# Consortium for Small-Scale Modelling

Technical Report No. 37

## *The COSMO Priority Project INSPECT Final Report*

January 2019

DOI: 10.5676/DWD_pub/nwv/cosmo-tr_37

*Deutscher Wetterdienst*
*MeteoSwiss*
*Ufficio Generale Spazio Aereo e Meteorologia*
*ΕΘΝΙΚΗ ΜΕΤΕΩΡΟΛΟΓΙΚΗ ΥΠΗΡΕΣΙΑ*
*Instytucie Meteorogii i Gospodarki Wodnej*
*Administratia Nationala de Meteorologie*
*ROSHYDROMET*
*Agenzia Regionale Protezione Ambiente Piemonte*
*Agenzia Regionale Prevenzione Ambiente Energia Emilia Romagna*
*Centro Italiano Ricerche Aerospaziali*
*Amt für GeoInformationswesen der Bundeswehr*
*Israel Meteorological Service*

*www.cosmo-model.org*

Editor: Massimo Milelli, ARPA Piemonte

# *The COSMO Priority Project INSPECT*

# *Final Report*


***Project participants:***

***A. Bundel***[1]‡, ***F. Gofa***[2]‡, ***D. Alferov***[1], ***E. Astakhova***[1],

***P. Baumann***[3], ***D. Boucouvala***[2], ***U. Damrath***[4], ***P. Eckert***[3],

***A. Kirsanov***[1], ***X. Lapillonne***[3], ***J. Linkowska***[5], ***C. Marsigli***[6],

***A. Montani***[6], ***A. Muraviev***[1], ***E. Oberto***[7], ***M.S. Tesini***[6],

***N. Vela***[7], ***A. Wyszogrodzki***[5], ***M. Zaichenko***[1], ***A. Walser***[3]

‡ Project Coordinator

[1] RHM

[2] HNMS

[3] MCH

[4] DWD

[5] IMGV

[6] ARPAE

[7] ARPAP

# Contents

# 1 Introduction

As numerical weather prediction models began to increase considerably in resolution, it became clear that traditional grid-point-by-grid-point verification methods did not provide material information about forecast performance. Traditional verification scores often indicate poor performance because of the increased small-scale variability. Numerous methods have been proposed in order to assess the value of very-high-resolution forecasts, including spatial verification methods. Availability of radar and merged radar-station observations contributed to the growth of popularity of these methods. Furthermore, the plethora of spatial verification methods has led to the need to analyse how these methods relate to one another, how each method works, what information could be gleaned from each method, and whether any given method actually conveys any useful information or not. The ICP international project and its second phase MesoVICT (Mesoscale Verification Inter-Comparison over Complex Terrain) were initiated to study how these methods provide feedback about the forecast skill through well-structured experiments (http://www.ral.ucar.edu/projects/icp/). The main objectives of MesoVICT international project can be summarized as follows:

- To investigate the ability of existing and newly developed methods to verify fields other than deterministic precipitation forecasts
- To demonstrate the capability of spatial verification methods over complex terrain and gain an understanding of the issues that arise in such cases
- To encourage community participation in the improvement of spatial methods
- To provide the community with a test-bed with common datasets but also to provide assistance in developing and testing these methods

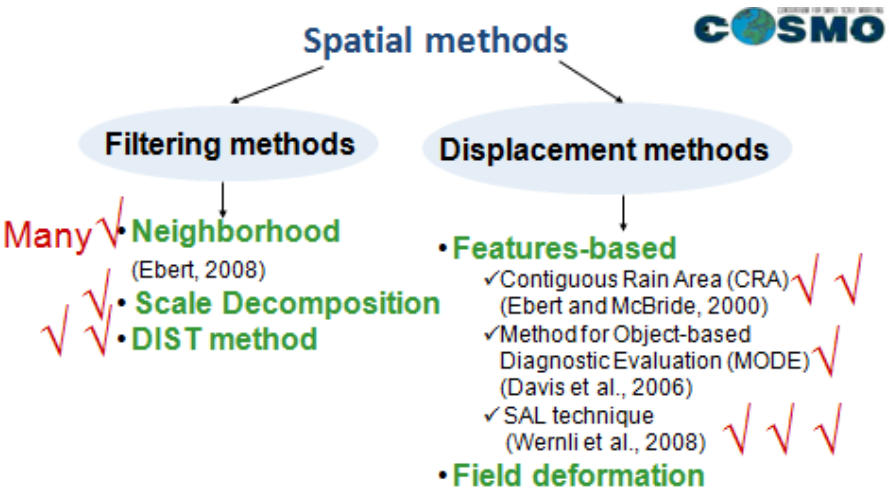



Figure 1: Classification of spatial verification methods and applications in COSMO within the INSPECT project (red signs).

The INSPECT project ran in parallel to MesoVICT to summarize the experience of applying spatial verification methods to COSMO forecast systems of very-high-resolution (1-3 km) compared to high-resolution models, providing criteria for deciding which methods are best suited to particular applications. In addition to targeting the goals of the MesoVICT project, PP-INSPECT provided COSMO users with more choice as to verification domains

and reference data, encouraging the participation of the COSMO community in the testing, development and improvement of spatial verification methods. Fig. 1 represents the classification of spatial verification methods and applications in COSMO within the INSPECT project indicated with red signs.

## 2 The main outcomes of PP INSPECT Summary

In this section, the main outcomes of the project are listed, while in the following paragraphs a detailed description will be given.

- A number of reruns of deterministic and ensemble forecast systems were performed for MesoVICT test cases (ARPAE-SIMC, MCH, RHM) (Table 1). All the reruns are interpolated into the same grid by M. Dorninger (Austria) in order to make easier the comparison, and are available by request at the MesoVICT site (http://www.ral.ucar.edu/projects/icp/).

- Scripts with examples for running neighbourhood, CRA, and SAL methods using free SpatialVx package are available at WG5 repository (HNMS, IMGW-PIB, RHM). Also examples of scripts to adapt radar and satellite data in a gridded format comparable with model output are available. For Neighbourhood (filtering) and SAL (object-based) methods, testing was performed by means of comparison of results from two independent software packages; bugs are reported to SpatialVx developer E. Gilleland (IMGW-PIB, HNMS). For Intensity Scale (IS) method (scale separation), MET software (NCAR) was also tested in addition to SpatialVx, but the graphical outputs of the packages were different and not comparable with those of SpatialVx. Further investigation is required for the differences in the application of the method.

- Several ways of compact visualization of neighbourhood, CRA and SAL methods were proposed (DWD, MCH, RHM). Especially for neighbourhood scores, such a cumulative framework can be implemented as part of the Common Verification activity.

- An object-based SAL (Structure-Amplitude-Location) method was found easier to implement than MODE and CRA methods as it does not require pair-wise matching of observed and forecast objects. The results however must be considered carefully as the method considers average characteristics over a domain. The accumulation precipitation range should not be small. It is recommended to calculate the SAL parameters with 6h precipitation range and higher, unless a highly convective case with significant precipitation amounts is tested. Object-based MODE and CRA methods provide more information compared to SAL about individual features of precipitation field, as they estimate matched pairs of observed and forecast objects. Smoothing of the fields can lead to the creation of bridges among different objects that are close to each other and unify them. This can significantly change the results of all object-based methods. On the other hand, smoothing can be undesirable for estimating objects with intense precipitation (or other variable of interest). The option to discard objects that are smaller than a certain size is found useful. When applied to observations, it eliminates small objects of any intensity that can be erroneous noise. Option for splitting objects is desirable sometimes. (HNMS, RHM, IMGW-PIB, ARPAE-SIMC).

- For MODE and CRA methods, it was found not feasible to identify one optimal universal matching function, in particular for high-resolution fields with objects of complex shape. For lower precipitation thresholds (and, consequently, wider features), more

| Institution | Forecast system |
|---|---|
| MCH (P. Baumann) | COSMO-1 reruns for ALL MesoVICT cases are done and available at WG5 repository |
| ARPAE-SIMC (A. Montani) | ECMWF-IFS reruns (51 member) for cases 1 and 2 (8 initial dates) |
| MCH (A. Walser) | COSMO-E reruns (21 member) for cases 1 and 2 (8 initial dates) |
| RHM (D. Alferov and E. Astakhova) | COSMO-Ru2-EPS (51 member) for case 1 (1 initial date) and case 2 (1 initial date) |

Table 1: Reruns of forecast system for MesoVICT cases.

reasonable results were obtained for a criterion based on matching observed and forecast objects if the distance between them is less than their average size. For higher thresholds, this criterion often does not allow for successful matches due to the small areas of features. A minimum boundary separation criterion seems more promising for intense events, but with a minimum boundary separation distance, beyond which, features should not be matched (RHM, IMGW-PIB).

- The IS scale separation method allows for the skill to be diagnosed as a function of the scale of the forecast error and intensity of the precipitation events. Results show that reduction of skill is mainly due to the small-scale misplacement of more intense (rare) precipitation events. Wavelet-based (Hier wavelets) scale-separation statistics are suitable for comparing models with different resolutions as the reference forecast accounts for the forecast variability. The method allows the analysis of precipitation instances, but it is not able to provide generalized information on the relative long term performance of a modelling system based on aggregated data.

- Applications of DIST, SAL and CRA methods to ensembles were made and new approaches on summarizing performance over various members and time accumulations were proposed.

- First results of experiments on introducing observation uncertainty into the spatial methods are given.

What was not completely fulfilled within the project and needs to be further studied:

- Introducing orography factor explicitly.

- Wind characteristics were only partially explored by M.S. Tesini (ARPAE-SIMC). The upscaling DIST method was applied, which analyses the statistical parameters of values in boxes of increasing size. For wind speed, the representative value of the box was defined as the median exceeding a predefined threshold; for wind direction, as the most populated category after having binned the direction in 8 categories. First results on DIST application to wind, were found not very satisfactory. The possible reasons were identified as follows:

  - possibly, the representative value of the box could be defined in another way;
  - the verification period was very short;
  - wind is too localized and the aggregation has benefit only if the boxes were chosen based on a different criteria that those of precipitation;
  - taking into consideration the orography (valley,) is required.

- Spatial applications to wind and other variables (e.g. Cloud cover) besides precipitation should be further developed.
- Only a limited number of applications to ensembles were performed. This direction of spatial method application is much demanded at present, and should be developed.
- Introducing observation uncertainty in the analyses was initiated by D. Boucouvala (HNMS) (see Sec. 4.1.3). This topic requires further study.
- Processing large amounts of data (for example nowcasting forecasts verification) will be of increased demand in the future. In Sec. 4.3.3, we describe an RHM experience of assessing the nowcasting scheme during the summer 2017 using the CRA method. The computational efficiency and the optimal organization of the data will be of utmost importance for processing large amounts of data.

Moreover, the PP INSPECT supported the efforts to increase COSMO visibility worldwide due to the active participation of COSMO members in the international project MesoVICT and its activities as well as in the WMO-JWGFVR workshops. Finally, it should be noted that the project was focused on the comparison of the spatial methods, and not on assessing the model performance, although some reflections to the advantages of different forecast systems were made.

# 3 Filtering methods

## 3.1 Neighbourhood Methods (Flora Gofa, HNMS)

### 3.1.1 Method applied (related to an INSPECT Task) and objectives

Neighbourhood verification is based on the principle of expanding the field of view to nearby (neighbours) data points in space, employing a spatial window around the forecast and observed points. In this way, the penalty for differences between modelled and observed values is relaxed. The size of this neighbourhood can be modified in order to verify model performance at several different scales, thus providing insight into the scales the model has the most skill at. Since the size of the neighbourhood can be varied, such approaches are well suited to verification of high-resolution models (Casati et al., 2008). In addition to neighbourhood size, the type and degree of filtering applied can also be modified. For example, extreme values may not be filtered out when examining severe weather events. Furthermore, in most cases, as in this research, the filter is applied to both the forecast and the observations. A more detailed review of neighbourhood approaches is available in Ebert (2008).

### 3.1.2 Short description of the dataset (forecast-observation data), adaptation required, software for the method application

The datasets used in the context of this study have been obtained from the Mesoscale Verication Inter-Comparison over Complex Terrain (MesoVICT) project. MesoVICT has been established in order to facilitate the application, capability and enhancement of spatial methods both for deterministic and ensemble forecasts (Dorninger, 2013). MesoVICT benefits from a huge data collection effort within the framework of the World Weather Research

Programme (WWRP) Project: Mesoscale Alpine Programme (MAP) D-PHASE over Central Europe in 2007. All six available test cases were analysed that cover a wide range of meteorological phenomena in and around the Alps. The example that is presented here is for the period 20-22 June 2007. With respect to the synoptic situation, the region of interest is ahead of a trough located over the British Isles, and warm moist air is being advected towards the Alpine Region. This leads to strong convective events on the evening of 20 June in the area north of the main mountain range. Subsequently, a cold front reaches the Alps from the west and moves to the east rather quickly while convective events are again observed ahead of the front (Fig. 2).

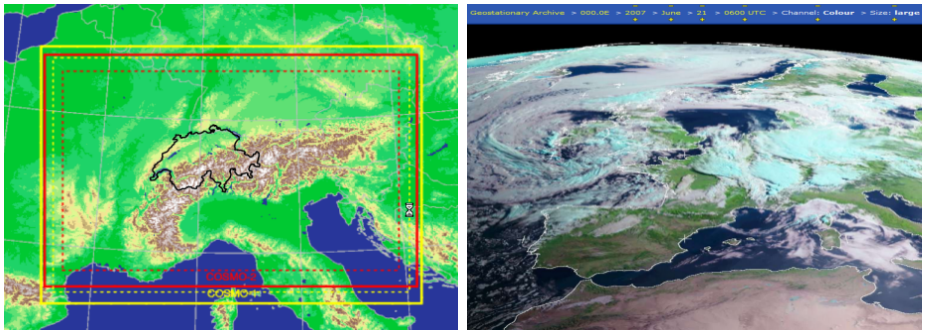


Figure 2: Left: COSMO-1 (dashed yellow), COSMO-2 (dashed red) integration domains, Right: Satellite image, cloudiness on 21.06.2007, 00UTC.

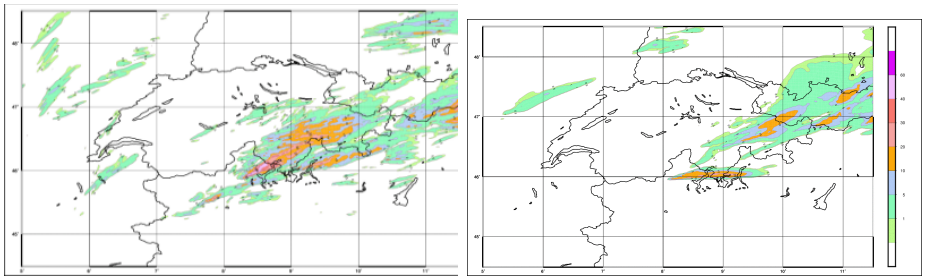


Figure 3: 12h accumulated precipitation at 21.06.08: 18UTC Left: COSMO-1, Right: COSMO-2.

From the multitude of NWP models with varying resolutions that are available through the MAP-D Phase, forecasts derived from COSMO-2 ( 2km res.) and COSMO-1 ( 1km res.) of the Swiss Meteorological Service were used. The domain of the COSMO model is shown in Fig. 3. Both models were nested on a coarser 7km COSMO model. Hourly precipitation forecasts from both models were upscaled prior to the application of neighbourhood methods in order to match the resolution of the observation fields in a final 8km grid spacing to match the observations.

### 3.1.3 Verification software

For the purposes of this research, the VAST (VERSUS Additional Statistical Techniques) software package, which was developed by the COSMO consortium and is based on Beth Eberts fuzzy verification IDL (Interface Design Language) code, was employed (Vela, 2017). VAST offers a number of neighbourhood verification tools, of which the following were tested

in the context of this research: upscaling, anywhere in the window, minimum coverage, fuzzy logic and joint probability. What distinguishes one method from another is the decision model, i.e. how strict or relaxed are the criterion that determine whether a forecast is successful or not. A large variety of methods and scores were calculated for a number of thresholds and window sizes, but only a selection of representative results are presented here.

### 3.1.4 Main findings (plots and explanation)

In this section, a short evaluation of the results of the application of some neighbourhood methods (Fig. 4) is given. Fractions skill score. The decision model is: "A forecast is useful if the frequency of forecast events is similar to the frequency of observed events", giving information on the spatial scales that the forecast resembles the observations, ranging from 0 to 1 (perfect). The score is most sensitive to rare events (e.g., small areas of precipitation); the useful scales are indicated in bold (Fig. 4). The FSS values for both COSMO forecasts are greater for light rain thresholds and larger scales, with useful skill displayed at spatial scales of around 130 km or larger for light rain and not at all for the heaviest rainfall rates. COSMO-1 forecasts exhibit similar behaviour with COSMO-2, being more useful for slightly smaller spatial scales ( 70km) for small to medium rainfall thresholds, while for higher thresholds COSMO-2 seems to be slightly more useful than COSMO-1.

Pragmatic approach. Instead of verifying the forecast probability within a neighbourhood against the observations within that same neighbourhood, the forecast is verified against the observed value in the central grid box. The decision model is that a useful forecast has a high probability of detecting events and non-events supported with the use of Brier skill score (BSS) to quantify the forecast success. The limited application of this method over the test case in Alps indicated that there is only minor improvement of the forecasts versus the reference, which is the sample climatology of the observations over the whole domain.

Upscaling method. It is the most widely used neighbourhood verification technique. Forecasts and observations are simply averaged to increasingly larger grid scales for comparison using a range of standard statistics. The decision model is that a good forecast has approximately the same mean rainfall amount as the observations. For this particular case, the ETS (equitable threat score) was calculated for each window size/threshold and, as shown, the scores generally improve with increasing scale and smaller rainfall thresholds, while the relative advantage of COSMO-1 forecasts is demonstrated.

Fuzzy logic approach. It is based on the fact that a forecast has a certain likelihood of being an event and a certain likelihood of being a non-event, and this is also the case for the observations. This likelihood is called the weight of support. This means that a forecast can be somewhat correct and somewhat incorrect at the same time (window). The decision model is, a forecast is useful if it is more correct than incorrect. From the categorical scores, BIAS is calculated in order to provide a more precise picture of the scales and thresholds that each model overestimates (BIAS $> 1$) or underestimates (BIAS $< 1$) hourly precipitation amounts. Other neighbourhood methods and scores for each method were calculated in an effort to explore the variability in the information that can be provided from their application to two different modelling systems which have not been included in this paper.

