different spatial resolutions. The results presented here are averaged over a rather long period of time (6 months). More information on the strength and problems of the individual models can be gained from conditional verification, e.g. stratified by weather types.

4 Verification of DWD-models against high resolution precipitation and SYNOP measurements (U. Damrath, DWD)

4.1 Introduction

Precipitation forecasts with high resolution models often show a high variability of forecasted patterns which have in some cases a characteristic size of only a few grid points. For those cases with observations with nearly the same characteristic size usually it cannot be expected that forecasted values and observed ones match correctly. Therefore, it is necessary to use verification methods that take into account that also right "signals" of forecasted events are judged beyond point by point verification which are further needful for weather elements with no high variability.

4.2 Methods

Traditional verification is carried out at DWD for nearly all weather elements which are observed near the ground and far from the ground. RMSE and BIAS are calculated for temperature, dew point depression, pressure wind (speed, direction and vector wind, no BIAS). Categorical scores are computed for cloud covers at different heights, gusts and precipitation in different classes. In order to check also the capability of the models to signal an event in a region of interest so called "fuzzy" techniques are used. These techniques are fully described by Ebert, 2008, Roberts and Lean 2008, and Casati, 2004. A lot of scores are defined within those frameworks and we have put our attention to the ETS from the upscaling method, the fractions skill score and a modification of the energy squared score. Using the energy squared score observed and forecasted fields are splitted in so called wavelets that contain the differences of values of a scale under consideration from the scale above. The modification relates to the original MSE Skill Score:

$$1 - \frac{2MSE(Y,X)}{En^2(X) + En^2(Y)} \tag{1}$$

to

$$1 - \frac{MSE(Y,X)}{En^2(X) + En^2(Y)} \tag{2}$$

MSE is the mean square difference between observed and predicted precipitation for a given wavelet, En^2 is the sum of the squared values for the given wavelet. The main difference between the wavelet method and the other methods is that with the wavelet method each scale can be separated but with the other methods an overview of all scales below the scale under consideration is given.

4.3 Data

For fuzzy verification forecasts of GME, COSMO-EU and COSMO-DE are compared on the same grid. That means that the forecasts of GME and COSMO-EU are transformed to the COSMO-DE grid. Observations are the values from the radar composite of DWD measurements which are also available on the grid of COSMO-DE. Traditional verification uses SYNOP observations and the forecasted value for the nearest grid point for COSMO-EU and an averaged value over 9 x 9 grid points for COSMO-DE in order to avoid scale dependent effects. Evaluations were made for all forecasts starting in an interval of 3 hours.

4.4 Results

Figs. 11 to 15 show exemplary the behavior of FSS, the ETS for August 2008 and these scores together with the modified energy squared score for January 2009. Also an example for traditional verification is presented (Figs. 16 and 17). These results are representative for most of the other months.

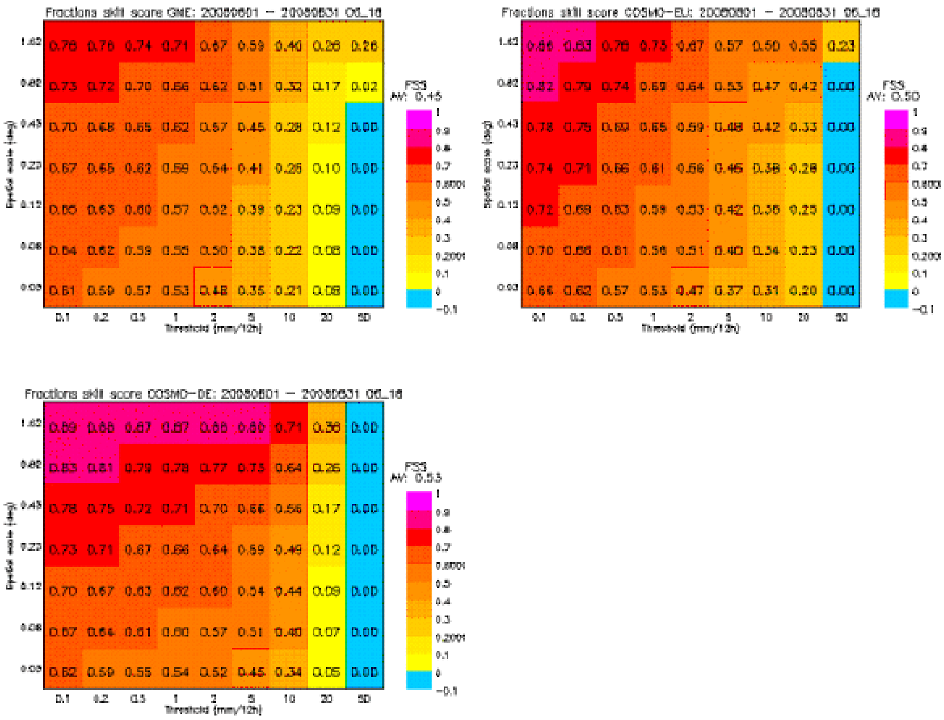

Figure 11: Fractions skill score for forecasts of GME, COSMO-EU and COSMO-DE for August 2008, forecast time 06-18 hours.

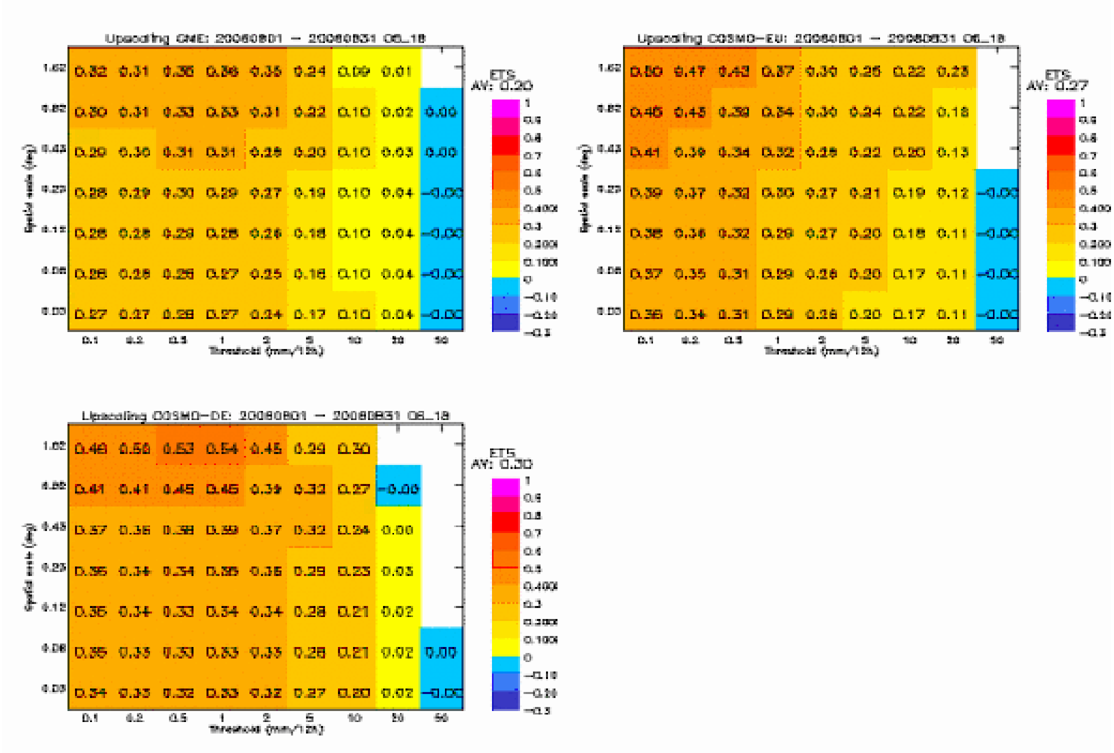

Figure 12: ETS from upscaling for forecasts of GME, COSMO-EU and COSMO-DE for August 2008, forecast time 06-18 hours.