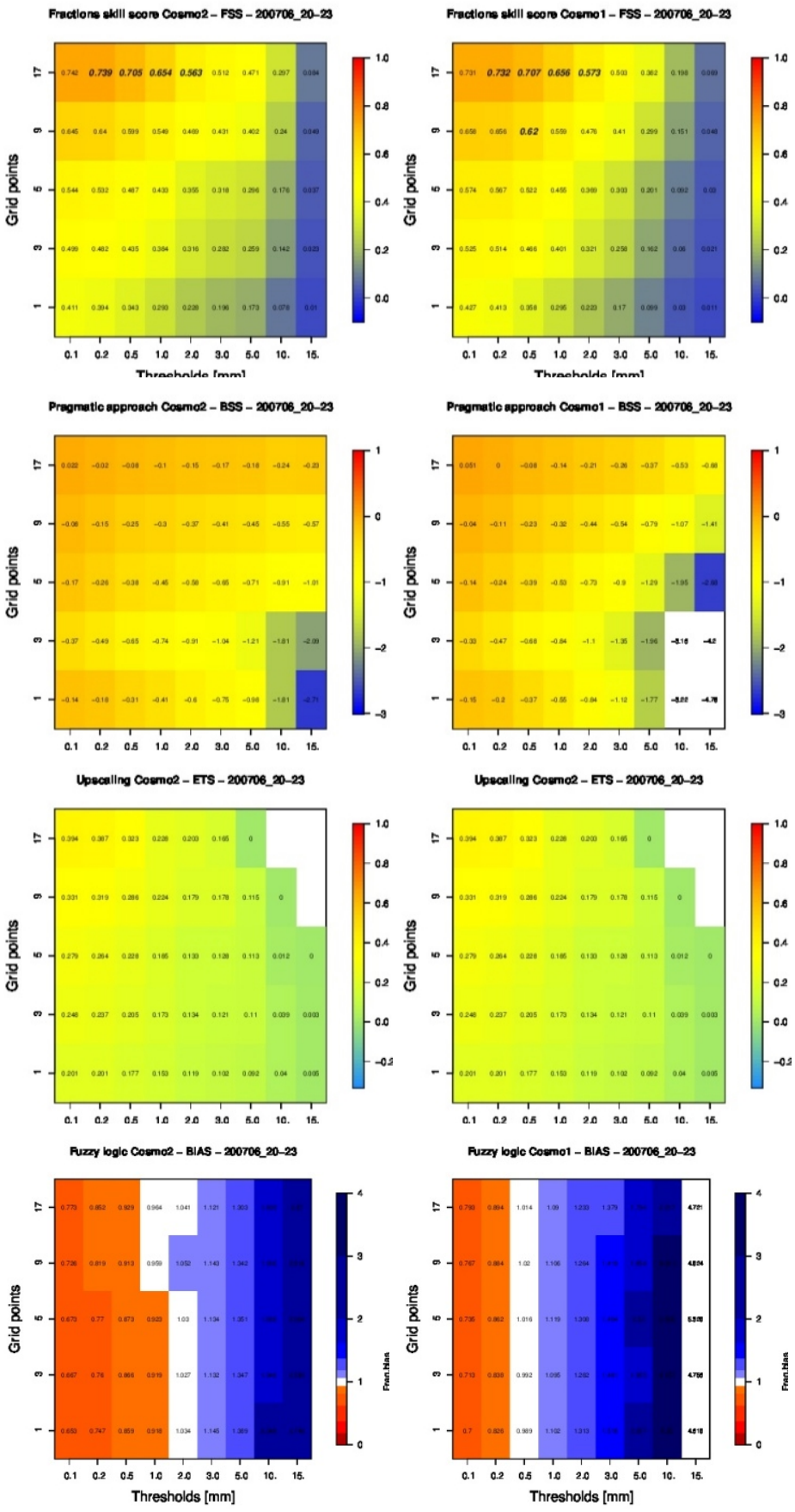



Figure 4: Fraction Skill Score (FSS) (first row), Pragmatic Approach (BSS) (second row), Upscaling (ETS) (third row), and Fuzzy Logic (BIAS) (fourth row) for COSMO-1 and COSMO-2.

#### 3.1.5 Characteristics of the method applied)

- Efficiency in calculation time: the efficiency in time is quite satisfactory with respect to VAST. A significant effort is required to adapt the datasets and make them identical in size (grid dimension and resolution) prior to the application of VAST. If the user keeps the default amount of space windows (5) the calculation is rather fast, while it increases dramatically with the addition of even one additional space window. The adjustment of thresholds is easier and faster. No application was performed on time windows so far.

- Ability to deal with different density of observations: the method requires a complete set of gridded observations and forecasts with no gaps in the domain of verification. In case of inhomogeneous density of observations, the user has to adapt the dataset accordingly.

- Stability against observation errors: neighbourhood methods do not account for observation errors, as by relaxing the requirement of exact match in point and the use of averaged values over space windows for both forecast and observations, the weight of such errors on the verification results is eliminated.

- Ability to assess the added value of high-resolution models: neighbourhood methods provide a tool to compare modelling systems of various resolutions and are particularly valuable in the case of high resolution forecasts. However, before deciding on the methodology or score that is more suitable for a model evaluation, the first step in this approach involves carefully defining the attributes of a good forecast and subsequently identifying the specific methods and their associated decision models best suited to the particular application. Neighbourhood verification is particularly valuable in the case of high resolution forecasts, providing useful feedback on the scale and intensity for which each model configuration is advantageous. Precipitation events on different spatial scales are produced by different physical processes (e.g. large-scale frontal systems or small-scale convective events). Verification at different spatial scales provides greater insight into model performance to simulate these different processes.

Although the value of a neighbourhood verification framework has been demonstrated for this particular test case, its most sensible use is for evaluating sets of forecasts to determine typical forecast performance. For example, it can be used to monitor monthly or seasonal forecast performance in a region. It can also be used to evaluate updated versions of a model to identify at which scales and intensities, if any, skill has been improved (Gofa, 2017).

### 3.2 Analysis of long time series of neighbourhood scores (in particular: FSS, upscaling with ETS and FBI) for precipitation. Further investigation into the most informative and compact representation of scores.

#### 3.2.1 DWD experience (U. Damrath)

At DWD, the spatial methods have been developed during many years. This section describes the analysis of long time series of the following scores: the Equitable Threat Score (ETS) for upscaling and the Fractions skill score (FSS). In (INTERP 2009), these scores were identified among the most useful. The analysis scheme is as follows:

- Getting the ETS for upscaling and FSS as monthly values from fuzzy verification

- No averaging over daily values is applied, but the calculation of scores from the contingency tables of the whole month is performed
- Calculation of running means of the results over one year
- Presentation of mean values and mean averages

Data sources that were used in this study:

- Precipitation forecasts of German COSMO-models and GME (with March 2015 ICON)
- Precipitation observations from radar data

Grid sizes and thresholds:

- Grid size from 0.025 (resolution of COSMO-DE) to 1.625 (65*resolution of COSMO-DE). For COSMO-EU the values are taken for the whole grid cell that is in this CDE-grid cell. Thus, there are about 9 points with equal values for the lowest window size
- Thresholds: 0.1, 0.2, 0.5, 1, 2, 5, 10, 20, 50 mm/12h or mm/24h

Fig. 5 gives an example of score visualization for the whole German territory. The red indicates that COSMO-EU is better, and the blue, that COSMO-DE. In this case, the higher resolution model advantage becomes evident only for the largest windows.

Overall, the results indicate that we are not able to make real forecasts for the lowest window size. A coarse value from the model with lower resolution has no double penalty.

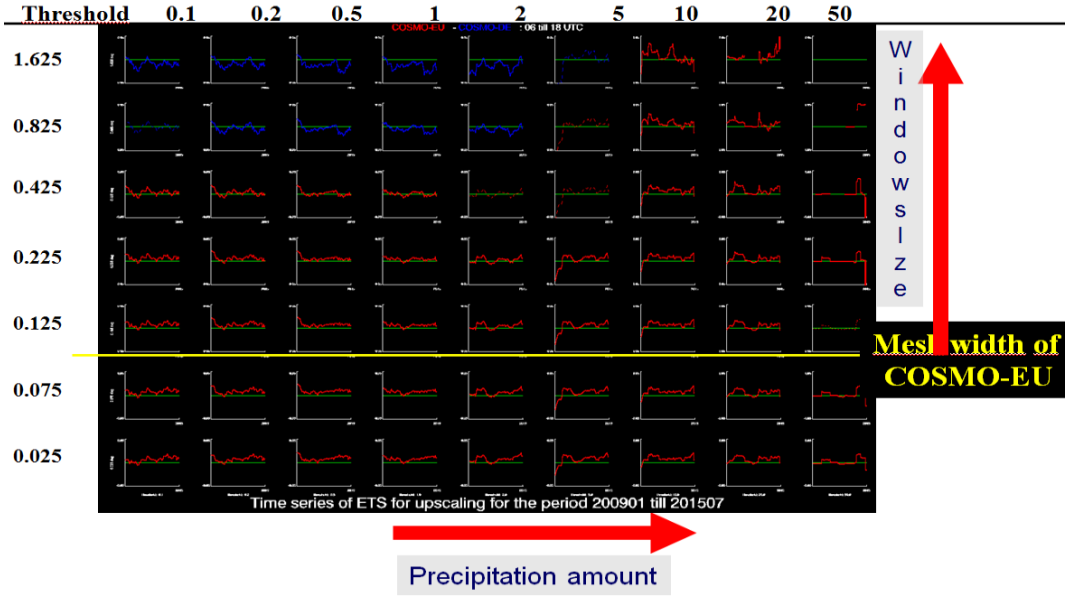



Figure 5: Comparison of COSMO-EU to COSMO-DE Differences of upscaling ETS. 06-18 UTC.

### 3.2.2 Evaluation of 4dVerif spatial verification visualization (X. Lapillonne, MeteoSwiss)

The use of spatial verification can bring new insight to the verification of model variables. It has a particular added value when it comes to verification of precipitation field since it can overcome the so called double penalty effect which occurs when a precipitation happen at a wrong geographical location in the model as compare to the observation. Because of its two dimensional structure it provides additional information as compared to stations based verification which needs to be visualized. In this report section we discuss some aspect of the visualization of these verification results and in particular how this information could be reduced so as to be easier to interpret.

**Fraction skill score with 4dverif at Meteoswiss for precipitation verification**

We focus here on the fraction skill score obtained with the 4dverif software run at MeteoSwiss. The fraction skill score (Roberts N.M., 2008) is one of possible available surface verification scheme. The basic idea of this score is to define an area or scale (here squares) and to count events in this region (fraction area) both in the model and in the observation. One then obtains a score for different scale see Fig. 6 and Fig. 7. A so called useful scale, see (Roberts N.M., 2008), is also defined above a certain score (in bold in Fig. 7).

The potential scale information introduces a new dimension in the verification parameter space, which is coming in addition to other existing parameters such as lead-time and initial forecast time. This provides potentially very detailed insight in the data which might be very useful for research purpose or to better understand the model but may be overwhelming for standard verification purpose, see Fig. 8.

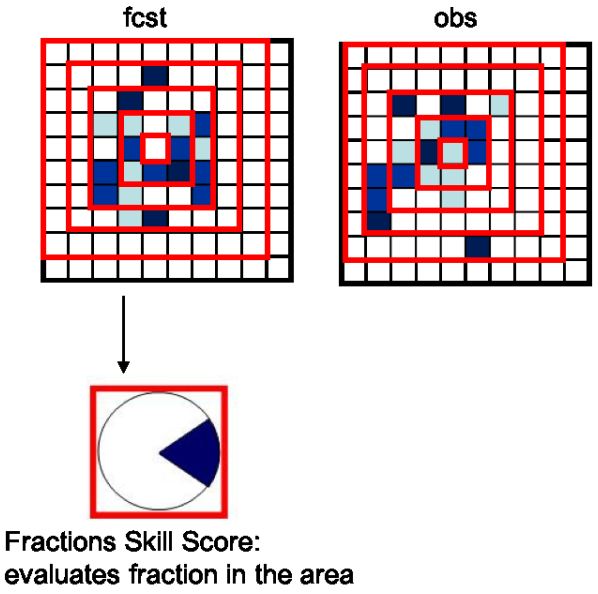



Figure 6: Surface evaluation area.

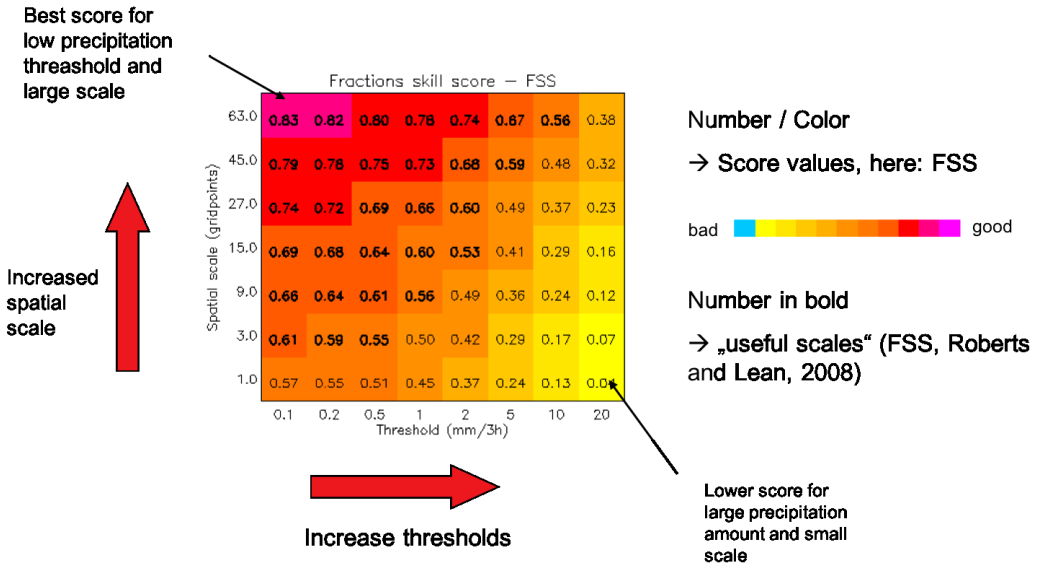



Figure 7: Fraction skill score (FSS).

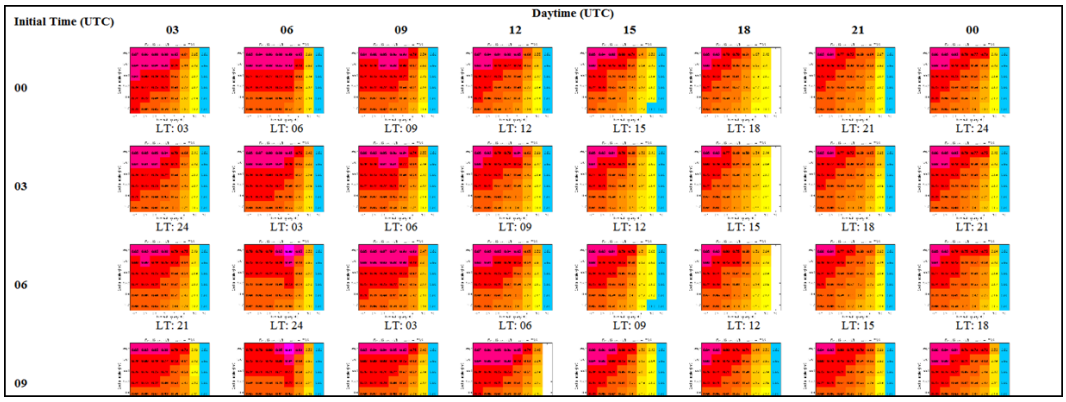



Figure 8: FSS for all lead times, initial time, scale and scores.

### Simplified visualisation

It is suggested to reduce the available information by showing the most relevant part for standard verification on a one dimension plot which resembles usual station based verification or in a table. For the considered score and variable we propose for example 2 different cross sections of the data. The first one is to select a single meaningful scale and to show for selected threshold the score as a function of leadtime. In Fig. 9, we show the FSS for two models for the scale 19.6 km which roughly corresponds to the size of a warning region in Switzerland.

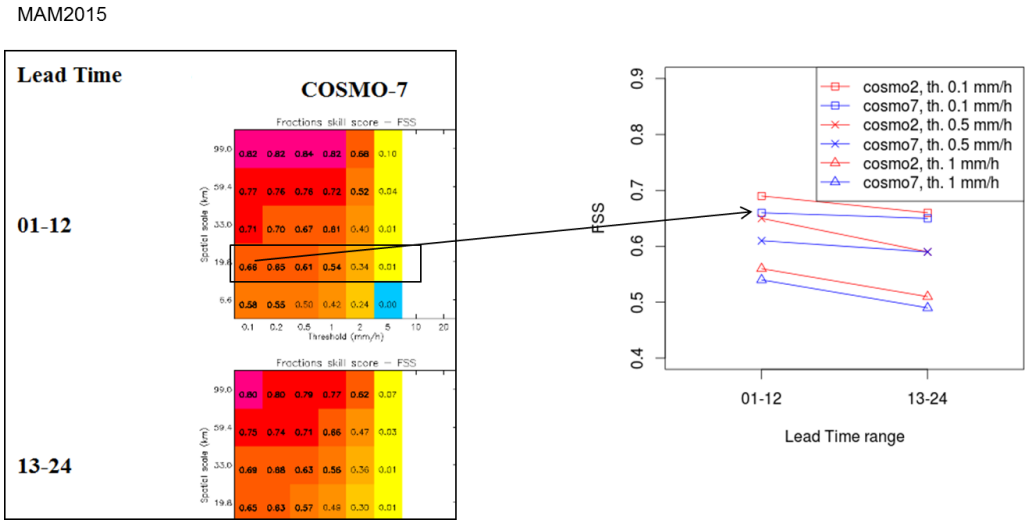



Figure 9: FSS for a selected spatial scale (19.6 km).

The second proposed way of displaying the data is to focus on the useful scale information. In Fig. 10, we show the useful scale as a function of lead time for different threshold. This could be interesting for example to help saying which model may be use for warning. Note that this information has a strong seasonal variability.

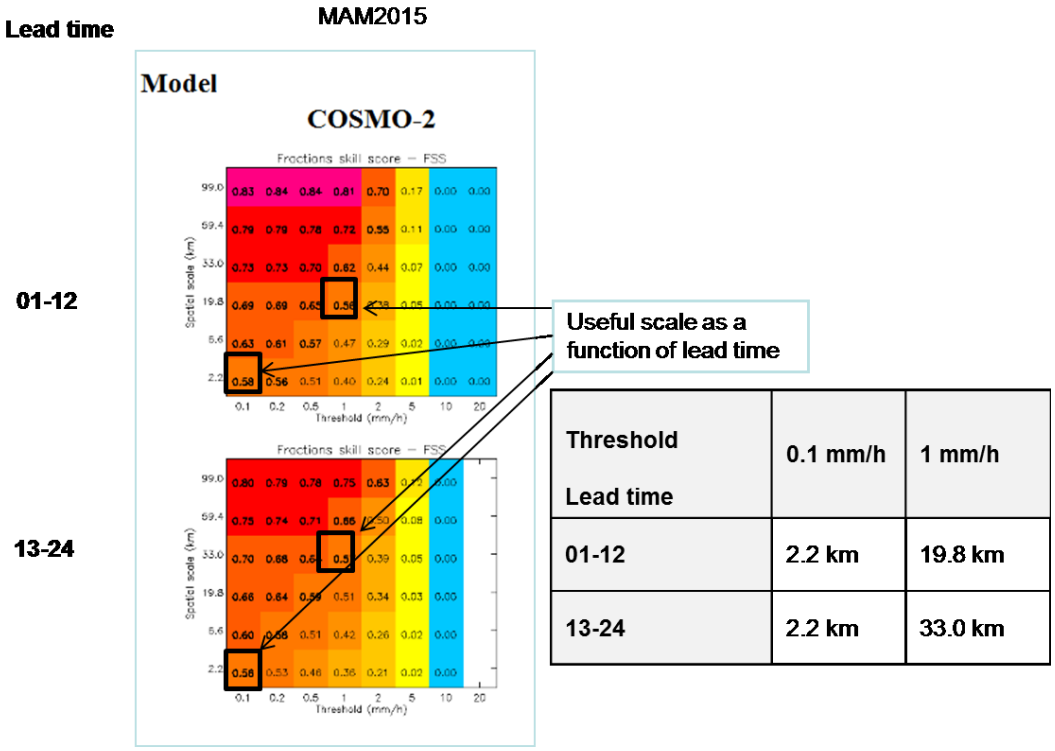



Figure 10: Useful scale as a function of lead time for different threshold.

## Conclusions

The spatial verification provides new insight and information in comparison to standard station based verification. This additional information may be for some usage overwhelming and some reduction this information may be necessary. Here we propose for the fraction skill score two different potential ways of displaying the score. The first one is to look at the score for a given spatial scale while the other one focus on the so-called useful scale for a given lead-time and threshold.

## 3.3 Wind verification with DIST method: preliminary results (M. S. Tesini, ARPAE-SIMC)

### 3.3.1 Motivation of the study

Current wind forecast verification at ARPAE-SIMC is not completely satisfactory. It provides too local results difficult to summarize for a large number of stations. The DIST methodology (Marsigli et al., 2008) used operationally to verify precipitation has pointed out some advantages with interesting results. DIST provides the scores for different distribution parameters calculated for model and observed variables in boxes of increasing size. One of its main advantages is that it can be performed using both sparse point observations and gridded observation against gridded forecast (even if the grids are different). The size and even the shape of the box can be freely defined (e.g. alert areas for hydrological purposes). It provides simple information to forecaster or hydrologist about the performances of models in a single area of interest (e.g. Alert Area) or over the whole model domain aggregating the results of all boxes. The MesoVICT project encourages the investigation of the ability of existing or newly developed spatial verification methods to verify fields other than deterministic precipitation forecasts, e.g., wind forecasts.