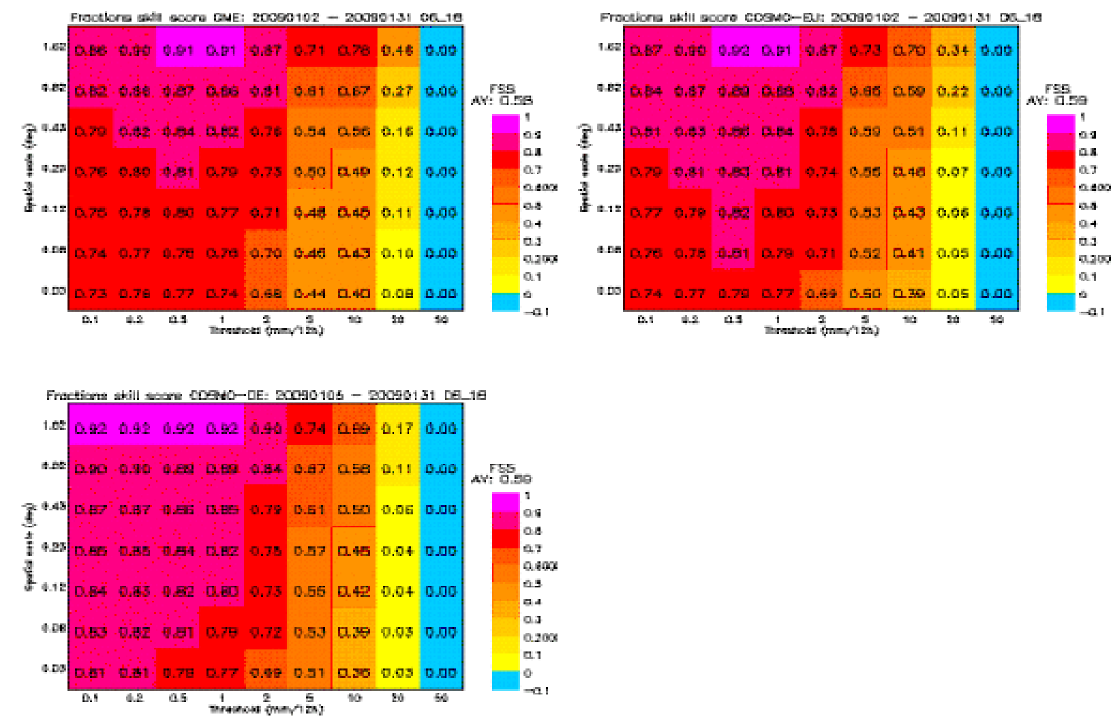

Figure 13: Fractions skill score for forecasts of GME, COSMO-EU and COSMO-DE for January 2009, forecast time 06-18 hours.

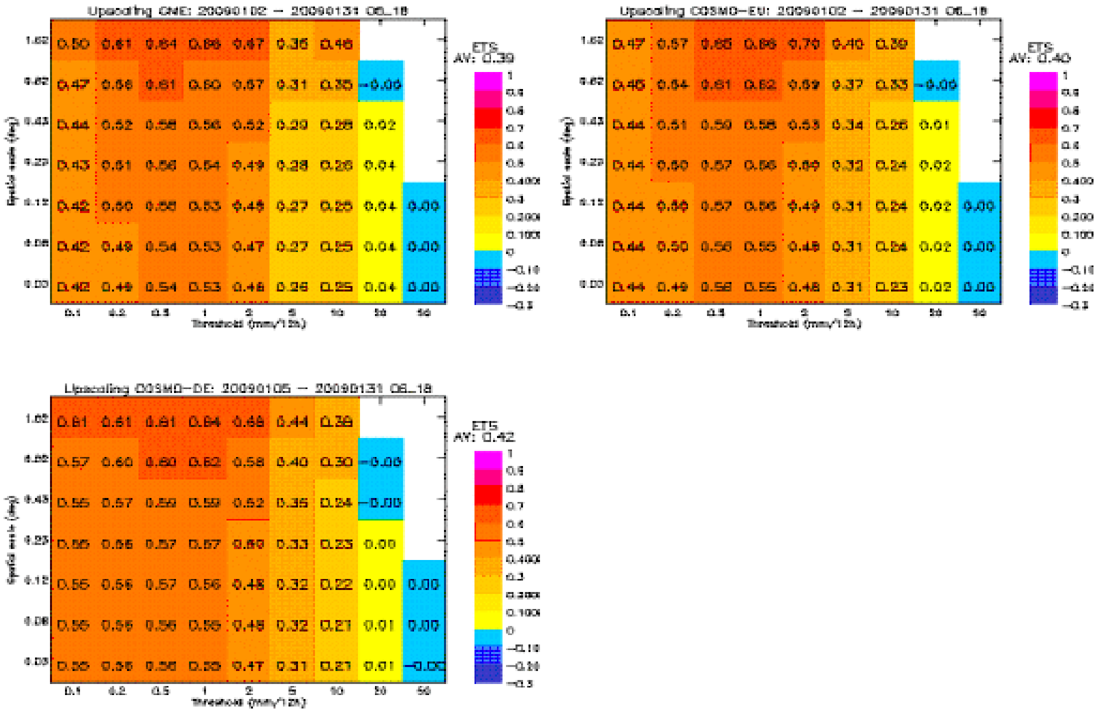

Figure 14: ETS from upscaling for forecasts of GME, COSMO-EU and COSMO-DE for January 2009, forecast time 06-18 hours.

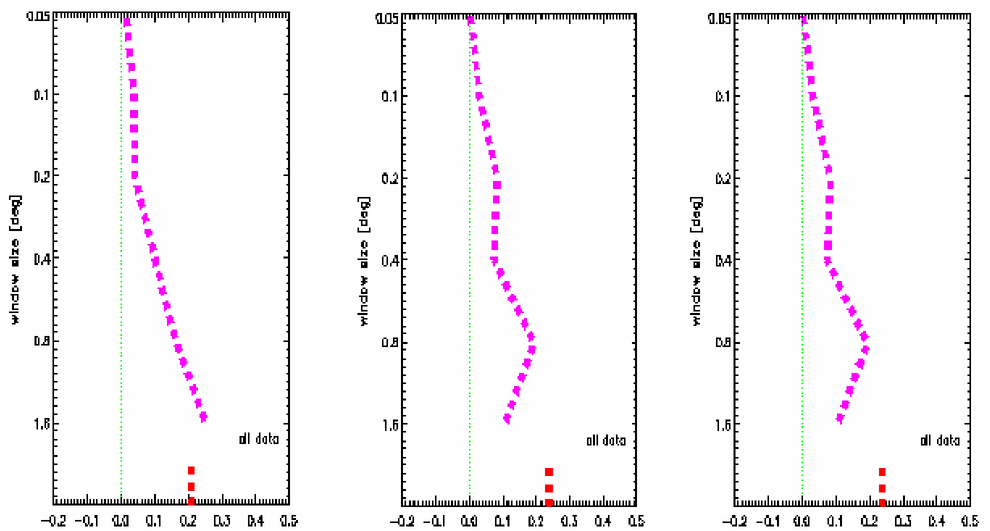

Figure 15: Modified energy squared score for forecasts of GME (left), COSMO-EU (center) and COSMO-DE (right) for January 2009, forecast time 06-18 hours; horizontal: skill score; vertical: window size; red mark: skill for all scale.

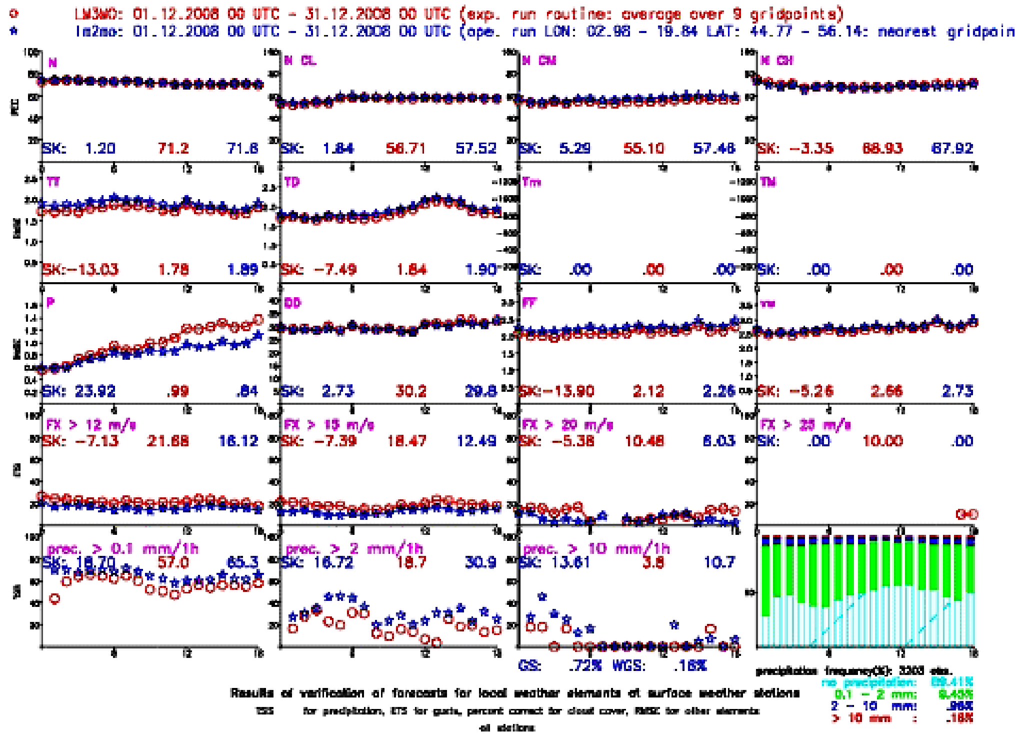

Figure 16: Example of traditional verification for surface weather elements for December 2008 (Start of forecasts 00UTC, scores measuring the quality of forecasts).

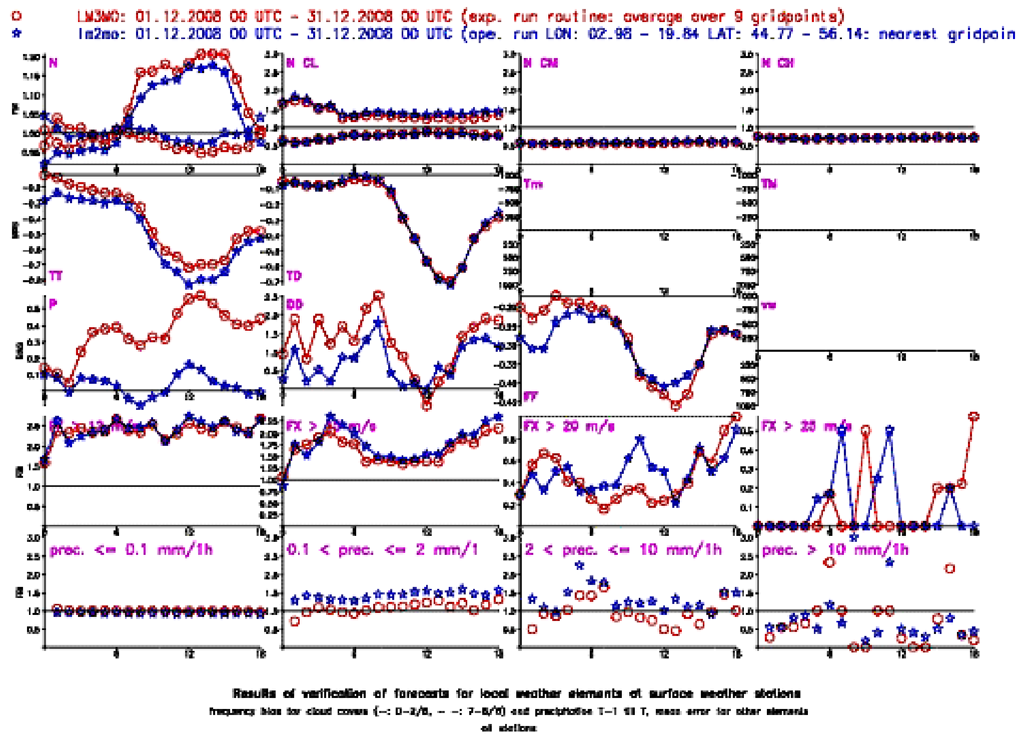

Figure 17: Example of traditional verification for surface weather elements for December 2008 (start of forecasts 00UTC, scores measuring the bias of forecasts).

This is a list of abbreviations which appear in Fig. 16:

- N: total cloud cover
- N CL: cloud cover of low clouds
- N CM: cloud cover of medium height clouds
- N CH: cloud cover of high clouds
- TT: 2m-temperature
- TD: 2m-dew point depression
- Tm: 2m-minimum temperature
- TM- 2m-maximum temperature
- P: pressure reduced to mean sea level
- DD: 10m-wind direction
- FF: 10m-wind speed
- vw: 10m-vector wind
- FX: gust over the given threshold
- Prec: precipitation for the given thresholds
- SK: skill of COSMO-EU agains COSMO-DE
- GS and WGS: global skill and weighted(different weights for each element) global skill over all elements
- red numbers: averages of the score for COSMO-DE
- blue numbers: averages of the score for COSMO-EU

4.5 Summary

Fuzzy verification methods show the following results which could be expected, but are assessed now:

- The quality of precipitation forecast is best for low precipitation values and large areas.
- For extreme precipitation values the dependency on the scale is as for low precipitation amounts. Higher quality occurs for larger areas. That means, indeed, that right "signals" often can be detected - especially by the models with finer resolution.
- Unfortunately, the number of those cases is relatively low and, therefore the statistical value of the scores is limited.
- The comparison between the models - and that follows from all the scores under consideration - indicate that the quality of forecasts increases in general with the degree of horizontal resolution.

- The question, whether the forecasts of a model has an economical benefit cannot be answered using these scores. It depends on the specific cost loss relationship. But if one can accept that a fractions skill score of above 0.5 or a ETS from upscaling is higher than 0.3 than one can find some areas in the graphics which indicate the horizontal scale and the amounts of precipitation that could be accepted as "good" forecasts.
- The example with the modified energy score shows that below a scale of 0.2° forecasts of GME have no skill. COSMO-models have a small skill below this range. For precipitation scales below 0.1 generally no skill can be expected.

Concerning the traditional verification it can be stated:

- Cloud covers are modeled with nearly the same quality for both models.
- COSMO-DE forecasts of temperature and dew point depression show small but systematic advantages against COSMO-EU.
- Although RMSE for pressure is for larger forecast times in COSMO-DE higher than in COSMO-EU (this is mainly related to the BIAS) wind forecasts of COSMO-DE are more trustable than those of COSMO-EU. This is also valid for gusts.
- QPF verification against SYNOP observations shows in general a frequency bias which is closer to 1 for COSMO-DE than for COSMO-EU. The small overestimation of COSMO-EU forecasts compared to COSMO-DE leads to a bit higher TSS for COSMO-EU. The evaluation of other scores that are not so strongly dependent on the frequency bias is under consideration.