### 3.3.2 Representative value of the box for wind characteristics and verification setup

It is important to define the representative value of the box for DIST application. Thinking to a more user-oriented verification we considered the median (e.g. the value below (or above) which 50% of the data may be found) and the 90th percentile (e.g. the value below which 90% of the data may be found, or above which 10% of data may be found) for wind speed. For wind direction, as a first step the values were binned into 8 category (N, NE, E, SE, S, SW,W,NW). Then the most populated category was taken as representative for the direction in the box. Since the direction for light wind may not be significant, another representative value has been evaluated considering only the direction for wind with intensity $> 3$ m/s. All the values of 3 consecutive hourly forecasts (and observations) that belong to the same area are put together to account for timing errors. The study was performed for all three MesoVICT cases (Dorninger, M.et al., 2013). The Swiss COSMO-2 and COSMO-1 model data were used. VERA gridded analysis was used as a reference data (Dorninger, M.et al., 2013).

### 3.3.3 First results

Fig. 11 represents three scores for events wind speed median exceeding a threshold. There is no clear increase of the forecast quality with increasing box size. In Fig. 12, the performance diagram is displayed for all three MesoVICT cases. In general, enlarging the box increases BIAS, but POD and FAR are slightly better. The scores move more or less linearly from 1

to 25 points in the box, but not from 100 to 400 pts. It is unclear if this behaviour depends on some scale of the considered phenomena or the data are not enough to produce consistent statistics. More investigation is needed to understand that. In Fig. 13, Wind speed threat score (TS) for COSMO-2 and COSMO-1 are given for boxes of increasing size. The event is defined here as 10% of points exceeding a predefined threshold. Scores (and trend of scores) considering boxes are nearly the same for the two models. COSMO-1 nearest grid point performs better than aggregation on 8 Km box. It is unclear if this can be explained with wind field characteristic or it is only an unlucky case. Maybe the choice of this percentile is not useful as was thought initially. Fig. 14 gives the PSS for wind direction for COSMO-2 and COSMO-1 for boxes of increasing size. The representative value is defined using all the data and considering only the direction for wind with intensity $> 3$ m/s. At larger scales, all the local information is filtered out. For boxes of smaller size, the information about local winds should be predominant. Another aspect is that it is not completely evident that VERA analysis is able to reproduce very local features.

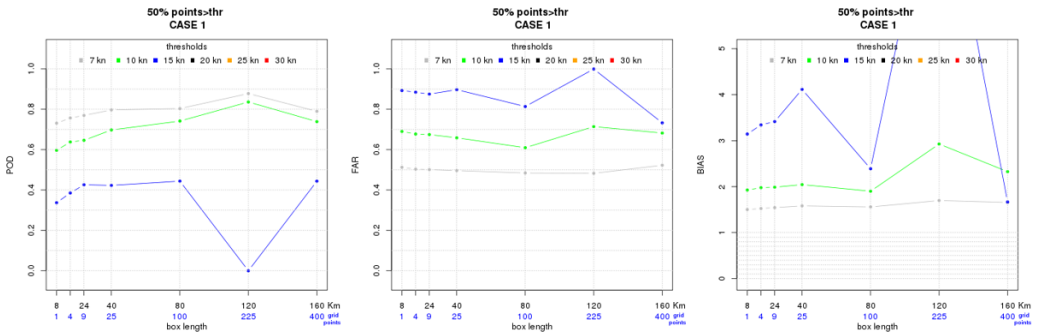



Figure 11: Probability of detection, False alarm rate, and frequency bias for wind speed (in knots), MesoVICT case 1, Cosmo-2 model. The event is defined as median exceeding a predefined threshold. The scores are plotted as a function of the box dimension.

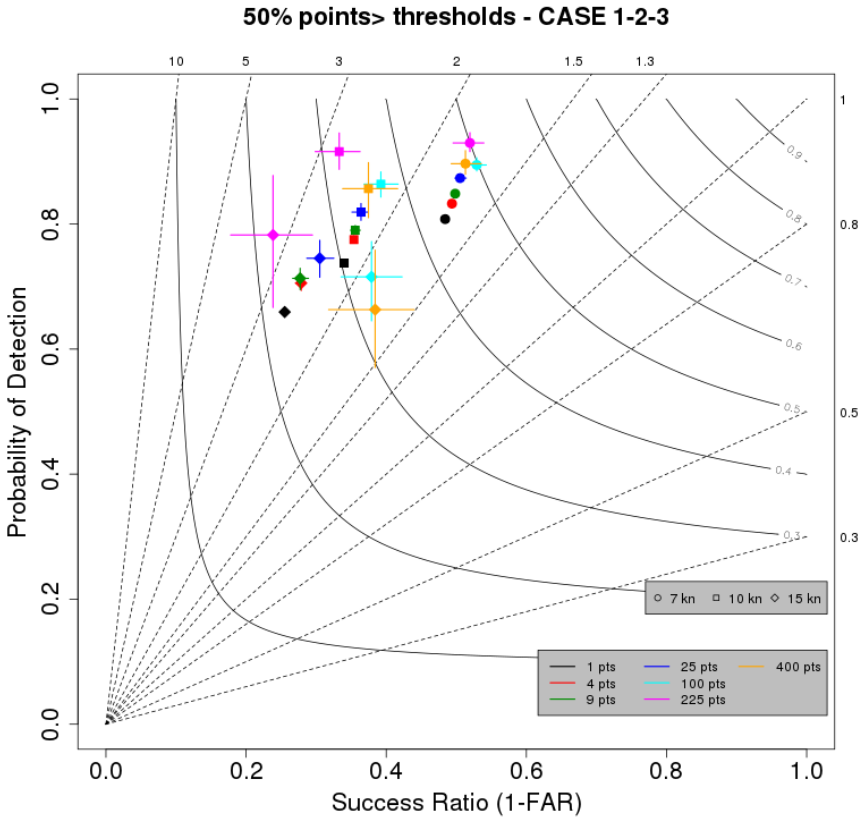



Figure 12: Performance diagram for wind speed (in knots), MesoVICT cases 1. 2. and 3, Cosmo-2 model. The event is defined as median exceeding a predefined threshold.

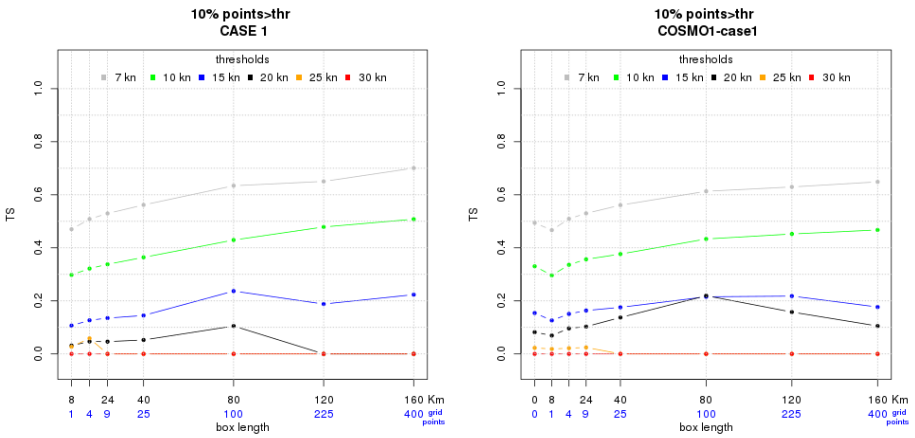



Figure 13: Wind speed threat score (TS) for COSMO-2 (a) and COSMO-1 (b) for boxes of increasing size. The event is 10% of points exceeding a predefined threshold.

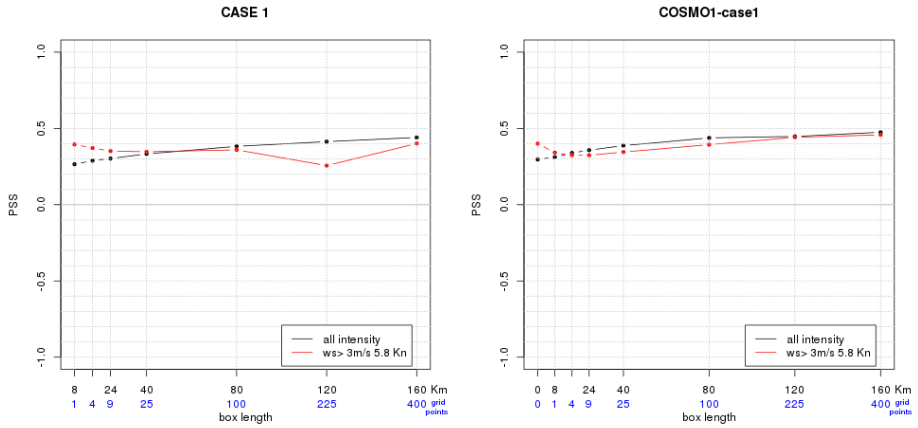



Figure 14: Wind direction, MesoVICT case 1, Cosmo-2 (a) and Cosmo-1 (b), PSS (Peirces skill score: What was the accuracy of the forecast in predicting the correct category, relative to that of random chance).

### 3.3.4 Conclusions

The first results on DIST application to wind were found not very satisfactory. The possible reasons were identified as follows:

- possibly, the representative value of the box could be defined in another way;
- the verification period was very short;
- wind is too local and the aggregation has is benefit only if the box were chosen differently;
- taking into account the orography (e.g., wind in valleys) is needed.

But before giving up tests are needed for other MesoVICT cases and using the JDC original observation (one of the main advantages of DIST was to deal with sparse point observations). It is also important to look at the geographical distribution of the scores.

## 3.4 Intensity-scale method (F. Gofa, HNMS)

### 3.4.1 Method applied (related to an INSPECT Task) and objectives

Scale Separation Methods: in general with these methods you decompose forecast and observation fields into the sum of spatial components on different scales by using spatial filters, and then you perform the verification on each scale component, separately. Verification on different scales can provide useful insight into NWP model representation of the different physical processes associated with phenomena on different scales. Scale-verification approaches aim to assess quality and skill of the forecasts for different spatial scales, analyse the scale-dependency of the forecast predictability (e.g. evaluate the no skill  skill transition scale), and assess the forecast ability to reproduce scale spatial structure of observed precipitation fields. Precipitation fields are characterized by the presence of features on different spatial scales triggered by different physical phenomena. As an example, events such as

frontal systems are driven by the mesoscale dynamics of the atmosphere, whereas smaller scale events such as showers can be developed by local convective motions. Verification on different spatial scales can provide a deeper insight in the model performance at simulating different dynamics and give useful feedback for improvements. The intensity-scale technique for verifying spatial precipitation forecasts was tested in the framework of PP INSPECT. The technique provides a way of evaluating the forecast skill as a function of intensity of the precipitation rate and spatial scale of the error. The forecasts are assessed using the MSE skill score of binary images, obtained from the recalibrated forecasts and analyses by thresholding at different precipitation rate intensities. The skill score is decomposed on different spatial scales using a two-dimensional discrete Haar wavelet decomposition of the binary error images. Wavelets are functions characterized by a location and a scale (Daubechies, 1992). Similar to Fourier transforms, wavelets can be used to represent functions on different spatial scales, and so can be used to investigate scale properties of physical phenomena. Because of their local properties, wavelets are more suitable than Fourier series for representing spatially discontinuous fields such as precipitation. Moreover, because of their locality, wavelets are more efficient than Fourier components at representing sparse images containing few non-zero values. Different types of wavelets exist. Each wavelet type is defined by a mother and a father wavelet, characterized by different shapes and mathematical properties (e.g. smoothness, symmetry, etc.). In this study, Haar wavelets are used, because of their square shape which best captures the difference in binary variables. Fig. 15 shows the one- and two- dimensional Haar wavelets. Note that the two-dimensional wavelets are generated simply as the Cartesian product of the one dimensional wavelets.

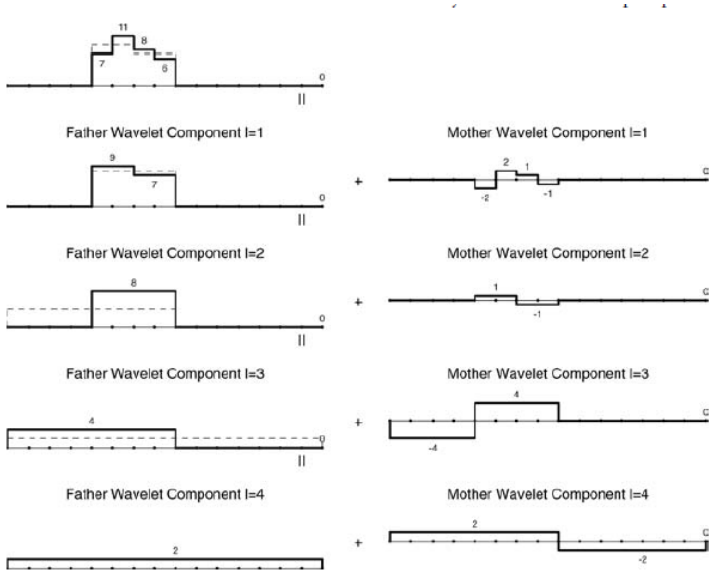



**Figure 11.** *Example of the one-dimensional discrete Haar wavelet filter applied to an example function (top left panel). At the first step the function is decomposed into the sum of a coarser* mean *function (the first father wavelet component) and a* variation-about-the-mean *function (the first mother wavelet component). At each step the Haar wavelet filter decomposes the father wavelet component obtained from the previous step into the sum of a coarser* mean *function (the $l^{th}$ father wavelet component) and a* variation-about-the-mean *function (the $l^{th}$ mother wavelet component). The $l^{th}$ father wavelet component is obtained from the initial function by a spatial averaging over $2^l$ pixels. The process stops when the largest father wavelet component (mean over the whole domain) is found.*

Figure 15: One- and two- dimensional Haar wavelets.

Steps followed in Intensity Scale method application:

- Binary error decomposition: Thresholding is used to convert the forecast and analysis into binary images for each of the thresholds. Binary error is the difference of this:

$Z = I_{y'} - I_x$

- Binary error image is then expressed as the sum of components on different spatial scales by performing a 2-dimensional discrete Haer wavelet decomposition

$$Z = \sum_{l=1}^{L} Z_l$$

- Most substantial binary error image of the mother wavelet components are calculated for various spatial scales (l =1,..L=7 that corresponds to X km). The spatial scales refer to the spatial scale of the error and not that of the precipitation features or their displacement as it happens in the neighbourhood methods

- The MSE of the binary error image is calculated from:

$$MSE = \sum_{l=1}^{L} MSE_l \qquad MSE_l = \overline{Z_l}^2$$

  while for each threshold the skill score can be calculated from:

$$SS = \frac{MSE - MSE_{random}}{MSE_{best} - MSE_{random}}$$

  where and $MSE_{random}$ is associated with a random forecast calculated from the bias and the base rate at each threshold

- Intensity scale verification technique is a spatial generalization of traditional binary verification (HSS, PSS).

### 3.4.2 Short description of the dataset (forecast-observation data), adaptation required, software for the method application

The datasets used in the context of this study have been obtained from the Mesoscale Verication Inter-Comparison over Complex Terrain (MesoVICT) project. All six available test cases were that analysed, cover a wide range of meteorological phenomena in and around the Alps. The example that is presented here is for the period 20-22 June 2007. From the multitude of NWP models with varying resolutions that are available through the MAP-D Phase, forecasts derived from COSMO-2 ( 2km res.) and COSMO-1 ( 1km res.) of the Swiss Meteorological Service were used. The domain of the COSMO model is shown in figure above. Both models were nested on a coarser 7km COSMO model. Hourly precipitation forecasts from both models were upscaled prior to the application in order to match the resolution of the observation fields in a final 8km grid spacing to match the observations. The software used was the R based SpatialVx and the waveIS routine. MET software (NCAR) was also tested but the graphical outputs of the methods were different and not comparable with those of SpatialVx. Further investigation is required for the differences in the application of the method.

### 3.4.3 Main findings (plots and explanation)

An example of the verification analysis is presented in Fig. 16 for 20070621-15UTC.

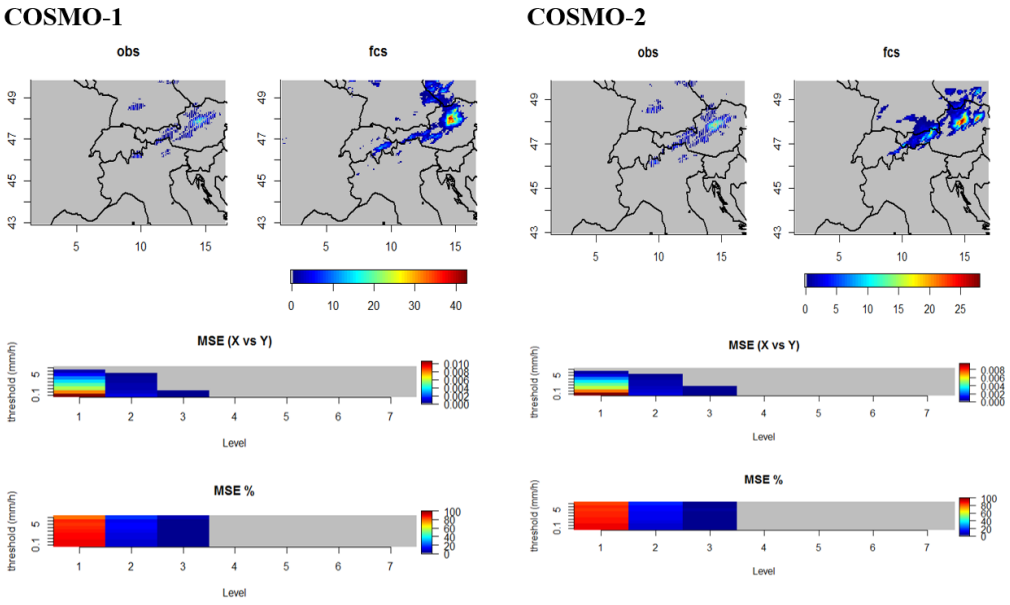



Figure 16: Example of the verification analysis for 20070621-15UTC.

From the analysis of this precipitation instant, it can be deducted that MSE almost disappears for error tiling of the order of 3x3 grid points. Comparing the two model performances, COSMO-1 shows slightly improved behaviour compared to COSMO-2. In Fig. 17, the results from the Skill score are given.

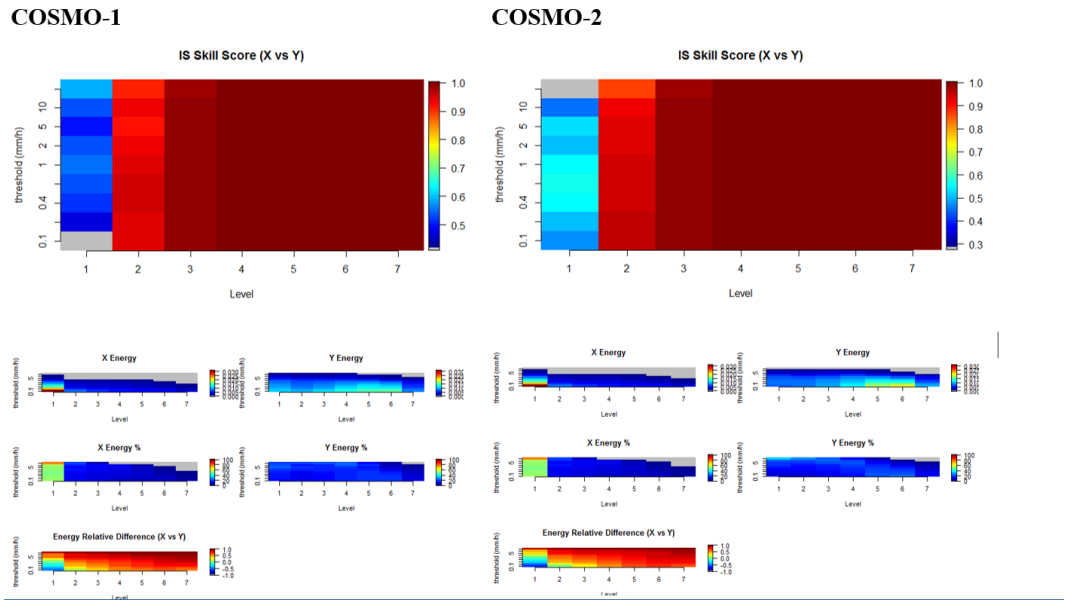



Figure 17: Results from the Skill score.

Small error scales (l=1) have skill close to zero as it is shown in the graphs, while slightly large scales exhibit large skill. COSMO1 ISS graphs exhibit that errors due to displacement of small spatial scale features are more important compared to those of COSMO-2. As a summary, both models have **small skills at the smallest scale** but skill improves when

considering larger spatial scales. ISS decreases as the precipitation threshold is increased and this is due to the **poor ability of the models to go beyond just the yes/no rain discrimination**.

#### 3.4.4 Characteristics of the method applied

This filtering method allows the skill to be diagnosed as a function of the scale of the forecast error and intensity of the precipitation events. Results show that reduction of skill is mainly due to the small-scale misplacement of more intense (rare) precipitation events.

Wavelet-based scale-separation MSE skill score and scale-separation statistics are suitable for comparing models with different resolutions as the reference forecast accounts for the forecast variability. IS method constraints are related to the request to have precipitation analysis available for each grid point of the forecast field, and to the fact that Haar wavelet decomposition is designed for a square domain. Computationally, the method was easy to be applied and to derive numerical and graphical results, while their interpretation is not so straightforward and easy to comprehend from a non-experienced user.

The method allows the analysis of precipitation instances but it is not able to provide a generalized information on the relative long term performance of a modelling system based on aggregated data.

# 4 Object-based methods

## 4.1 SAL Method for deterministic and ensemble precipitation verification at HNMS (D. Boucouvala)

### 4.1.1 Description of the Method

SAL method (Wernli et. al 2008, 2009) is a spatial three component object- based quality measure which quantifies a precipitation forecast in terms of 3 parameters which correspond to a global field measure of Structure (S), Amplitude (A) and Location (L). An object is defined when it exceeds a fixed or statistically defined threshold value. The objects of the observed and forecast fields do not require one-to one matching and are identified separately. The fields are then transformed into binary representations of 1 (grids exceeding the threshold) or 0 (grids not exceeding the threshold).

The **A** component represents a normalized difference between the domain-averaged forecast D($R_{mod}$) and observation fields D($R_{obs}$) and it is the only one that **is independent of the identification of features** as it depends on the total precipitation amount. A positive value indicates overestimation of total precipitation, and negative indicates underestimation. The value of A is in the range of [-2,2], with 0 value corresponding to the perfect forecast for a system-averaged precipitation intensity. Values close to -2 show almost missed events and values close to 2 almost false alarms. A value of A=1 indicates an overestimation by a factor of 3.

$$A = \frac{D(R_{mod}) - D(R_{obs})}{0.5(D(R_{mod}) - D(R_{obs}))}$$

The **S** component compares the total of volumes of the normalized precipitation objects (scaled over the maximum value for each object), and provides information about their size

and shape. The range of $S$ is [-2,2]. A positive value indicates that modelled precipitation objects are too large or too flat, and a negative value indicates that objects are too sharp and too small.

$$S = \frac{V(R_{mod}) - V(R_{obs})}{0.5(V(R_{mod}) - V(R_{obs}))}$$

The **L** component combines information about the distance of predicted and forecast mass centres and the error of a weighted average distance between the precipitation objects and centre of masses. It consists of two parts. $L = L_1 + L_2$. $L_1$ measures the normalized distance between the mass centres of the forecast x($R_{mod}$) and observation fields x($R_{obs}$), where d is the maximum distance found in the given domain between two boundary points. Its value range is [0,1]. The value of 0 indicates that the two fields mass centres are identical. However many different precipitation fields can have the same mass centres without being identical, therefore $L_1 = 0$ does not necessarily indicate a perfect forecast. $L_1$ **is independent of identification of features**.

$$L_1 = \frac{|x(R_{mod}) - x(R_{obs})|}{d}$$

$L_2$ is the difference of weighed mean normalized distance between the mass centre and the individual precipitation objects over observed r($R_{obs}$) and forecast r($R_{mod}$) fields and is a measure of difference of scattering of identified objects between the two fields. $L_2$ ranges between [0,1]. Therefore, $L$ ranges between 0 and 2.

$$L_2 = 2\frac{|r(R_{mod}) - r(R_{obs})|}{d}$$

A perfect forecast is therefore characterized by zero values for all three SAL components. A **taSAL** index $=(|S| + |A| + |L|)$ is suggested by Lawson et al. (2016) (as an objective skill score in order to quantify the forecast quality by means of one only parameter. The bigger the index is the worse is the forecast.

### 4.1.2 SAL Calculation

**Dataset**

- In order to apply the method, precipitation objects need to be identified within a **gridded** verification domain which should be the same for both observed and forecast fields. In the case of MesoVICT data used for INSPECT project, gridded data in ASCII are already available. However, if the method is used for comparison with grib observations and bufr model data, a software for reading the files in such format should be available. A file with latitude and longitude of the gridded fields is also necessary.
- **The domain size is important**. It should be identical for both observed and forecast fields. It is not recommended to be too large as components from different precipitation regimes can override each other.
- **The accumulation precipitation range should not be small**. Hourly precipitation fields are often noisy, with not well defined objects. It is recommended to calculate the SAL parameters with 6h precipitation range and up, unless a specific case with significant precipitation amounts is tested.

### Software required

In order to calculate the SAL components, after reading the input files, the following software are available:

- The original code (in Fortran) which can be provided by the authors.
- The libary SpatialVx (Gilleland, 2017) which is part of the free R language software with available documentation online. The advantage of this method is that by using featurefinder (function for defining objects) a the user can easily view objects and specify thresholds, smooth fields and discard small scale noise. The SAL parameters are then calculated by using the smaller function.

### Object Identification Methodology and examples of SAL calculation

The selection of a threshold value for objects identification is required for this method and it can be critical if the fields contain objects with different local maxima. A small change in the threshold can lead to a different number and size of objects and therefore different $L_2$ and S values. The possibilities of setting a threshold R are the following:

- Constant fixed user defined threshold (e.g. 2 mm) for forecast and observed fields. (This method is more subjective)
- $R = f \cdot R^{max}$: threshold is a fraction of $R^{max}$ (the maximum value of the field) and $f$ is subjectively chosen factor with a value of 1/15 suggested by Wernli et al. (2008) as the most appropriate to identify objects. Higher values of $f$ result in threshold increase
- $R = f \cdot R^{95}$: threshold is a fraction of the $95^{th}$ percentile of all gridpoint values in the domain which are larger than 0.1mm. (This is the method suggested by Wernli et al., 2009) and is used in the original code. This is sensitive to outliers. (eg. Single grid points with very intense precipitation)

Examples of SAL are given in Fig. 18, (calculated with method 3) where CMH S value is positive indicating that objects are flatter than observed (precipitation is more stratiform), and negative S for COSMO-2 means that modelled objects are sharper than observed. Positive (negative) values of A for CMH (COSMO-2) indicate overestimation (underestimation) of total domain precipitation. L values are slightly higher for COSMO-2 as there are fewer and more localized objects. TaSAL is lower for CMH, which indicates overall better forecast.

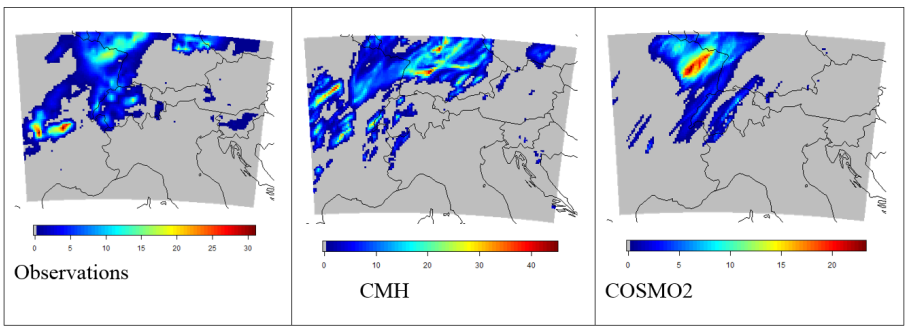



Figure 18: CMH and COSMO-2 forecasts versus Observations for MESOVICT case of 12 hourly Precipitation of 18h for 19/7/2007.

| Model | S | A | L | taSAL |
|---|---|---|---|---|
| **CMH** | 0.22 | 0.48 | 0.18 | 0.88 |
| **COSMO-2** | -0.12 | -0.55 | 0.25 | 0.92 |

**Object Identification Options**

- Smoothing: smoothing option is available on featurefinder function by setting do.smooth=TRUE and giving a value to smoothpar (which can also be different for observed and forecast fields). Specifically, by applying smoothing, the value at each grid is replaced by the mean value over a disk with a radius of a number of grid points given by the parameter smoothpar. The convolved data is **then** thresholded yielding a binary mask, so mostly affected are the object boundaries which are smoothed. Also small scale noise is filtered out. The selection of smoothing radius is not straightforward (Weniger et al., 2016) and it is not obvious how to select it for a given set of data. It is also relative to grid size. So, the same smoothpar can lead to light smoothing when applied to a grid of 1km but significant when applied to 7km. Smoothing for example can lead to bridges creation among different objects that are close to each other and unify them. If the selected smoothing factor affects one domain (forecast or observations) more than the other, then $S$ and $L_2$ parameters may completely change with only a small smoothpar change. In general, the increase of smoothing radius does not necessarily improve the results of the parameters, as it may unify or separate objects depending on the threshold, therefore alter the number of objects. Examples of effects of smoothing for the case of Fig. 18 are shown in Fig. 19. The effects of smoothing on $S$ in this example are apparent when moving from smoothpar=3 (Fig. 19b) to smoothpar=4 (Fig. 19d) as the unification of the objects in observed field results in a complete $S$ change (even on sign). Flatter objects were predicted in Fig. 19b due to the larger modelled yellow object, while the observed objects unification in Fig. 19c resulted in slightly sharper objects now predicted by the model. The bigger smoothing in Fig. 19d resulted in unification of almost all objects and made small objects disappear.

- Omission of small objects: the min.size option of featurefinder (min.size=c(n_min_obs, nmin_fcs)) defines objects as cohesive threshold exceedances and objects with less than n.min are omitted. The interpretation of this parameter is more straightforward that smoothing (Weniger et al., 2016) and it is easier to foresee the consequences of a particular choice of a parameter value on the objects as it does not unify or separate objects. It does not affect shape of large objects. If the selection of factor removes small objects and affects object spread for one domain (forecast or observations) more than the other, the $S$ and $L_2$ parameter may increase or decrease. Examples of effects of using min.size are shown in Fig. 20, where omitted modelled objects resulted in more widespread forecast precipitation with a slight and stable increase of $S$ and $L_2$ values with min.size increase. User can visualize objects after using the function featurefinder in order to decide about the best threshold and smoothing selection, but this selection is more straightforward for one specific case and not for a series of data with different precipitation regimes, that will be plotted afterwards on a SAL Plot.

  **Errors in observations** (too big values) can be discarded in the calculations when threshold is a fraction of a percentile of all gridpoint values in the domain. Also, by applying the min.size option in the observed field, small objects of any intensity that can be erroneous noise vanish. However, erroneous values within normal range can lead to wrong observation objects and therefore SAL components.

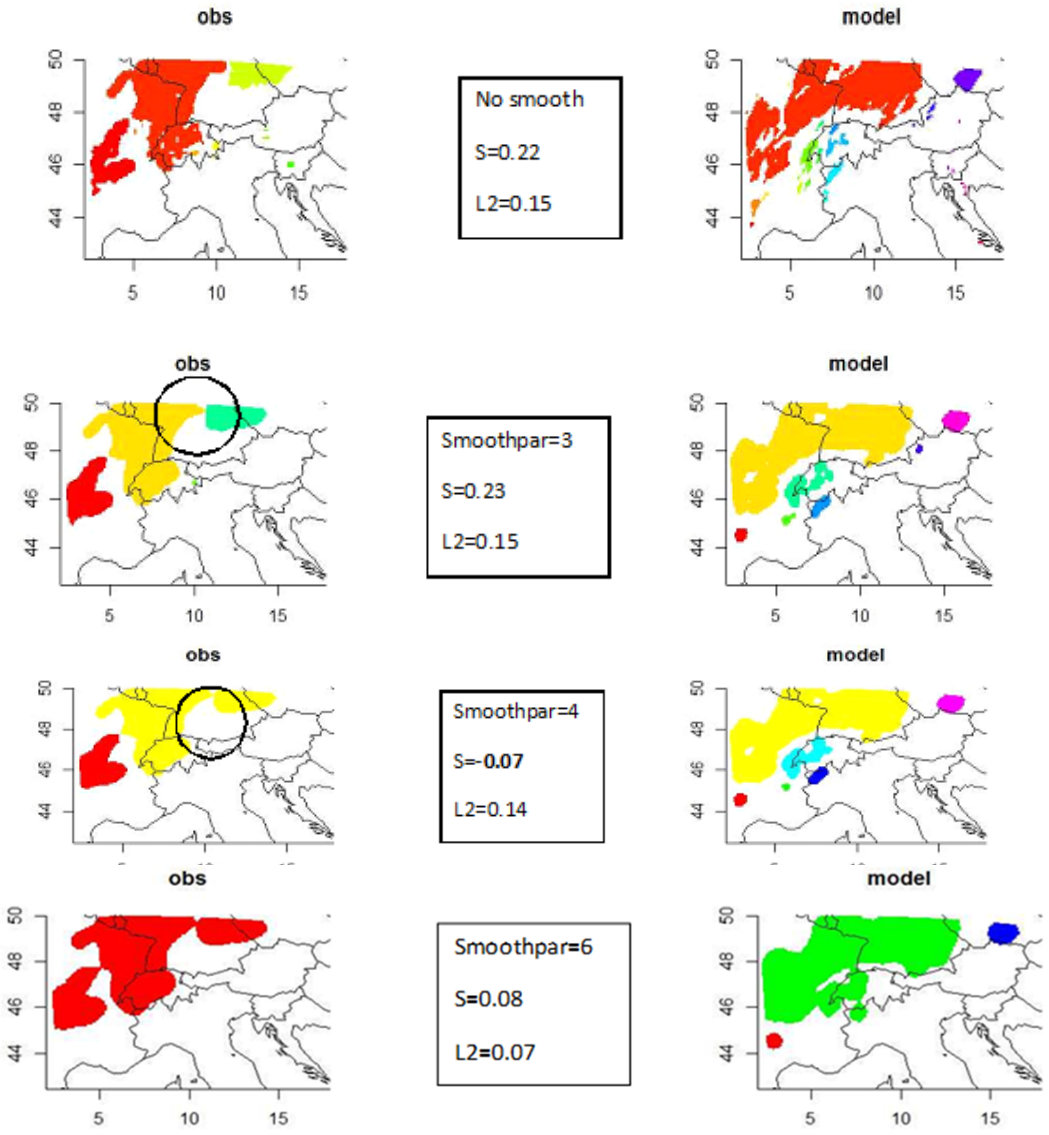



Figure 19: Examples of different smoothing factors and their impacts on objects and $S$ / $L_2$ applied on the case of Fig. 18.

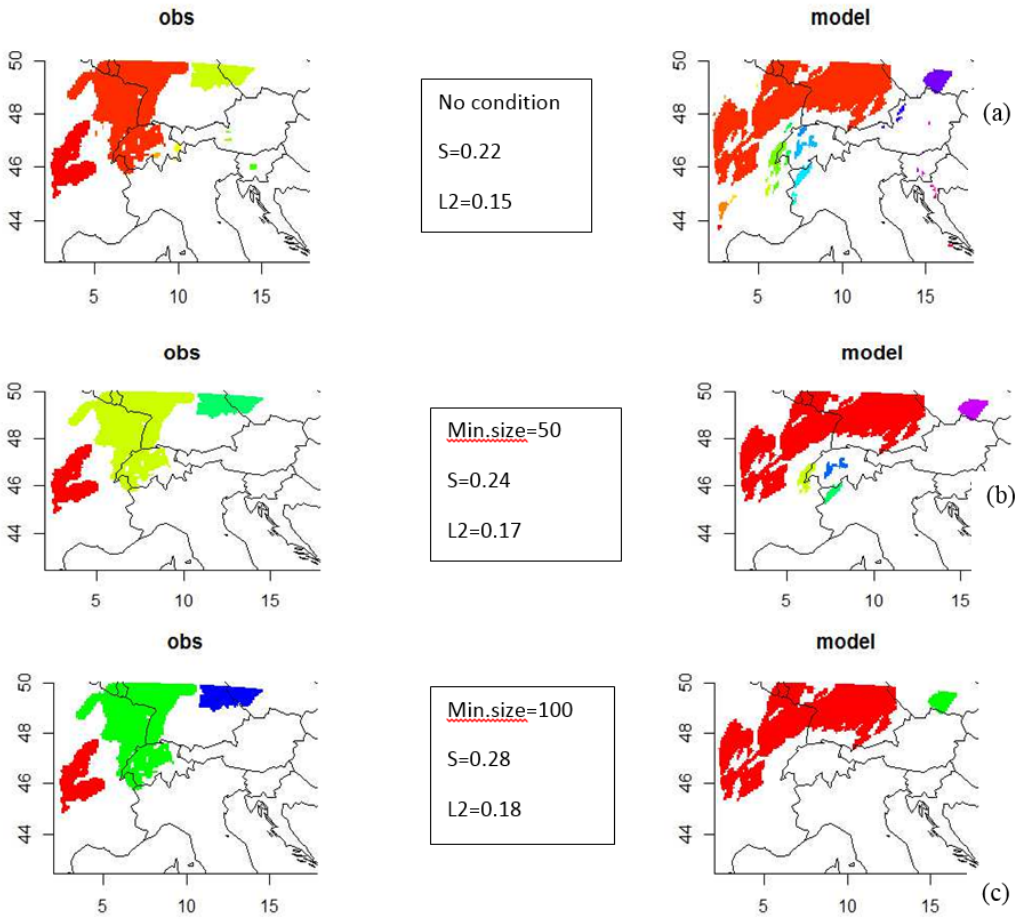



Figure 20: Examples of different min.size options and their impacts on objects and $S$ / $L_2$ applied on the case of Fig. 18.

### SAL Plots

The SAL method can be applied to a set of forecasts performed by the same model in order to define a general model tendency of predicting each of the three SAL components. This is done by the SAL diagram in which each point represents one case. The median value of $S$ and $A$ together with the $25^{th}$ to $75^{th}$ percentile box can also be plotted. The $L$ component value for each point is depicted with a colour scale. Example of this plot is shown in Fig. 21.

**Warning**: if no features are found in either or both forecast and observed fields, the SAL values cannot be defined. In this case (no objects found by featurefinder in forecast, observed, or both fields), saller function cannot be executed and the case is defined as Miss, False Alarm, or Correct Negative respectively. The use of extreme values of SAL parameters (-2.2) in case of lack of objects is not recommended and cases are simply omitted.

Statistically Specified thresholds (with selection of appropriate scaling factor) can be used when dealing with a significant number of cases in one SAL plot, as the precipitation amounts are different for each one, and a fixed threshold can be too big for some cases of them and exclude them. However, for one particular case, or for a number of cases with similar situation (ex. only convective) or if only extreme events exceeding a specified threshold are studied, then a fixed threshold can be used. SAL parameters can also be used to test the performance

of an EPS ensemble by plotting the SAL values of each of the members on a SAL plot and get an estimate of the ensemble forecast for one particular time (Barrett et al, 2015).