5 Post-processing of COSMO-7 and COSMO-2 Precipitation with the Neighborhood Method (P. Kaufmann, MeteoSwiss)

5.1 Introduction

With increasing resolution, the numerical weather prediction models advance towards scales that are less and less deterministically predictable. This is especially true for precipitation with its high temporal and spatial variability. The time of onset and the location of a single convective cell for example cannot be predicted exactly due to the random nature of convection. This uncertainty in time and space needs to be accounted for when using high-resolution model output fields. The neighborhood method (Theis et al., 2005) accounts for the probability related to the spatial and temporal uncertainty on the scale of a few model grid cells. It solves the above-mentioned predictability problem by sampling a spatio-temporal neighborhood around each grid cell to derive a probabilistic forecast for that cell. Doing this, the neighborhood method provides a way to derive a probabilistic forecast from a deterministic model. In the present study, precipitation forecasts of the two Swiss models, the COSMO-7 with a grid spacing of 7 km (6.6 km after Feb. 2008), and the COSMO-2 with 2.2 km, are used as input.

5.2 Optimizing the neighborhood

5.2.1 Method

The neighborhood size has an important impact on the magnitude of the predicted probabilities. Theis (2005) found that she could not optimize the neighborhood size based on objective verification alone. With this unpromising perspective, I based the optimization for the Swiss COSMO models on case studies of precipitation cases. The optimization of the neighborhood uses a graphical comparison of the forecast probabilities with actual precipitation measurements of 2006 - 2008. An objective evaluation of the neighborhood determined in this way remains highly desirable and will be done this year (2009).

5.2.2 Spatial radius of neighborhood

Theis et al. (2005) tested three different sizes for the neighborhood: a small, a medium and a large neighborhood with a spatial radius of 3, 6 and 10 grid points, respectively. In this study, four sizes of 1, 3, 5, and 10 grid point radius are compared for COSMO-7 (Fig. 18). While the smallest radius retains most of the structural details, it also produces probabilities that are either very high or very low with much fewer points in between than are meaningful for a probabilistic forecast. The largest radius on the other hand only produces low probabilities, blurring the structural information very strongly. The ideal neighborhood radius is a trade-off between these extremes, lying for COSMO-7 somewhere in the range of 3 to 5 (Kaufmann, 2008) and for COSMO-2 between 8 to 15 grid points radius.

5.2.3 Shape of the neighborhood

Over homogeneous, flat terrain, a circular shape in space is the natural choice. In complex terrain or along surface property discontinuities however, it seems advantageous to account for the anisotropy. Theis et al. (2005) tested different anisotropic spatial shapes. The impact

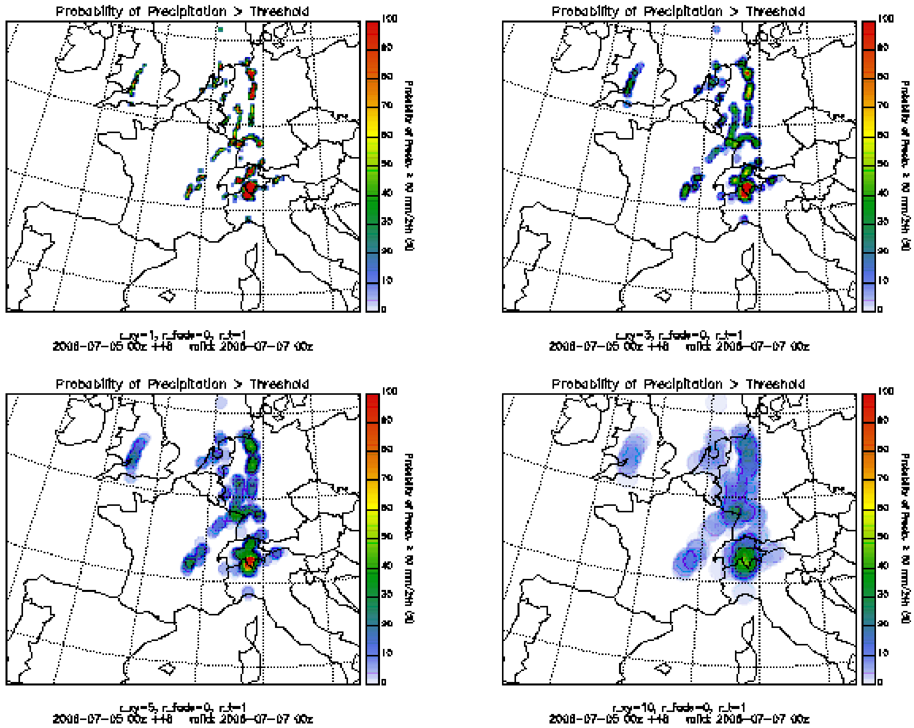

Figure 18: Probabilities of precipitation derived from COSMO-7 with a spatial neighborhood radius of 1, 3, 5, and 10 grid points, in this order.

proved to be quite small and at times even detrimental. Therefore, a circular neighborhood was chosen here, without further experimenting with the spatial shape.

The ellipsoidal shape of the neighborhood used by Theis et al. (2005) means that the spatial radius depends on the temporal distance to the central point. This implies a space-time dependency. It has at the same time the practical effect of smoothing the edges of the regions with high probabilities, because grid points near the central point contribute to the neighborhood at more time levels than those further away. If it is assumed that the time and space dimensions are independent, the spatial radius is constant in time, leading to a cylindrical neighborhood. This assumption is true if the ambient wind is nearly zero, and becomes increasingly inaccurate with increasing advection. The cylindrical neighborhood has the practical disadvantage of showing unreasonably sharp boundaries along the edges of precipitation. Smoothing of the edges has to be added to ameliorate this. A smoothing with linearly fading weights is introduced in the next section.

5.2.4 Linearly fading weights

The cylindrical neighborhood shape, in combination with localized strong precipitation peaks, leads to a dotted plot of probabilities, with a circular dot surrounding each major precipitation peak (Fig. 18). Such an unwanted feature could be smoothed out in an additional step, but it is more meaningful to do this as part of the neighborhood processing. This is achieved by adding an additional zone around the equally weighted neighborhood, with decreasing weights. For simplicity, the weights are chosen to decrease linearly. The meteorological reasoning behind such a zone is that given a convective rain forecast at the central point, it is equally likely to appear within a certain radius, but the likelihood decreases beyond this radius and finally reaches zero at some larger distance. Fig. 19 shows the weight depending on the spatial distance of a grid cell to the neighborhood center for three neighborhood sizes.

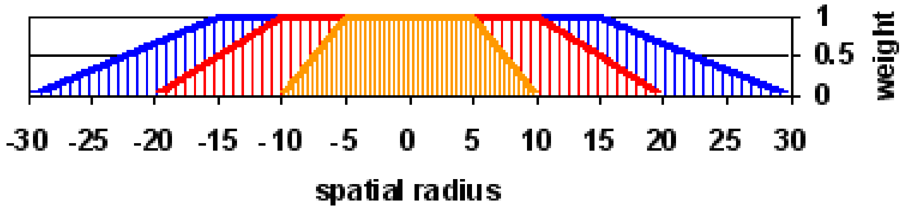

Figure 19: Weights of grid cells depending on the distance from the central cell, for tree different sizes of the neighborhood. Blue: Zone of equal weight and fading zone with 15 grid points radius each; red 10; orange: 5.

The result of adding such a fading zone is shown in Fig. 20. Here, the probability fields of the lower panels of Fig. 18 are repeated, but this time calculated with an additional linear fading zone of the same radius as the neighborhood itself. The areas are much smoother than in Fig. 18, and the peak values are somewhat decreased. This is because the total radius including the fading zone is now doubled.

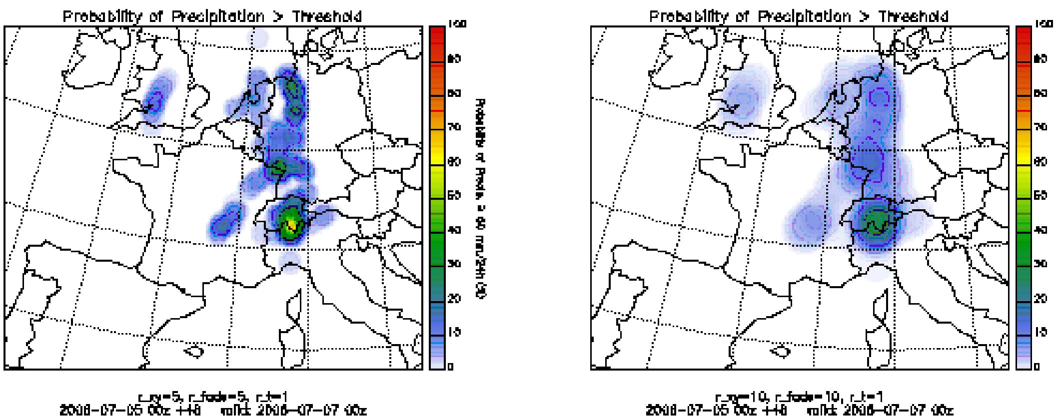

Figure 20: Same as lower two panels of Fig. 18 but with an additional fading zone of the same radius as the neighborhood, (left) 5 and (right) 10 grid points, added.

The cylindrical neighborhood shape with a fading zone is approximately equivalent with an ellipsoidal neighborhood with a larger temporal but smaller spatial radius (Fig. 21). With the same temporal radius, the linear fading provides smoother edges than the discrete weights of the ellipsoidal shape, while the values within the edges are retained. The smoother areas with non-zero probabilities are more realistic than the patchy areas of the ellipsoidal shapes.

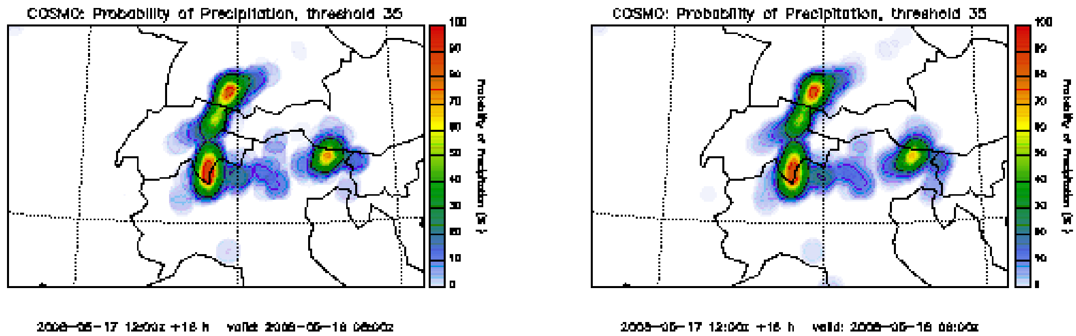

Figure 21: Approximately equivalent neighborhood shapes for COSMO-2. Left: ellipsoidal with 13 grid point radius, right: cylindrical with 8 grid point radius plus 8 grid point fading zone giving smoother transitions and edges. Shown is the 12h sum ending 18 May 2008 at 6:00 UTC.

5.2.5 Temporal radius

The temporal radius of the neighborhood represents the uncertainty in time of a rainfall prediction. For convective precipitation, it is known to span several hours. However, tests with temporal radius of 0, 1, 3, and 6 hours showed relatively little influence. The rainfall accumulation of 6 hours and even more so of 24 hours already has a smoothing effect on the time dimension, rendering the temporal extent of the neighborhood less critical. From the computational point of view, a smaller temporal radius is favorable because it is less costly.

5.3 Examples of products

5.3.1 COSMO-7, threshold 25 mm / 6h

Fig. 22 displays an example for a COSMO-7 product with a threshold of 25 mm per 6 hours. The situation shown occurred in August 2006, with a small region of strong precipitation over southern Switzerland. The probabilities for precipitation above the threshold level of 25 mm / 6h reach 50% in that region. Although probabilities and single precipitation events cannot be compared directly, large probabilities should coincide with rainfall above the threshold for a majority of cases. The probabilities generally decrease with increasing neighborhood size, thus a frequent occurrence of moderate probabilities with actual rainfall above the threshold as in Fig. 22 would indicate that the chosen neighborhood size is still slightly larger than optimal.

5.3.2 COSMO-2, threshold 25 mm / 6h

Fig. 23 shows on the left the product for COSMO-2 with a 25 mm per 6 hours threshold. On the right side, the actually occurred rainfall amount as measured at Swiss stations is shown. This is during the same episode already presented in Fig. 21, but 12 hours earlier.

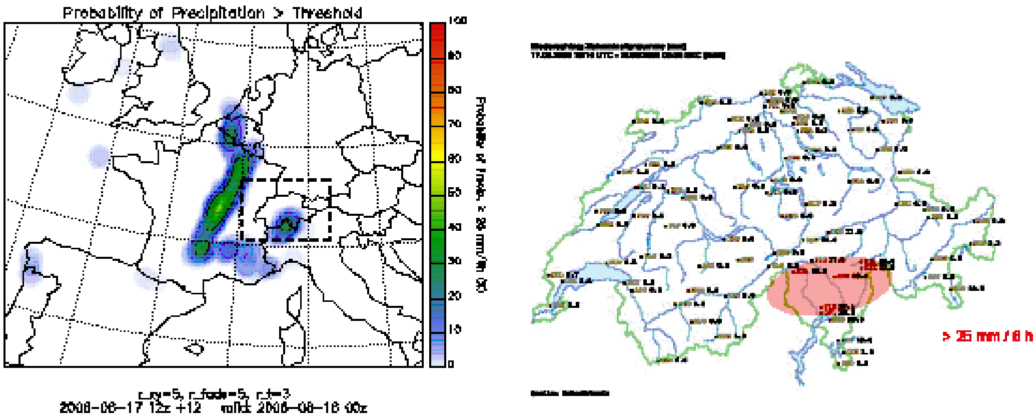

Figure 22: Probabilities predicted by the neighborhood method (left) and actual rainfall sum (right) for 2006-08-18 00:00 UTC.

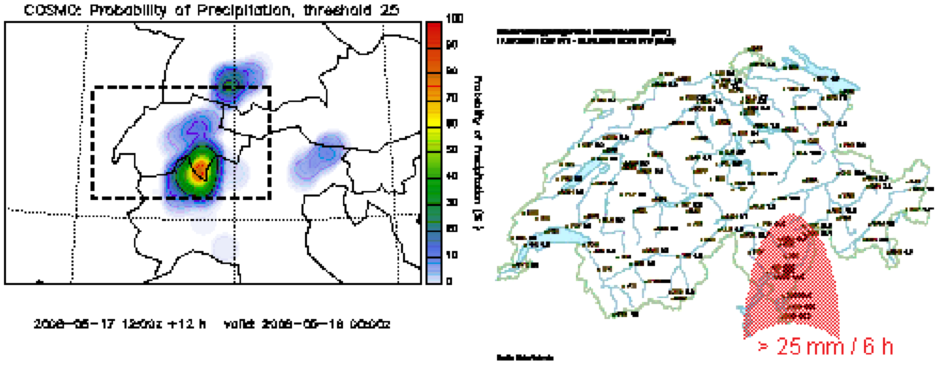

Figure 23: Probabilities predicted by the neighborhood method (left) and actual rainfall sum (right) in Switzerland (within dashed box in left figure) for 2008-05-18 06:00 UTC.