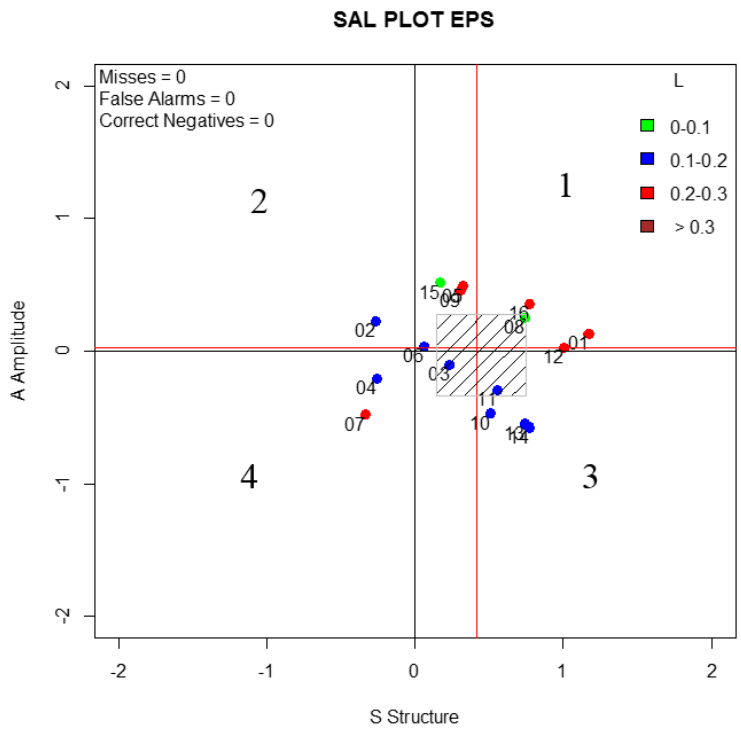



Figure 21: SAL diagram for COSMO-LEPS 3h precipitation for 21/6/2007 12UTC. The red lines denote the median values of $A$ and $L$, the gray box is the area of $25^{th}$ to $75^{th}$ percentile.

In Fig. 21, an example of a SAL diagram with 16 members of COSMO-LEPS indicates that median $S$ is positive, meaning that as an ensemble, predicted objects are flatter and larger than observed. The $A$ median parameter is close to zero and slightly positive which means that amplitudes of observed and modelled fields are comparable. Members in quadrant 1 produce too much rain with too large or flat objects. $L$ values are relatively low and less than 0.5. The forecast fields for each member may be significantly different. Therefore, fixed thresholds applied to a set of members may be too big for some of them, resulting in no object formation for some members, therefore $S$ and $L$ may be calculated with fewer of them. On the other side statistically specified thresholds for each member, imply a different threshold for each of them when compared to the observed field. It is up to the user and the specific case to select the appropriate method.

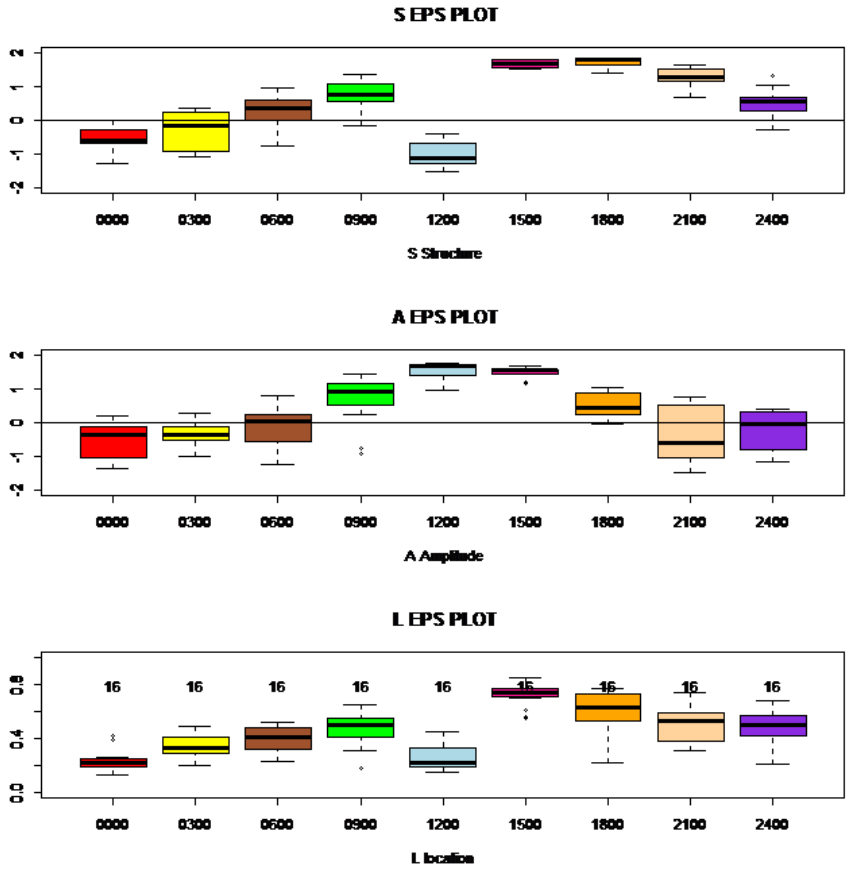



Figure 22: COSMO-LEPS hourly variation of ensemble SAL parameters (Statistically Specified threshold) for 20/6/07 with a Run beginning on 19/6/07.

EPS forecasts hourly variation can also be plotted as boxplots (each one representing the distribution of the members) of each parameter as a function of time lead (Fig. 22). The $L$ parameter distribution is now also better represented than SAL Plots. The R SAL plot and boxplots are written in R language. The calculation time can be significant if the program reads a large number of files and plots many cases for one SAL Plot In order to perform a SAL plot over a set of many different forecasts of an EPS model, there are two possibilities:

- **Medians** of the SAL parameters over the members for each case are calculated and then plotted on a SAL plot. The number of points will be equal to the number of cases;
- **Medians** of the SAL parameters over all cases for each member are calculated and then plotted on a SAL plot. The number of points will be equal to the number of members.

Examples for MESOVICT COSMO-LEPS 3 hourly forecasts of the Run beginning on 18/07 12 UTC (from 20 to 22/7) with the two methods given in Fig. 23. The two methods give comparable $A$ and $S$ median values with different spreads, as expected.

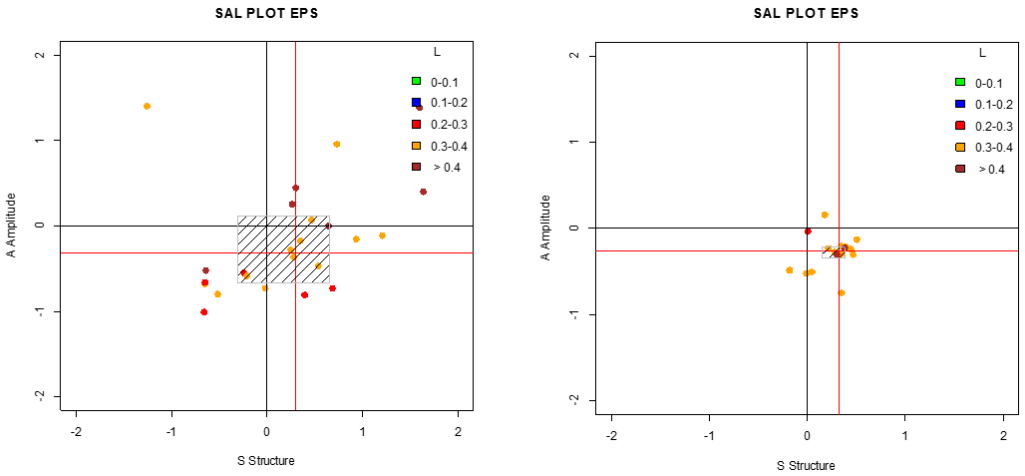



Figure 23: COSMO-LEPS SAL parameters for all 3-hourly forecasts from 20-22/7/07 with a run beginning on 18/6/07.

Further research on SAL applied to EPS models includes comparison of objects in fields of probability (Radanovics et al, 2015), and consideration of observation uncertainty.

### 4.1.3 EPS in terms of probability (a research topic)

Further research on SAL applied to EPS models includes comparison of objects in fields of probability (Radanovics et al, 2015) with consideration (if available) of observation uncertainty. In an effort to apply this concept to MESOVICT dataset the following ideas (which require further research) are presented in this document:

- *when the observation domain is constant and uncertainty is applied only for LEPS*

  First, the setting of precipitation threshold to be tested is needed (eg. 2mm). The fraction of the 16 LEPS members that predict precipitation above this specified threshold applied on every grid point consists the probability field in the model domain (range 0-1). In Fig. 24, different colours represent the different probabilities and the brown objects are of probability 1(all the 16 members predict precipitation above the threshold). When uncertainty is applied only to the model domain, these objects (in order for the comparison to be fair) can be compared to the observation objects which reflect the real conditions, and can also be represented by probability 1 (Fig. 24). In this case, A parameter will be only calculated for the 2 fields by taking into account only the objects with value 1 and setting to zero the remaining grid points for observed and modelled domains (because in the observations field the remaining grid points cannot be represented as probabilities) . Also, $S$ parameter will reflect only their size and not sharpness. These objects are detected by the featurefinder function (Fig. 25) and the resulting SAL parameters will be a measure of how the objects over a specified threshold can be forecasted by all 16 members of EPS model. In this particular case $S$=1, $A$=0.38, $L$=0.3;

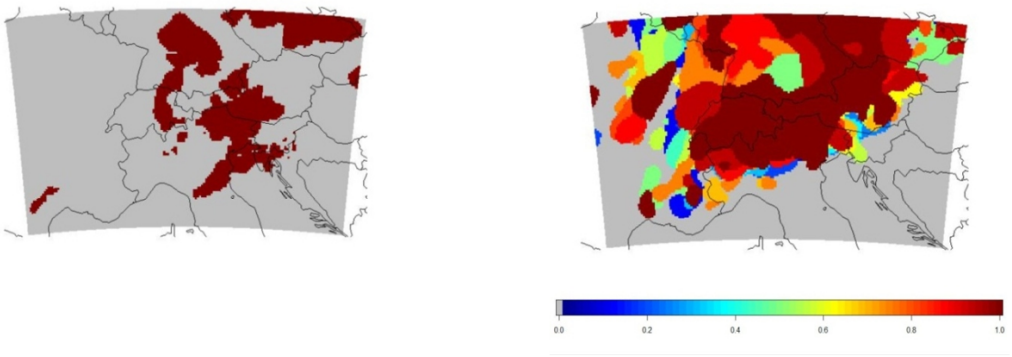


Figure 24: Probabilities of 3h precipitation threshold > 2mm for observations (left) and LEPS (right). 21/06 12UTC. The colour scale is the probability range.

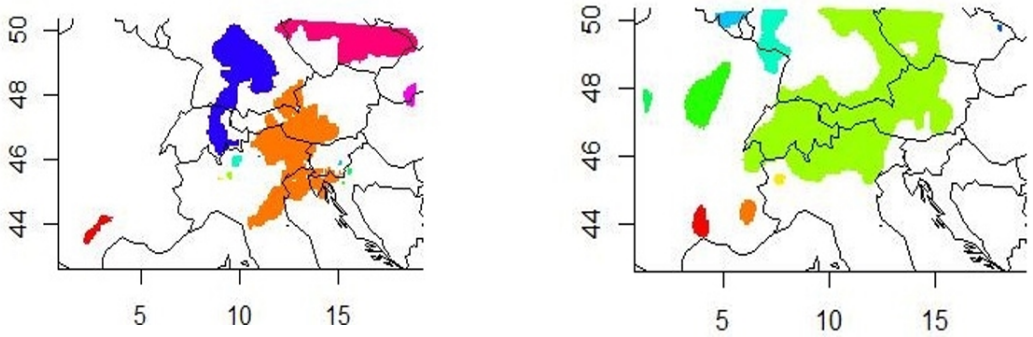



Figure 25: Objects of probability=1 as detected by the featurefinder function for observations (left) and model (right). Colours indicate different objects.

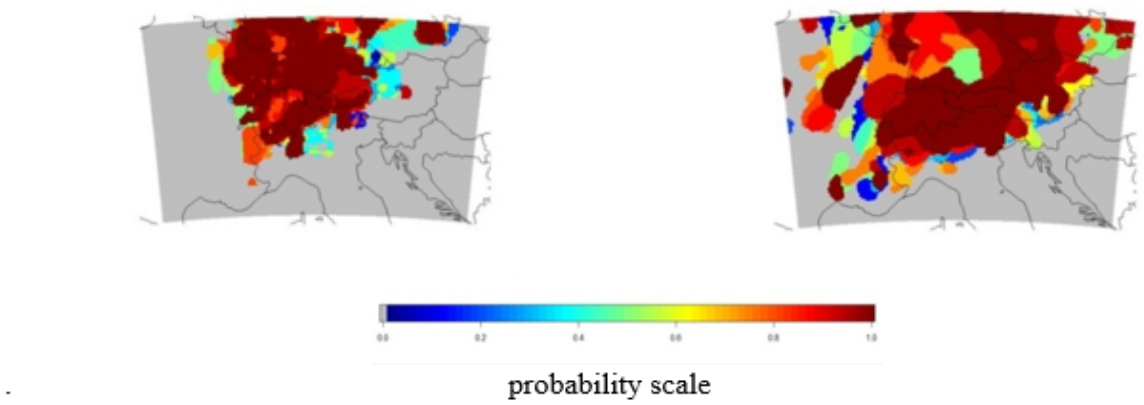
- *when uncertainty is applied also to observations (in this case 16 members of VERA ensemble)*

  When uncertainty is also applied to observations (Gorgas and Dorninger, 2012a), the concept of the fraction of the members exceeding the specified threshold can be applied in the same way to observed and LEPS domains, and fields of probability can now be created in both of them. In this case, probability thresholds of less than 1 can also be tested. For example, for precipitation threshold 2mm, the objects of each field created by the different probabilities exceeding it, are shown in Fig. 26a. In Fig. 26b, the objects of probability=1 (all members in observation and VERA fields predict precipitation above 2mm) detected by featurefinder in both domains are shown. In this case, $S$=0.37, $A$=0.6, $L$=0.1. The objects of probability threshold $< 1$ (0.5 in the case of Fig. 26b, which denote that at least half of members predict precipitation above the threshold) are bigger and include also higher probabilities can be compared with the SAL method for the two domains. In this case, $S$=0.65, $A$=0.5, $L$=0.05. Further research on LEPS SAL parameters includes the introduction of a SAL index (Radanovics, 2017a) for LEPS (Ensemble SAL) with average means over the members of the equation parameters for observed and model fields. This formulas that may be introduced in SpatialVx (Radanovics, 2017b), will be a more efficient way of SAL parameters estimation for Ensemble Forecasts.

$$L_1 = \frac{|\langle x(rr_{mod})\rangle - \langle x(rr_{obs})\rangle|}{d}$$

$$L_2 = 2 \times CRPS(P(\frac{r_{mod}}{d}), P(\frac{r_{obs}}{d}))$$

$$S = \frac{\langle V(mod)\rangle - \langle V(obs)\rangle}{0.5\langle V(mod)\rangle + \langle V(obs)\rangle}$$




a)


b)


c)

Figure 26: a) Probabilities of 3h precipitation threshold $>$ 2mm for VERA ensemble (left) and LEPS (21/06 12h). b) Objects of probability threshold=1 as detected by the featurefinder function for VERA ensemble (left) and LEPS (right). Colours denote different objects. c) Same as b with probability threshold=0.5.

## 4.2 MODE, CRA, SAL: IMGW-PIB experience

The experience described in this section is unique in INSPECT, as IMGW-PIB applied three different object-based methods, which facilitated their intercomparison.

### 4.2.1 Method applied (related to an INSPECT Task) and objectives

MODE - Method for Object-Based Diagnostic Evaluation, its objective is to identify localized features of interest in scalar fields, merge and/or match features and compare features in two fields to identify which features best correspond to each other. CRA Contiguous Rain Area, identifies features of interest, uses pattern matching techniques to determine the location error, errors in area, mean and maximum intensity, and spatial pattern. The total MSE (Mean Square Error) is decomposed into components due to location, volume, and pattern error (MSE displacement, MSE volume, MSE pattern) (see also Chapter 4.3) SAL - Structure, Amplitude, and Location, requires a preselection of a domain of interest, a definition of a threshold to identify objects in the observational data and model forecast, one-to-one matching between the identified objects in the observed and forecasted fields is not required. A forecast is perfect if S = A = L = 0. (see also Sec. 4.1.1.)

### 4.2.2 Short description of the dataset (forecast-observation data), adaptation required, software for the method application

- Radar data
  - OPERA (Operational Programme for the Exchange of Weather Radar Information), 1 hour rainfall accumulation. The composites cover the entire Europe in a Lambert Equal Area projection. For the project data in HDF5 format was used. The HDF5 files are read directly in the R software by using "HDF5 interface to R". The data are adopted to geographical coordinates (latlon geographical projection) in terms of verification against COSMO PL data
  - Polish national RADAR (POLRAD network) data composite. The data are available in Rainbow binary XML format and can be processed with additional php5 tool available at IMGW-PIB to extract rainfall accumulation and geographical locations.

  Both data sets require a procedure of matching to COSMO PL model domain.
- Vienna Enhanced Resolution Analysis - VERA data
- Forecast data
  - COSMO-PL 7 km, COSMO-PL 2.8 km
  - COSMO-2 data - interpolated on the VERA grid.

### 4.2.3 Main findings (plots and explanation)

The all plots and statistics below were obtained using SpatialVx R-package developed at NCAR (Gilleland, 2015).

- MODE example. MODE graphical output shows matched objects (e.g. Hits) and unmatched objects (e.g. False alarm in forecast field, Misses in observation field), which are based on the interest value. The interest value is an overall measure of similarity between objects in the observed and forecast fields, ranges from 0 to 1. The value of 0.70 is used for objects to be considered matched. COSMO-PL7 precipitation model output was verified against OPERA radar data. The result of MODE method shows a good correspondence of the precipitation objects between COSMO-PL7 and radar data.

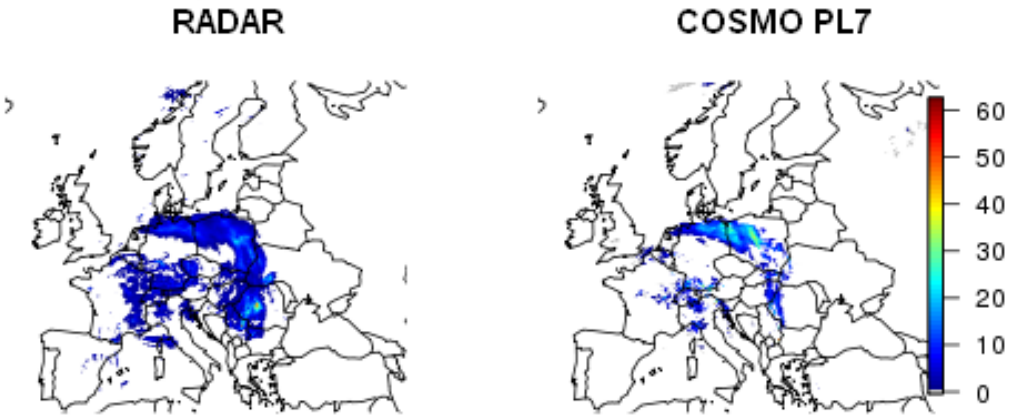



Figure 27: 24h QPF ending at 00 UTC on 05.05.2017.


Figure 28: Identified, matched objects, threshold 5mm, Matched objects are indicated in the same colour in both the observed and the forecast fields .Grey colour corresponds to unmatched objects.