5.3.3 COSMO-7, threshold 50 mm / 24h

With a forecast range of just 24 hours, COSMO-2 cannot be used with the neighborhood method for 24h precipitation sums, unless the temporal radius is set to zero. The corresponding product for COSMO-7 (Fig. 24) shows a case with strong precipitation over northern Switzerland. The region of probabilities of 50% and more for precipitation above the threshold level of 50 mm / 24h nicely fits with the observed precipitation.

5.4 Conclusions and Outlook

The neighborhood method is used to derive a probabilistic forecast from a deterministic model. The probability is related to local-scale spatial and temporal model uncertainty. The resulting plots are compared to real cases and are optimized in this way. For COSMO-7, a spatial radius of 5 grid cells (33 km) with an additional zone of 5 grid cells with decreasing

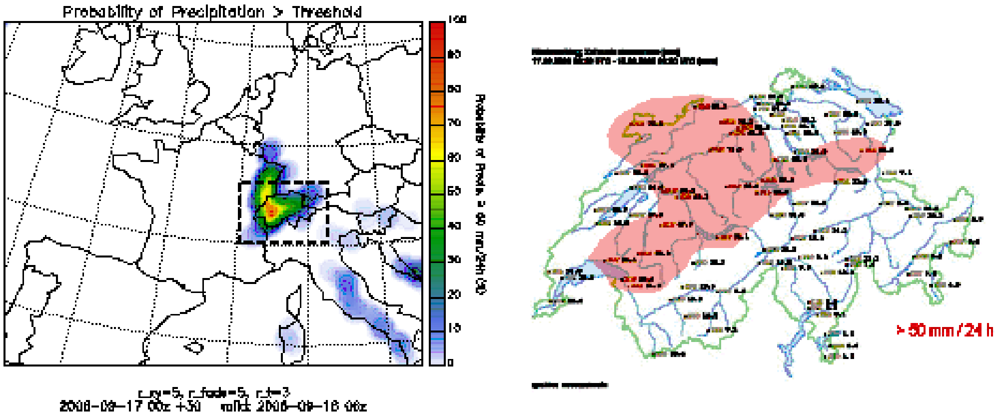

Figure 24: Probabilities predicted by the neighborhood method (left) and actual rainfall sum (right) for 2006-09-18 06:00 UTC.

weights and a temporal radius of 3 hours gives probabilities which are in accordance with the observed events and non-events. The radius of the zone with equal weight and of the zone with fading weights combine to an ”effective” radius of approximately 7.5 grid cells (50 km). The temporal radius is currently chosen to be 3 hours for COSMO-7; however the extent of the neighborhood in time is of lesser importance. The respective values optimized in the same way for COSMO-2 are 10 grid cells (22 km) radius plus additional 10 grid cells for the fading zone, with a temporal radius of 1 hour. The two zones of the neighborhood add up to an ”effective” radius of 15 grid cells (33 km). Compare this to the German Weather Service (DWD) which uses an ellipsoidal neighborhood with 13 grid point radius (36 km) for its 2.8 km resolution model. A throughout validation however, using objective verification, will have to follow to prove the quality of the probability forecast with the chosen neighborhood. Such a verification is planned for this year (2009). The neighborhood method is able to represent the local uncertainty concerning the prediction of the exact position and onset of precipitation. It does not account for the uncertainty related to synoptic forcing, which can be done using a model ensemble forecast. A combination of the ensemble and the neighborhood methods would combine both synoptic-scale and small-scale uncertainties.

6 Statistical properties of precipitation in boxes of various sizes (M. S. Tesini, C. Cacciamani, T. Paccagnella, ARPA-SIMC).

Model verification at ARPA-SIMC, in particular QPF, is operationally performed by using an upscaling method of both forecasted and observed fields over areas of selected size. The domain is divided in many squared areas and, for each area and for the chosen period of time, some indicators deduced by the distribution of the forecasted precipitation field (such as mean, maximum value, 90th percentile), are compared to those deduced by the distribution of the observed rainfall. Results concerning models performance depend on the considered distribution parameter; for example categorical verification scores for COSMO-I7 are better if the chosen indicator is the maximum value of precipitation in the area rather than those obtained selecting the mean value. Nevertheless all the distribution parameters are equally useful because they provide different information on the ability of the model in reproducing the precipitation field. Our interest is therefore addressed to the study of the distribution of the precipitation field, in order to analyze the capability of the models in reproducing precipitation statistical properties. Currently we focused on COSMO-I7 and ECMWF verification. The results suggest that verification methods based on the statistical description of model QPF and observations give more credit to high resolution model than the traditional standard verification, based on the comparison of model grid-point against observation (for example the choice of nearest grid point). We have applied a methodology based on the estimation of the distribution function of the observed/predicted precipitation field considering all points that fall in each area throughout the examined periods. Each area contains a lot of observation points as well as model grid-points: we assume that the occurrence of precipitation in each of these points is equiprobable. In order to consider well-founded this assumption, areas should be chosen carefully taking into account orographical and geographical features. The dimension of the selected areas is a critical point too. We followed a pragmatic approach for the choice of the dimension on the box: we should have enough forecast and observation points to perform significant statistics in each area. We chose squared areas of about 100 km x 100 km (see Fig. 25), containing about 200 COSMO-I7 grid-points or 16 IFS-ECMWF grid-points.

There are more than 1000 rain-gauges, belonging to regional/local administrations and provided by Italian Department of Civil Protection over the Italian territory, and the number of rain-gauges varies from 20 to 130 depending on the selected region. Tests on smaller area have been performed, but the results are very sensitive to the density of the observation points. We approached the problem mostly from a qualitatively point of view using a boxplot representation. A boxplot is a simple graphical summary of a dataset that can be useful to display differences between populations without making any assumptions of the underlying statistical distribution. It is possible to obtain information about the spread of the distribution and a roughly description of key measures that define it, such as the median, the quartiles, maximum and minimum values, and a range of values identified as "outliers". In this way we can, at least qualitatively, describe the type of precipitation: for instance the resulting distribution of a convective rainfall system possesses a long "tail" while the distribution of a widespread light rain event presents a distribution with a peak near a low amount of rainfall and little or no "tail". In this context, COSMO-I7 and IFS-ECMWF are used as representative of models with different horizontal resolution without focusing on their quality. In Fig. 26 the time-series of the daily forecasted/observed rainfall distribution are shown, obtained considering all the model grid-points and all the observed rain that fall daily (00-24UTC) within each area for ECMWF (blue), COSMO-I7 (red) and observations (green) during Autumn 2007. The number of stations in this area (131) is comparable with

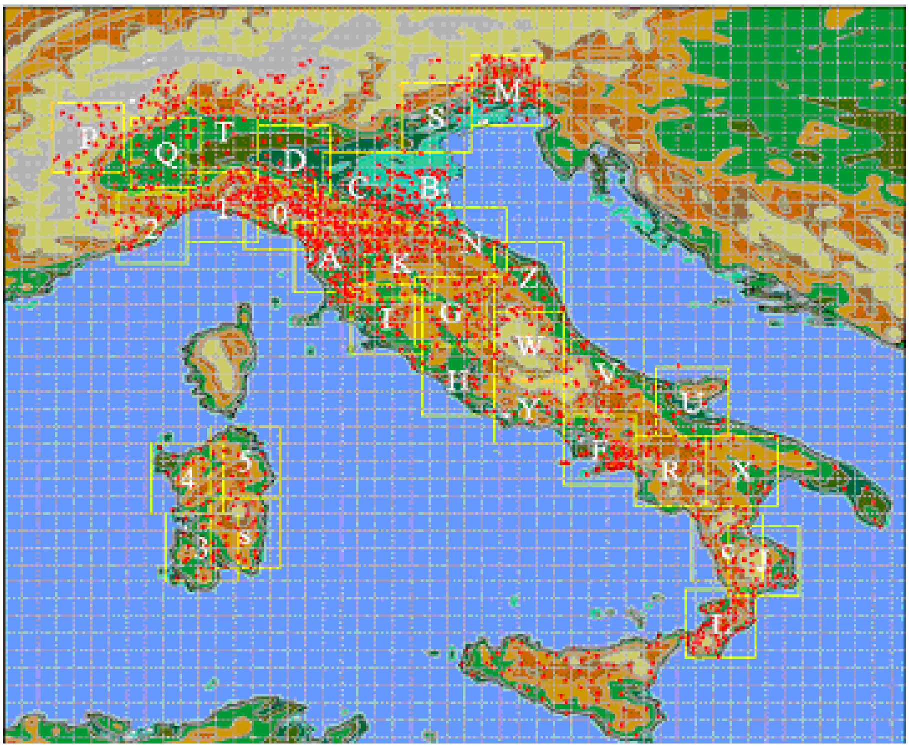

Figure 25: Box distribution.

COSMO-I7 grid-points (192), while ECMWF presents only 16 grid-points.

In general, we observe that COSMO-I7 is more realistic than ECMWF in reproducing the intra-box variability. However, COSMO-I7 presents both a large number of false alarms and high "spikes". On the other hand, ECMWF shows a greater number of missed alarms, especially if we chose high reference thresholds. According to most standard verification measures, COSMO-I7 forecast may have poor quality, but it might be very valuable to the forecaster since it provides additional information on the distribution and variability of the rain field over the considered region. The role of horizontal resolution is very important since it allows a more realistic simulation of the precipitation distribution, in particular for the "tails", that nevertheless represent the most interesting values of the rain field, especially for the emission of warning alerts to the Civil Protection. A better description of the distribution function of the observed precipitation field over each area represents a fundamental step in order to apply this methodology also to COSMO-I2, as the difference in the density of model grid-points and observations should represent a critical point. In fact, if we consider an area in which the number of model grid-points is significantly greater than the number of observations, such as the one shown in Fig. 27 where there are 34 stations, 212 COSMO-I7 grid-points and 16 ECMWF grid-points, we can see that "tail" of COSMO-I7 distribution is overestimated compared to the "tail" distribution of observations. Is it a sampling problem or does it depend on the behavior of the model in that area ? The question is still open.

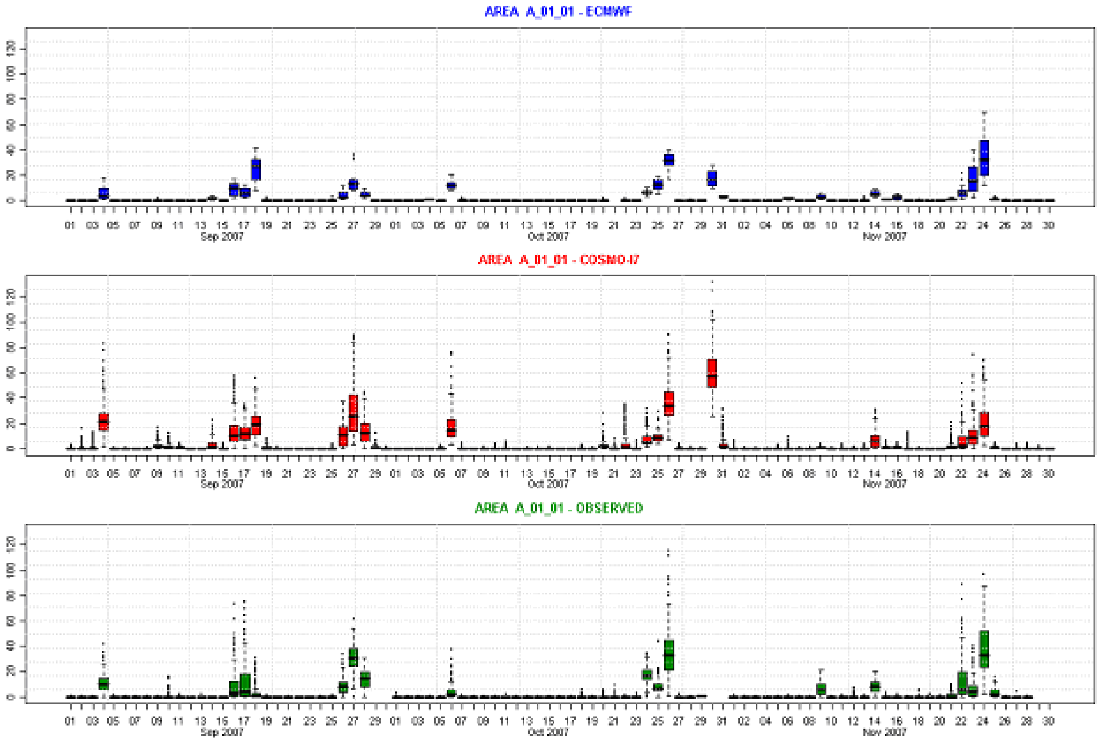

Figure 26: Time-series of the daily forecasted/observed rainfall.

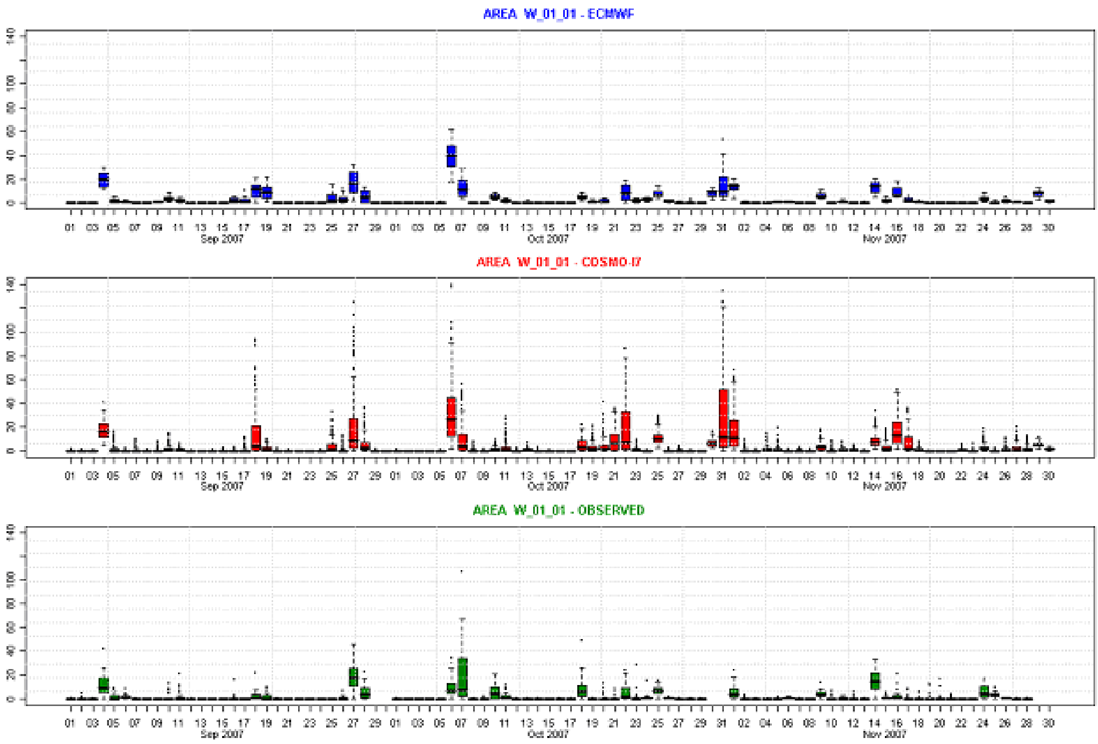

Figure 27: Example of area in which the number of model grid-points is significantly greater than the number of observations.