- CRA example. CRA verification was used to verify forecast model COSMO-2 against VERA data. One of the cases was chosen for the verification. For object pairs 2 and 4 the total error is associated with the error in location, while for object pairs 1 and 3 the total error is associated with the fine scale structure.

| Observation feature | Forecast feature | Total interest |
|---|---|---|
| 1 | 1 | 0.85 |
| 2 | 2 | 0.7 |
| 2 | 1 | 0.46 |
| 1 | 2 | 0.42 |

Table 2: Ranking of feature pairings based on total interest.

| | MSE.total | MSE.disp | MSE.volume | MSE.pattern |
|---|---|---|---|---|
| 1 | 0.01222 | 0.00232 | 7.405e-05 | 0.00983 |
| 2 | 0.00173 | 0.00123 | 4.615e-07 | 0.0005 |
| 3 | 0.07752 | 0.01059 | 3.919e-03 | 0.06301 |
| 4 | 0.00166 | 0.00124 | 3.645e-07 | 0.00041 |

Table 3: Total MSE and its components.

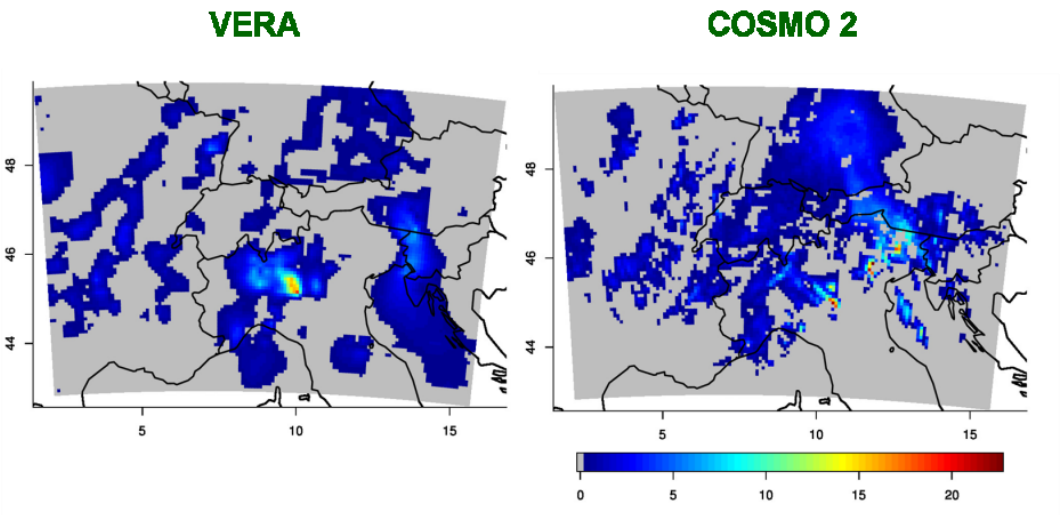



Figure 29: 1h accumulated precipitation 26.09.2007, 18UTC.

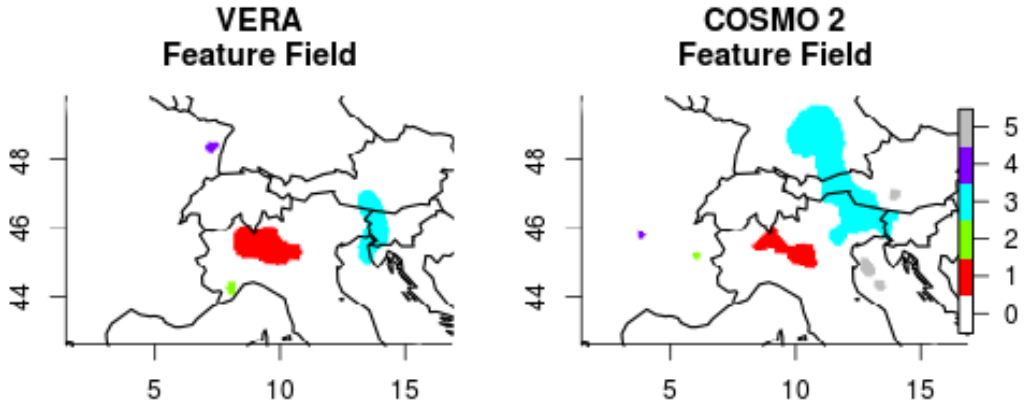



Figure 30: Objects, threshold 2 mm, 26.09.2007, 18 UTC.

main.tex

- SAL example. SAL verification was used to verify 1 hour accumulated precipitation output of forecast model COSMO 2.8 against radar data of POLRAD network. The verified period was from 05 18UTC to 07 06UTC August 2016. Almost all SAL entries are found in the top right quadrant of the diagram. The bottom right quadrant contains only a few of them. For all entries component S is positive which means that the predicted forecast objects are too large or widespread with respect to the observed objects. For most entries component A is positive which indicates an overestimation of the domain-averaged precipitation coming from the model. There are only a few entries when the model underestimates the precipitation (negative A component). No entries were found in top and bottom left quadrants.

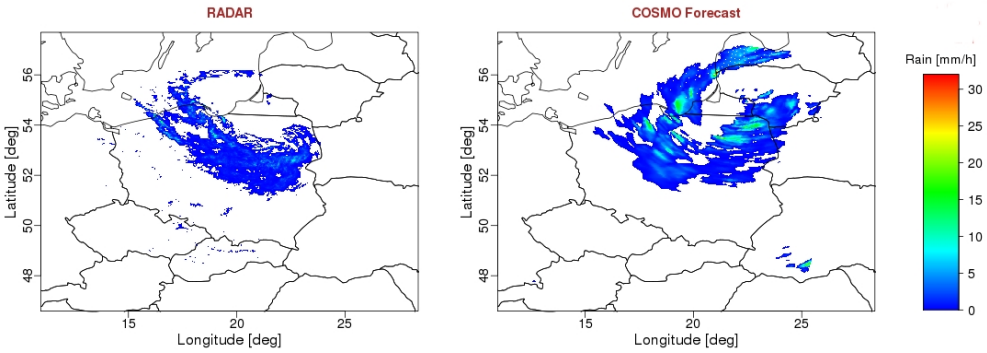



Figure 31: 1h accumulated precipitation, Radar data composite (POLRAD) against COSMO-PL 2.8, 06.07.2016, 20 UTC.

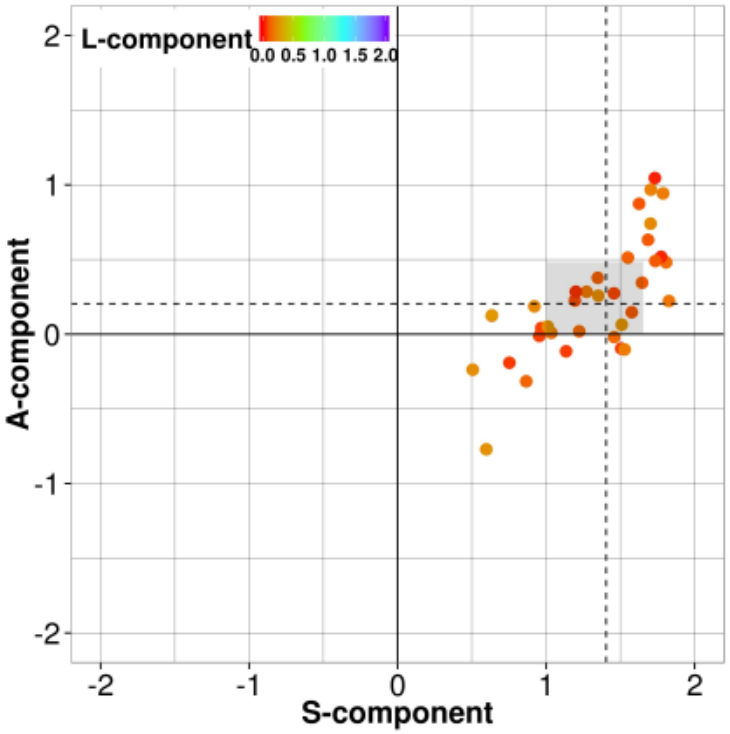



Figure 32: SAL verification, 1h accumulated precipitation, 05 18UTC  07 06UTC August 2016. Every dot shows the values of the three components of SAL for a particular forecast. L component is indicated by the colour of the dots. The grey area extends from the 25th to the 75th percentile of the distribution of S and A, respectively. The dashed lines depict the median values of S and A components.

### 4.2.4 Characteristics of the method applied

Identification in MODE is done separately in observation and forecast fields by using two parameters such as convolution radius and threshold. The method looks at characteristics of the objects such as intensity distribution, centroid location, area, curvature, orientation, and attempts to match objects in the forecast and observed field based on these characteristics. Objects nearby can be merged. CRA method gives useful information of the forecast errors. The total mean square error is decomposed due to location, pattern and intensity. The method depends on pattern matching. To find the optimal rigid transformation between two features can be difficult. SAL components correspond to aspect of amplitude, location and structure of precipitation field in a preselected area. All three methods MODE, CRA, SAL require identification features first. Applied method calculation time using SpatialVx library can depend on chosen merging/matching function, additionally all these methods are able to deal with different density of observation. Depending on the model resolution object features may have different pattern complexities and fine structures which further affects merging algorithm performance and computed scores.

## 4.3 CRA experiments in Roshydromet for MesoVICT (A.Bundel and A. Muraviev))

In the framework of MesoVICT (Mesoscale Verification Intercomparison over Complex Terrain, phase 2 of the ICP, https://ral.ucar.edu/projects/icp/), a set of cases is provided to compare various spatial verification methods. The object-based CRA (Contiguous Rain Area) method is used (Ebert, McBride, 2000). First, the objects are identified in observed and forecast fields. Then, the objects pairs are identified using some matching criterion based on the distance between the objects. Then, an optimal shift of the forecast object to the observed object is found by minimizing the error. Here also, different criteria can be used, such as the correlation coefficient and the mean squared error (MSE). In this study, the MSE criterion was used. Then, the difference is found between the initial MSE (represented as $MSE.total = MSE.displacement + MSE.volume + MSE.pattern$) and the MSE after the shift. This difference is the error due to forecast displacement: $MSE.displacement = MSE.total - MSE.shifted$. The MSE, which is left, consists of the squared difference between the mean precipitation volume in the forecast and observed object $MSE.volume = (\mathbf{F} - \mathbf{X})^2$ and of fine scale pattern discrepancies $MSE.pattern = MSE.shift - MSE.volume$. To calculate the CRA scores, free R SpatialVx package (https://cran.r-project.org/web/packages/SpatialVx/index.html) was used. It is developed by E. Gilleland contains most part of existing spatial methods, including identifying, matching, and merging features in observed and forecasted fields.

### 4.3.1 Deterministic study

**Setup of experiments**

- Mesovict core case: 20-22 June 2007;
- 1-h precipitation accumulations;
- Vienna Enhanced Resolution Analysis (VERA) observation analysis (Dorninger, M.et al., 2013) is used as reference data in this study;
- the Swiss COSMO-2 deterministic model is used as model data (2.2 km grid step).

Fig. 33 shows the time series of domain precipitation maximum during the MesoVICT core case. It can be seen that the absolute VERA maximum on 20 June 2007, 21h UTC, (49.84 mm/h) falls within the COSMO-2 domain. Most of other observed maximums were within the COSMO-2 domain. Overall, COSMO-2 reproduces well the time series of maximums, but gives the highest maximum 12 hours later, on 2007062020 much later, on 2007062108.

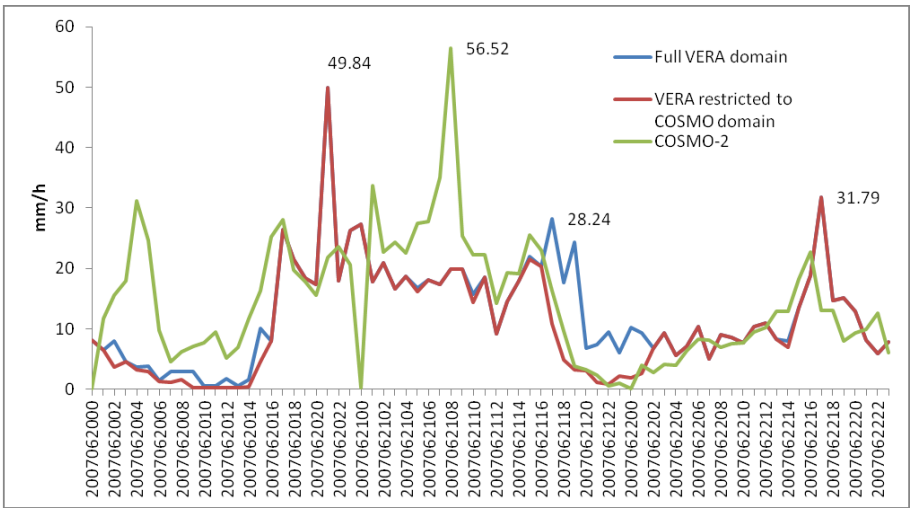



Figure 33: Time series of domain 1h precipitation maximums during the MesoVICT core case of 20-22 June 2007.

### Identification of objects in R SpatialVx

Function FeatureFinder is used to identify objects in both observed and model fields. First, the field is smoothed using a convolution smoother, and then it is set to a binary image where everything above a given threshold is set to one (Davis et al, 2006). Features are identified by groups of contiguous events (or connected components in the computer vision/image analysis literature). Option min.size enables eliminating features with a size less than determined value from the analysis. Through a set of experiments, the reasonable value was determined as min.size = 20 grid points or ∼36*36 km.

### The effects of smoothing

By applying smoothing, the value at each grid is replaced by the mean value over a disk with a radius of a number of grid points given by the parameter smoothpar (see also Chapter 4.1 about the effects of smoothing). The FeatureFinder option uses disc kernel smoother from the R package smoothie.

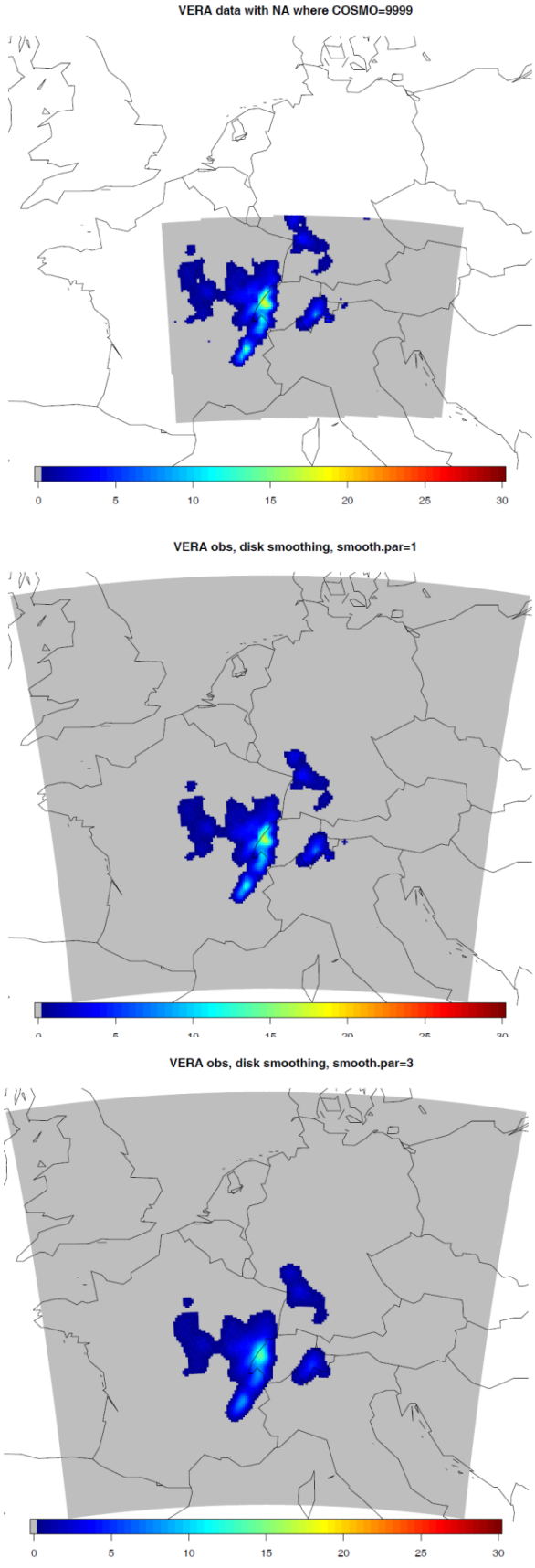



Figure 34: VERA gridded observation analysis on 2007062106: no smoothing (top), smoothing parameter =1 (middle), smoothing parameter = 3 (bottom).

Smoothing spreads the maximum over wider areas (Figure 4.3.1.2). Therefore, no or minimal degree of smoothing is needed in estimating intense precipitation areas. Smoothing parameter = 1 is chosen for the experiments described in this chapter. Functions for matching objects in R SpatialVx used in this study are as follows:

- Minboundmatch, the minimum boundary separation is calculated by first finding the distance map for every feature in the observed field, masking it by each feature in the forecast field, and then finding the minimum of the resulting masked distance map (Eric Gilleland, SpatialVx Manual). The distance map is found using distmap. The function distmap of point pattern computes the distance from each pixel to the nearest point in the given point pattern. If type is single, then the features are matched by the smallest minimum boundary separation per feature in each field. If type is multiple, then every feature is matched so long as their minimum boundary separation (measured in grid squares) is less than or equal to a specified value
- Centmatch is based on the method proposed by Davis et al. (2006a). It is possible for more than one object to be matched to the same object in another field. Objects are matched, if the centroid distance D is less than
  - the sum of the sizes of the two objects in question (size is the square root of the area of the object) (Centmatch 1)
  - the average size of the two objects in question (Centmatch 2)

  Centmatch does not merge objects explicitly, but determines possible merges applied if MergeForce function is run after centmatch (used in this study). The merging algorithm is described in (Eric Gilleland, SpatialVx Manual, deltamm and centmatch functions). It merges objects according to the minimum of distance metric between the objects.

**Examples of CRA application**

In Fig. 35 1h precipitation maps are given for 2007062021, VERA max precipitation during MesoVICT core case.

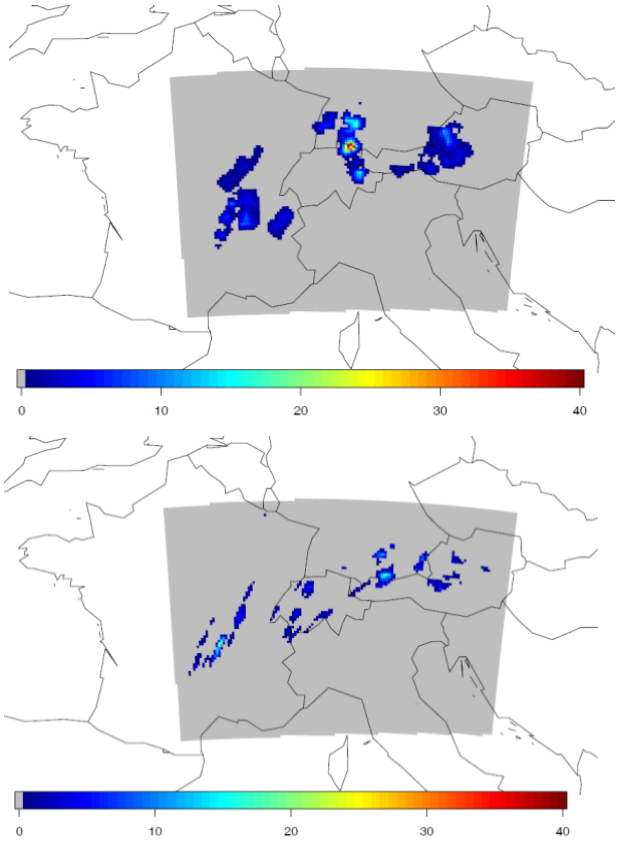



Figure 35: 1h precipitation maps from 2007062021 in mm/h: VERA observations (top), COSMO-2 (bottom).

Fig. 36 gives the matched features for a precipitation threshold 0.5 mm/h as an output of different matching functions for 2007062021.

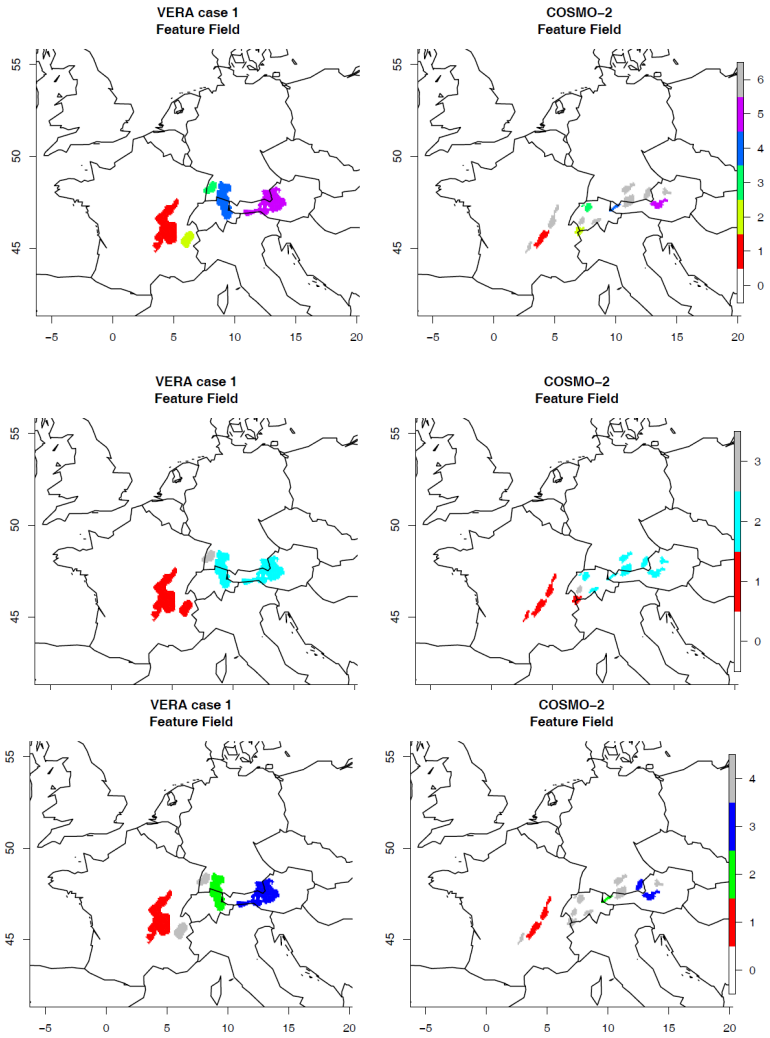



Figure 36: Matched objects pairs using (for precipitation $> 0.5$ mm/h) using different matching functions for 2007062021: Minboundmatch (top), Centmatch 1 (middle), Centmatch 2 (bottom). Same colours indicate paired objects.

It can be seen from Fig. 36 that it is difficult to objectively choose a best matching function. All methods are acceptable. The choice depends on the user preferences. Centmatch 1 makes implicit merging. In Fig. 37 the objects matched using minboundmatch are given for the same date, 2007062021, but for precipitation threshold $> 5$ mm/h. The functions Centmatch 1 and 2 do not make any matching in case of intense precipitation. Thus, these matching criteria are too strict for small features.

| | displacement | volume | pattern |
|---|---|---|---|
| first pair | 0.0006 | 0 | 0.0006 |
| second pair | 0.0007 | 0 | 0.0018 |

Table 4: CRA error components

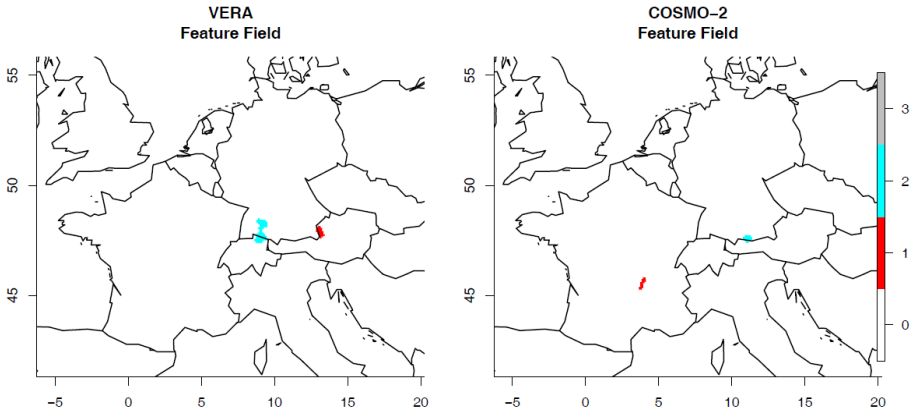



Figure 37: Matched objects pairs using minboundmatch, 2007062021, precipitation threshold > 5 mm/h.

The CRA scores (Table 4) show that the fine scale pattern error contributes most to the total mean squared error in this case. Let us consider another time slot from the MesoVICT case 1, 2007062115, when COSMO-2 made a good forecast (Fig. 38) with higher maximum intensities at the northeast of the domain. Fig. 39 for precipitation threshold 0.5 mm/h demonstrates that Centmatch 1 merges several features in the observed field to match them with one big precipitation object in the COSMO-2 field, while Centmatch 2 and Minboundmatch, which imply stricter distance criteria, do not merge separate objects. The CRA scores show that the fine scale pattern error is most important also in this case.

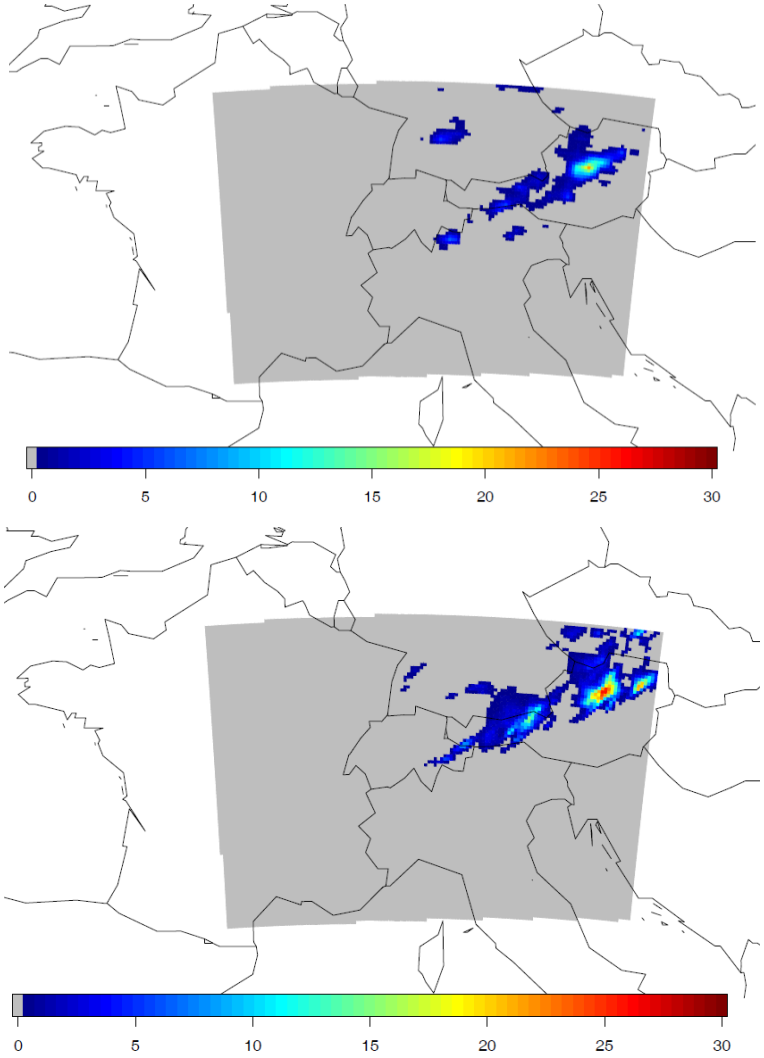



Figure 38: Matched objects pairs using (for precipitation $> 0.5$ mm/h) using different matching functions for 2007062021: Minboundmatch (top), Centmatch 1 (middle), Centmatch 2 (bottom). Same colours indicate paired objects.

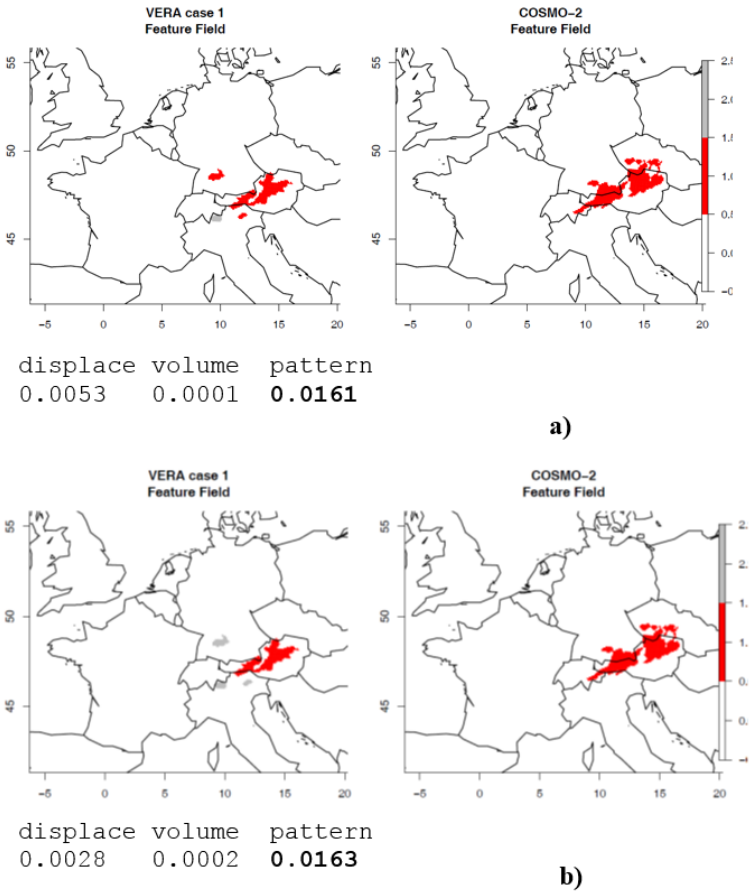



Figure 39: Matched objects pairs and corresponding CRA error components using (for precipitation $> 0.5$ mm/h) using different matching functions for 2007062115: a) Centmatch 1 b) Centmatch 2 and Minboundmatch (give same matches in this case), same colours indicate paired objects.

For higher precipitation threshold (Fig. 40), all matching functions give the same result. The best match is detected (the central object in COSMO-2 field with precipitation maximum). It should be noted that there were another two intense precipitation areas forecasted, which represent false alarms. In this method, the false alarms are discarded.

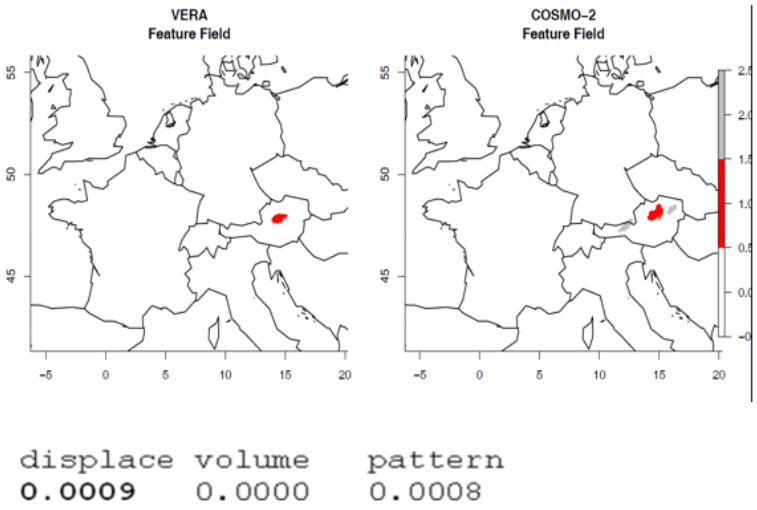



Figure 40: Matched objects pairs and corresponding CRA error components (for precipitation > 5 mm/h) for 2007062115. All matching functions (Centmatch 1, Centmatch 2, and Minboundmatch) give the same matches in this case), same colors indicate paired objects.

Other time slots and synoptic situations were considered, which are not given here.

**Conclusions**

Smoothing can be undesirable for intense precipitation estimation, although it depends on the user requirements. Option for splitting objects is desirable sometimes. For lower precipitation thresholds and thus wider features, centmatch 2 gives more reasonable results overall. Centmatch 1 makes more merging. For higher thresholds, centmatch often leaves all the objects unmatched due to small areas of features. Minboundmatch seems more promising, but with a minimum boundary separation distance beyond which features should not be matched. According to CRA, most of the error usually comes from the fine structure of the fields (MSE.pattern) for lower precipitation thresholds. For higher thresholds, displacement error contribution increases. The optimal choice of matching procedure can be rather difficult. Each case should be considered before application of particular matching function. Aggregation of results is difficult. The SAL method seems easier to apply because it does not require pair-wise matching.

**Ideas for future work**

It seems useful to formalize the complexity of situations when the application of metrics is useless. Below is the list of possible approaches to that. It should be noted that it is only an idea at the present state:

- by the maximum number of different features in the field;
- by the complexity of their boundaries (fractals?);
- simple spatial dispersion of the field;
- by the number of holes in a feature (Betti number?);
- other criteria?

### 4.3.2 An object-based approach to assess the MesoVICT ensemble data

This chapter contains an exploratory study with a goal to transfer the object-based method to ensemble data. The results are preliminary. The starting point was the method proposed in (Johnson and Wang, 2012). The forecast probability is generated for each forecast object as the fraction of ensemble members forecasting the object of interest. In this case, the objects of interest are the observed objects unlike (Johnson and Wang, 2012) where one of the ensemble members was taken as reference to calculate object probabilities. In our approach, performance of the ensemble system could be assessed a posteriori, but the probabilities of objects defined in this way could not be forecasted in the real time, as the observations are not yet known at the time of the forecast. If we want to assess the real forecasts, we need to choose the objects of interest from the ensemble. Then, the member defining the forecast objects should be randomly chosen so that the results were not improved by attempts to select a "best" member. Then, the probability of each observed object can be estimated using the traditional probabilistic scores, the BSS, for example, but verification was not yet implemented. The setup of the experiments is as in Sec. 4.3.1, but COSMO-E ensemble of Swiss meteoservice is taken as model data. In Fig. 41, the upper map represents the VERA observed objects, and the pairs of maps below are the matched observed and model objects for the first six ensemble members from a 21st member ensemble. The colours indicate matched pairs. Then we can calculate the probability of predicting each of observed objects. For example, the VERA object painted in yellow colour in the upper map was matched to an object in 20 of 21 ensemble fields, thus, its probability is 20/21.

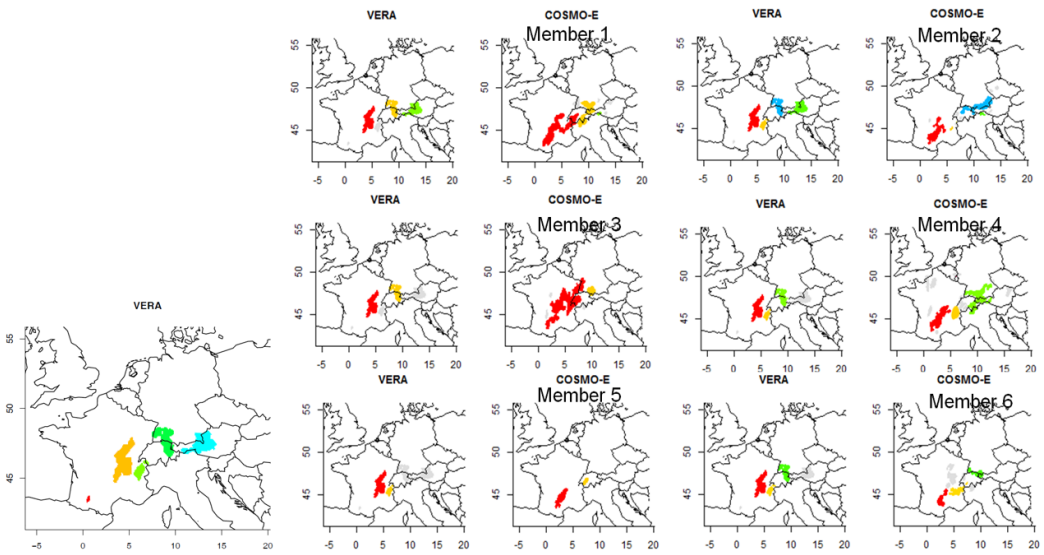



Figure 41: Minboundmatch: 2007062021, COSMO-E ensemble, first 6 of 21 members, precipitation threshold > 0.5 mm/1h. Probabilities of each of 5 observed objects: 1/21 20/21 10/21 19/21 14/21.

A deficiency of such an approach is that the CRA method itself cannot be applied, only the usual probabilistic measures, such as the BSS, can be calculated when a large number of forecasts is aggregated. Another drawback is that no merging of objects is possible as the list of observed objects must be the same for matching with all ensemble members, thus centmatch matching function, which produces merging, cannot be applied. To apply CRA, the possible approaches could be:

- to calculate location, volume, fine pattern errors for each ensemble member, and to average them;
- to identify objects using the probability threshold (Gallus, 2010).

An example of the second approach to object-based method is given in Sec. 4.1 (by D. Boucouvala).

### 4.3.3 Processing nowcasting forecasts using CRA at RHM (A. Muraviev)

The studies on test cases, such as MesoVICT cases, were conducted having in mind the main goal to verify operational forecasts, in particular, short-range and very-short-range forecasts. Below are the first results of verification of a nowcasting system implemented at the Hydrometcentre of Russia of Roshydromet (Muravev et al., 2018). The core of the system is the statistical STEPS scheme (Short Term Ensemble Prediction System) (Bowler N. et al., 2006) constructed as a multiplicative cascade model with an optical flow technology (the radar fields sequences are considered as an optical flow). Its motto is "seamless, stochastic, scaling, spectral, self-learning". STEPS can run in deterministic and ensemble modes. The results below are obtained with deterministic version. This chapter focuses on verification technology rather than on the scientific findings described in (Muravev et al., 2018).

**Verification setup**

Period: May 1, 2017 September 30, 2017(∼22000 forecasts). Nine radars over the Central Russia with a total area of about 1100x1300 km were used. For each radar, the area is 500*500 km was considered. The STEPS precipitation intensity forecast calculates fields up to 2.5 h lead time with 10-min time step (15 consequent fields for each forecast) with about 2km grid step. For verification, only 169 situations with intense precipitation were chosen by visual analysis. The SpatialVx package was used to identify objects and to calculate CRA scores. The objects with areas less than 1225=35*35 grid points (about 7070 km) and lager than 16384=128128 grid points (about 250250 km) were excluded from analysis using min.size and max.size option in FeatureFinder function. The radius of averaging for convolution smoothing was chosen empirically as 9 grid points (18 km).

**Analysis**

Let us consider a case from 17 May 2017. It can be seen from Fig. 42 that the nowcasting system forecasted well the rain object at the first lead time, but by the end of the forecast, the single object in the radar field splitted in two. In Fig. 43, it is reflected in the abrupt change in x_shift (shift in object mass centre in longitude) from positive to negative values at the last two lead times (14 and 15), because the system of object recognition and matching switched to a new object when the single object in the radar field (Fig. 42, lower right field) splitted in two. It follows from the CRA scores that the nowcast object is shifted eastward relative to the corresponding radar object. This case demonstrates the difficulties in matching objects during intense convection cases, as the objects appear and disappear chaotically and it is difficult to follow their history in both reference and forecast fields.