7 Open points

- The fact that precipitation patterns cannot be trusted at small scales is obviously dependent on the synoptic configuration. When a lot of uncertainty in the positioning is existent in purely convective situations, it is well possible that precipitation patterns are reproduced correctly in situations where orographic forcing is dominant (as long as the orography is correctly represented). Verification stratified by a weather type classification is thus recommended.
- By building statistics of precipitation values in a neighborhood of a gridpoint it is possible to build probabilistic precipitation forecasts. The same can be provided by Ensemble Prediction Systems (EPS), especially convection permitting EPS. Both methods should be compared by classical or fuzzy verification. Both systems can also be combined by applying the upscaling or the neighborhood method to the EPS.

8 Contributing scientists

The authors of the four articles are the contributing scientists.

9 References

Casati, B., Ross, G., Stephenson, D.B., 2004: A new intensity-scale approach for the verification of spatial precipitation forecasts. Meteorological Applications 11, 141-154.

Ebert, E.E., 2008: Fuzzy verification of high resolution gridded forecasts: A review and proposed framework. Meteorological Applications, 15, 51-64.

Kaufmann, P., 2008: Postprocessing of the aLMo [former name of COSMO-7] precipitation with the neighborhood method. COSMO Newsletter No. 7, DWD, Germany, 57-60. Available from http://www.cosmo-model.org.

Roberts, N.M., Lean, H.W., 2008: Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. Monthly Weather Review 136, 78-97.

Theis, S., 2005: Deriving probabilistic short-range forecasts from a deterministic high- resolution model. Ph. D. thesis, University Bonn, Germany.

Theis, S., Hense, A., and Damrath, U., 2005: Probabilistic precipitation forecasts from a deterministic model: a pragmatic approach. Meteorol. Appl., 12, 257-268.

Weygandt, S.S., Loughe, A.F., Benjamin, S.G., Mahoney, J.L., 2004: Scale sensitivities in model precipitation skill scores during IHOP. 22nd Conference Severe Local Storms. American Meteorological Society, Hyannis, MA, 4-8 October 2004.

Zepeda-Arce, J., Foufoula-Georgiou, E., Droegemeier, K.K., 2000: Spacetime rainfall organization and its role in validating quantitative precipitation forecasts. Journal of Geophysical Research, 105(D8), 10, 129-10, 146.

List of COSMO Newsletters and Technical Reports

(available for download from the COSMO Website: www.cosmo-model.org)

COSMO Newsletters

No. 1: February 2001.

No. 2: February 2002.

No. 3: February 2003.

No. 4: February 2004.

No. 5: April 2005.

No. 6: July 2006.

No. 7: April 2008; Proceedings from the 8th COSMO General Meeting in Bucharest, 2006.

No. 8: September 2008; Proceedings from the 9th COSMO General Meeting in Athens, 2007.

No. 9: December 2008.

COSMO Technical Reports

No. 1: Dmitrii Mironov and Matthias Raschendorfer (2001):
Evaluation of Empirical Parameters of the New LM Surface-Layer Parameterization Scheme. Results from Numerical Experiments Including the Soil Moisture Analysis.

No. 2: Reinhold Schrodin and Erdmann Heise (2001):
The Multi-Layer Version of the DWD Soil Model TERRA_LM.

No. 3: Günther Doms (2001):
A Scheme for Monotonic Numerical Diffusion in the LM.

No. 4: Hans-Joachim Herzog, Ursula Schubert, Gerd Vogel, Adelheid Fiedler and Roswitha Kirchner (2002):
LLM ‾ the High-Resolving Nonhydrostatic Simulation Model in the DWD-Project LITFASS.
Part I: Modelling Technique and Simulation Method.

No. 5: Jean-Marie Bettems (2002):
EUCOS Impact Study Using the Limited-Area Non-Hydrostatic NWP Model in Operational Use at MeteoSwiss.

No. 6: Heinz-Werner Bitzer and Jürgen Steppeler (2004):
Documentation of the Z-Coordinate Dynamical Core of LM.

No. 7: Hans-Joachim Herzog, Almut Gassmann (2005):
Lorenz- and Charney-Phillips vertical grid experimentation using a compressible non-hydrostatic toy-model relevant to the fast-mode part of the 'Lokal-Modell'

No. 8: Chiara Marsigli, Andrea Montani, Tiziana Paccagnella, Davide Sacchetti, André Walser, Marco Arpagaus, Thomas Schumann (2005):
Evaluation of the Performance of the COSMO-LEPS System

No. 9: Erdmann Heise, Bodo Ritter, Reinhold Schrodin (2006):
Operational Implementation of the Multilayer Soil Model

No. 10: M.D. Tsyrulnikov (2007):
Is the particle filtering approach appropriate for meso-scale data assimilation ?

No. 11: Dmitrii V. Mironov (2008):
Parameterization of Lakes in Numerical Weather Prediction. Description of a Lake Model.

No. 12: Adriano Raspanti (2009):
COSMO Priority Project "VERification System Unified Survey" (VERSUS): Final Report

No. 13: Chiara Marsigli (2009):
COSMO Priority Project "Short Range Ensemble Prediction System" (SREPS): Final Report

No. 14: Michael Baldauf (2009):
COSMO Priority Project "Further Developments of the Runge-Kutta Time Integration Scheme" (RK): Final Report

No. 15: Silke Dierer (2009):
COSMO Priority Project "Tackle deficiencies in quantitative precipitation forecast" (QPF): Final Report

No. 16: Pierre Eckert (2009):
COSMO Priority Project "INTERP": Final Report

COSMO Technical Reports

Issues of the COSMO Technical Reports series are published by the *COnsortium for Small-scale MOdelling* at non-regular intervals. COSMO is a European group for numerical weather prediction with participating meteorological services from Germany (DWD, AWGeophys), Greece (HNMS), Italy (USAM, ARPA-SIMC, ARPA Piemonte), Switzerland (MeteoSwiss), Poland (IMGW) and Romania (NMA). The general goal is to develop, improve and maintain a non-hydrostatic limited area modelling system to be used for both operational and research applications by the members of COSMO. This system is initially based on the COSMO-Model (previously known as LM) of DWD with its corresponding data assimilation system.

The Technical Reports are intended

- for scientific contributions and a documentation of research activities,
- to present and discuss results obtained from the model system,
- to present and discuss verification results and interpretation methods,
- for a documentation of technical changes to the model system,
- to give an overview of new components of the model system.

The purpose of these reports is to communicate results, changes and progress related to the LM model system relatively fast within the COSMO consortium, and also to inform other NWP groups on our current research activities. In this way the discussion on a specific topic can be stimulated at an early stage. In order to publish a report very soon after the completion of the manuscript, we have decided to omit a thorough reviewing procedure and only a rough check is done by the editors and a third reviewer. We apologize for typographical and other errors or inconsistencies which may still be present.

At present, the Technical Reports are available for download from the COSMO web site (www.cosmo-model.org). If required, the member meteorological centres can produce hard-copies by their own for distribution within their service. All members of the consortium will be informed about new issues by email.

For any comments and questions, please contact the editors:

Massimo Milelli
Massimo.Milelli@arpa.piemonte.it

Ulrich Schättler
Ulrich.Schaettler@dwd.de