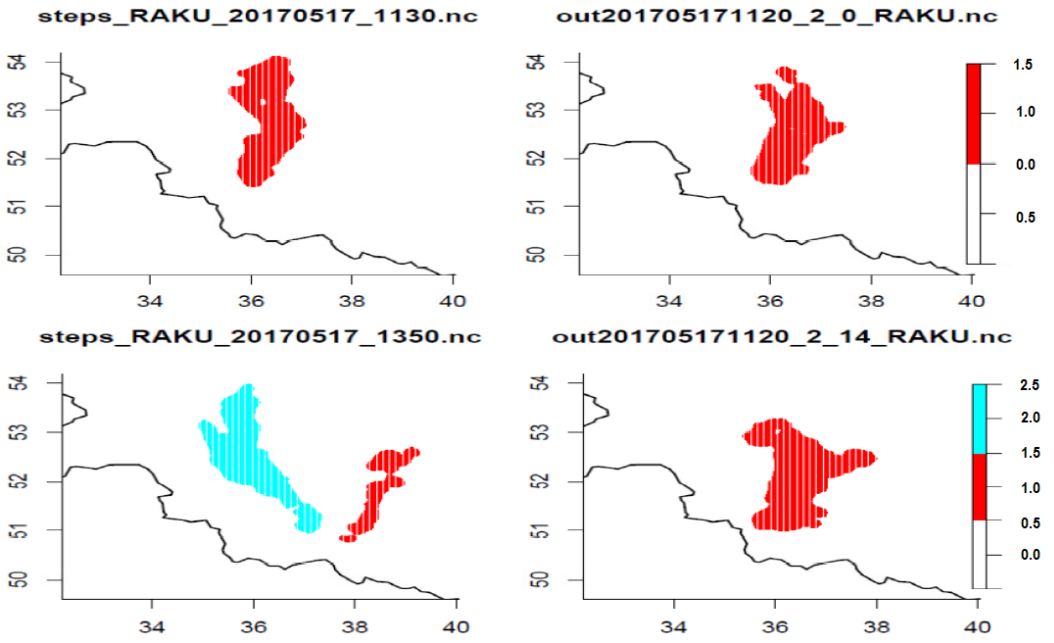



Figure 42: Objects matched using SpatialVx matching function Centmatch 1 (see Sec. 3.2.1) on 17 May 2017. Upper panel: the first nowcast time (11:30 UTC), lower panel: the penultimate nowcast time (13:50 UTC); on the left, the radar fields, on the right, the nowcast fields. Colours indicate matched pairs. Kursk (RAKU) radar.

R-SpatialVx craer output

| ilead | x_shift | y_shift | MSE.total | .shift | .displ | .volume | .pattern |
|---|---|---|---|---|---|---|---|
| 1 | 1.89 | -10.49 | 0.4387 | 0.6986 | -0.2599 | 0.0000 | 0.6986 |
| 2 | 2.83 | -11.04 | 0.6058 | 1.0097 | -0.4040 | 0.0002 | 1.0096 |
| 3 | 0.34 | -10.44 | 0.7205 | 0.8957 | -0.1753 | 0.0024 | 0.8933 |
| 4 | -1.41 | -13.42 | 1.2529 | 1.3553 | -0.1024 | 0.0074 | 1.3479 |
| 5 | -5.91 | -12.93 | 1.7393 | 1.7697 | -0.0304 | 0.0140 | 1.7557 |
| 6 | 4.20 | -18.01 | 0.9870 | 1.2413 | -0.2543 | 0.0044 | 1.2369 |
| 7 | 3.50 | -13.52 | 0.9584 | 1.0873 | -0.1288 | 0.0014 | 1.0859 |
| 8 | 2.36 | -6.45 | 1.0874 | 1.0787 | 0.0087 | 0.0001 | 1.0787 |
| 9 | 3.24 | -7.06 | 1.1609 | 1.1567 | 0.0042 | 0.0002 | 1.1565 |
| 10 | 5.49 | -6.60 | 1.2815 | 1.2475 | 0.0340 | 0.0013 | 1.2463 |
| 11 | 6.65 | -7.82 | 1.2953 | 1.2648 | 0.0305 | 0.0026 | 1.2622 |
| 12 | 10.41 | -14.45 | 1.3093 | 1.1718 | 0.1375 | 0.0036 | 1.1682 |
| 13 | 11.33 | -16.40 | 1.3937 | 1.2378 | 0.1559 | 0.0085 | 1.2293 |
| 14 | -13.02 | -12.33 | 2.3734 | 2.4361 | -0.0627 | 0.0008 | 2.4353 |
| 15 | -16.07 | -14.96 | 2.6149 | 2.7596 | -0.1446 | 0.0010 | 2.7586 |

Figure 43: CRA scores for the case of Fig. 42.

### Aggregation of the scores and conclusions

To assess the forecast (or nowcast) system, we need to aggregate a large number of cases. The following method was proposed in this study: the distribution parameters of CRA scores were calculated: the mean, median, upper and lower quartiles, and minimum and maximum values. In Fig. 44, an example for the x_shift (longitudinal shift of the forecast object relative to the reference one) is given. A critical shift error value was introduced from the following

reasoning: the larger object areas intersect, the better. If the smallest object size is 35 by 35 grid points, the reasonable shift error in each of two horizontal axes could be chosen as 35 grid points (about 70 km in this case), that is the radius of round objects of 35 by 35 grid points. The green frame indicates that no less than 50% of object shifts do not exceed the critical value. The red frame indicates that all forecast object shifts do not exceed the critical value. It follows that not less than half of the objects are forecasted with an admissible shift in longitude (green frame) up to 90 min. In Fig. 44, the scores are displayed for one radar only (radar in the Kursk town, RAKU). The same aggregation was performed for nine radars in the Central region of Russia. This analysis enabled the following conclusions about the systematic object shift in longitude: the forecast objects are shifted westward (thus going too fast) for radars RAKU and RAVN and eastward (going too slow) for radars RUDK, RUDN, and RUWJ.

RAKU: CRA x_shift statistics OVER SITUATIONS

| lead | min | q25 | med | mean | q75 | max |
|---|---|---|---|---|---|---|
| 1 | -15.280 | -7.620 | -2.680 | -0.609 | 3.990 | 16.730 |
| 2 | -23.480 | -9.408 | -2.505 | -1.519 | 5.140 | 18.240 |
| 3 | -25.920 | -9.505 | -5.345 | -3.125 | 0.588 | 30.640 |
| 4 | -33.920 | -9.970 | -0.940 | -0.467 | 6.923 | 32.660 |
| 5 | -50.030 | -17.730 | -5.530 | -6.011 | 7.145 | 32.350 |
| 6 | -43.530 | -19.230 | -4.490 | -1.933 | 18.115 | 30.640 |
| 7 | -44.900 | -19.790 | 0.800 | -2.318 | 13.760 | 41.780 |
| 8 | -44.210 | -13.613 | -2.990 | -2.133 | 10.455 | 41.160 |
| 9 | -48.090 | -12.640 | -8.150 | -5.302 | 10.428 | 27.000 |

Figure 44: Statistical parameters (mean, median (med), upper (q75) and lower (q25) quartiles, and minimum and maximum values) of CRA x_shift score (longitudinal shift of the forecast object relative to the reference one) calculated over the whole verification period May 1, 2017 September 30, 2017. The green frame indicates that no less than 50% of object shifts do not exceed the critical value, the red frame indicates that all forecast objects do not exceed the critical value. Kursk (RAKU) radar.

The same analysis for the latitudinal shift error showed that the systematic shift is to the north for all nine radars. This can be a feature of the optical flow organization in STEPS. The error does not exceed the empirical critical value of 35 km up to 90 minutes in both latitude and longitude for all the radars.

## 4.4 SAL deterministic study in ARPAE-SIMC (M. S. Tesini and D. D'Alessandro)

### 4.4.1 Method applied (related to an INSPECT Task) and objectives

The SAL method has been applied; it is an object-oriented method proposed by Wernli et al. in 2008 and provides information about structure (S), amplitude (A) and location (L) of a quantitative precipitation forecast (QPF) (See also Sec. 4.1.1) The definition of a threshold allow to identify the precipitation features inside the domain; a smooth parameter is used to filter small scale noise.

### 4.4.2 Short description of the dataset (forecast-observation data), adaptation required, software for the method application

The JDC dataset has been used as observations. This dataset consists of reports from more than 12.000 stations over Central Europe which results in a mean station distance of approximately 16 km. With the aim to interpolate irregularly distributed observations to a regular grid in mountainous terrain, the Vienna Enhanced Resolution Analysis (VERA) scheme has been used (horizontal grid resolution of 8 km). The forecast precipitation fields comes from the run 00 of COSMO-2 (horizontal grid resolution of 2.8 km). Due to the different resolutions, the forecast fields have been interpolated on the VERA grid. The COSMO domain extension has been adapted to the observational domain. The R software (SpatialVx package) has been used to apply the SAL method to six interesting cases occurred during the summer of 2017.

### 4.4.3 Main findings (plots and explanation)

The main result which comes out from the SAL application is that its verification ability is not constant but changes relatively to the considered configurations, in particular it seems to be higher with the decrease in the precipitation fields complexity. A single parameter to evaluate the structure, the amplitude or the location error in forecast is not enough in that cases in which the precipitation field complexity is too high; this becomes evident in Fig. 45: the precipitation intensity is overestimated in Germany and underestimated in France; these errors in forecast compensate each other and determine an amplitude value, referred to the total domain, which is approximately null.

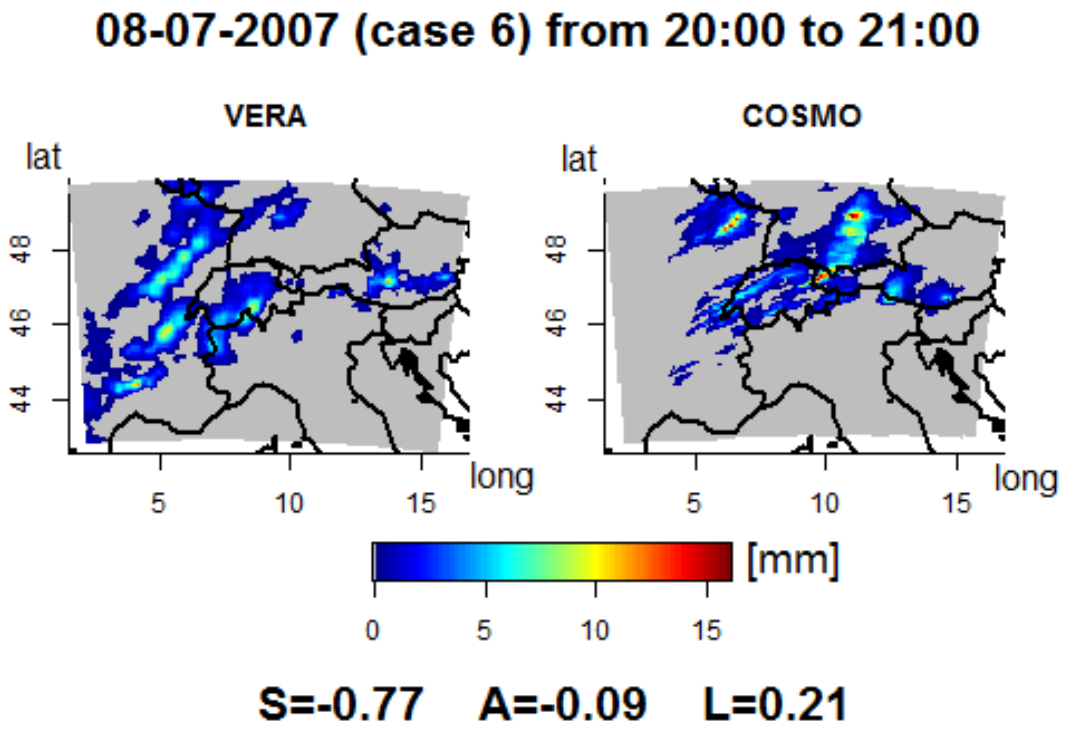



Figure 45: Precipitation intensity and SAL scores, MesoVICT case 6.

As evident from Fig. 46, the SAL scores derived from domains with reduced dimensions are more reliable because they refer to a restricted number of precipitation systems with similar characteristics, avoiding the problem of the loss of information over small scale due to the averaging over the complete domain. This suggests that the choice of the verification domain extension should be done taking it into account. It is clear that this kind of issue emerges for verification over limited accumulation period (1-3h); considering longer periods, the precipitation structures tend to be less fragmented and articulated, allowing the SAL to better evaluate the forecast.

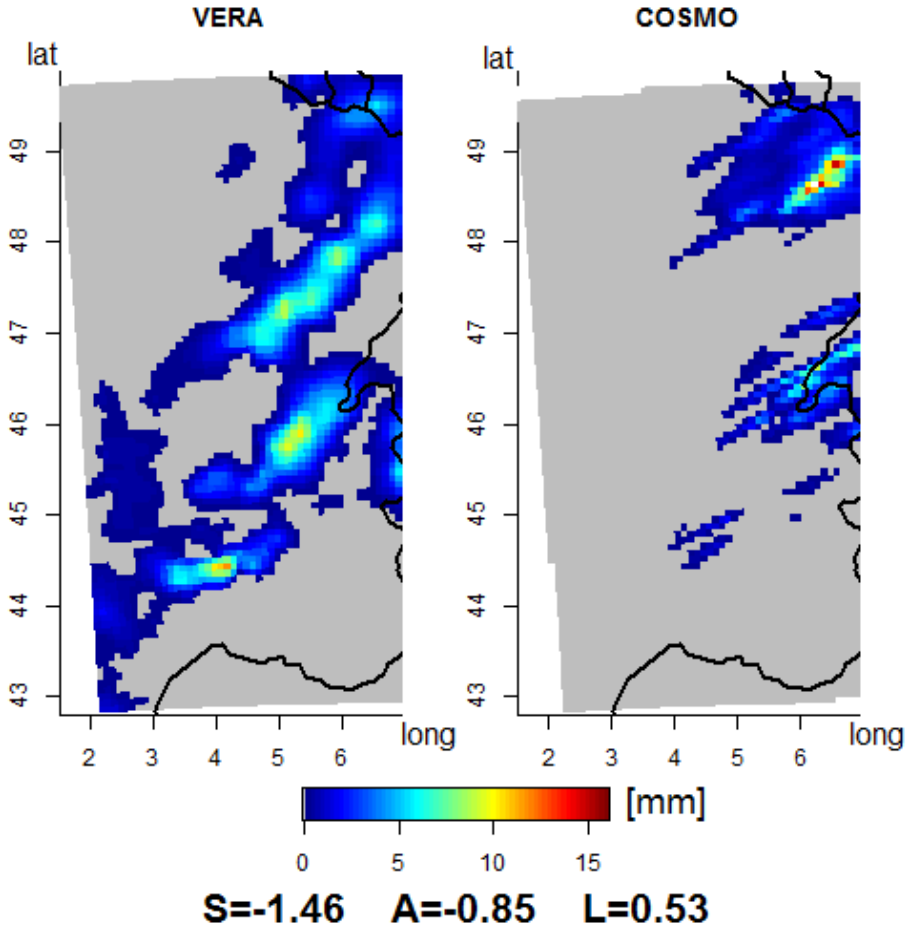



Figure 46: Precipitation intensity and SAL scores, MesoVICT case 6, smaller domain.

### 4.4.4 Characteristics of the method applied

The SAL has not required an excessive computational time for the verification of the considered cases.

The three scores provided by this method allow pointing out the characteristics of a high resolution model. For instance, the structure component S clearly emphasizes the ability of a high resolution model to predict more realistic precipitation features compared to coarser models. The threshold allows identification of the precipitation objects in the domains; the SAL is able to provide information about the location, amplitude and structure errors of these detected features. With respect to the location component L, this is provided as a sum of a L1 and L2 components; L1 gives a measure of the distance between the centres

of mass of the modelled and observed precipitation fields, while L2 takes into account the average distance between the centre of mass of the total precipitation fields and individual precipitation objects. Both L1 and L2 ranges from 0 to +1, so the final component L lies within the range [0,+2]. Based on this definition, different situations can yield the same value of L, in particular there is no sensibility to the rotation around the centre of mass. Furthermore the L value is strictly connected to the maximum dimension of the considered domain. In practical terms it means that, considering the same absolute displacement error, the smaller the domain dimensions, the bigger is L. The amplitude component A is defined as the difference of the domain-averaged precipitation values. The values of A are within [-2,+2] with A = 0 representing the best forecast. Although a zero value is desirable, it is important to remark that it does not necessarily represents a perfect forecast due to the infinity number of possible situations which can lead to identical values of the domain-averaged precipitation. A better indication is obtained integrating the structure component S. The threshold parameter can be a fixed value or a flexible one. The advantage of a fixed threshold is that verification can focus on a particular category, for instance, of intense events, and the statistical results are not blurred by weak events that might be of less interest. However, specification of a fixed threshold excludes poor forecasts from an object-oriented verification in situations in which the threshold is not exceeded in either the model (missed events) or the observations (false alarms).

# 5 Sensitivity of COSMO-LEPS forecast skill to the verification network: application to MesoVICT cases (A. Montani, C. Marsigli, T. Paccagnella, ARPAE-SIMC)

## 5.1 Overall aims

- To test the forecast skill of COSMO-LEPS in terms of total precipitation for different verification networks and different verification methods;
- to understand the meaning of the differences in the verification scores.

## 5.2 Verification datasets

Reference data are the verification networks available in 2007 (Fig. 47):

- **JDC** (**J**oint **D**Phase-**C**ops) dataset. About 12000 observations (mean station distance $\sim$ 12 km);
- **VERA** (**V**ienna **E**nhanced **R**esolution **A**nalysis), gridded analysis at the resolution of 8 km (Dorninger M. et al., 2013).

The model data come from the COSMO-LEPS suite available in 2007 at ECMWF (Montani et al., 2011).

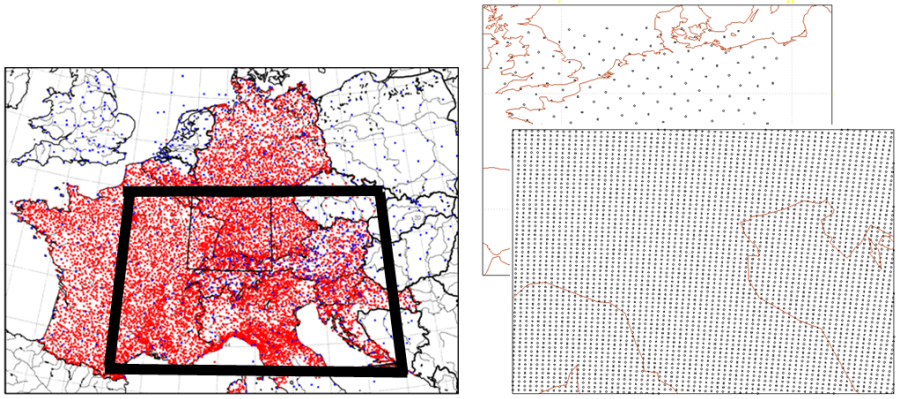


Figure 47: Verification area: DPHASE area (43-50N, 2-18E) (bold black frame).

### 5.3 Verification setup

- variable: 6h cumulated precipitation (0-6, ..., 18-24 UTC);
- period: all 6 mesoVICT cases (Jun  Sep 2007);
- region: 43-50N, 2-18E (D-PHASE area);
- method: nearest grid point, bilinear interpolation, boxes of different sizes;
- forecast ranges: 0-6h, 6-12h, ..., 126-132h;
- thresholds: 1, 5, 10, 15, 25, 50 mm/6h;
- probabilistic scores: ROC area, RPS, Outliers.

### 5.4 Results

In Fig. 48, the precipitation ROC area scores are displayed for two interpolation methods using two different reference datasets (JDC and VERA). For precipitation threshold 1mm/6h (Fig. 48a), the system performance is similar with respect to the 2 verification networks. For precipitation threshold 10 mm/6h (Fig. 48b), higher skill is observed when COSMO-LEPS is verified against VERA gridded analysis. There is almost no impact of the verification technique (nearest grid point versus bilinear interpolation) for both thresholds.

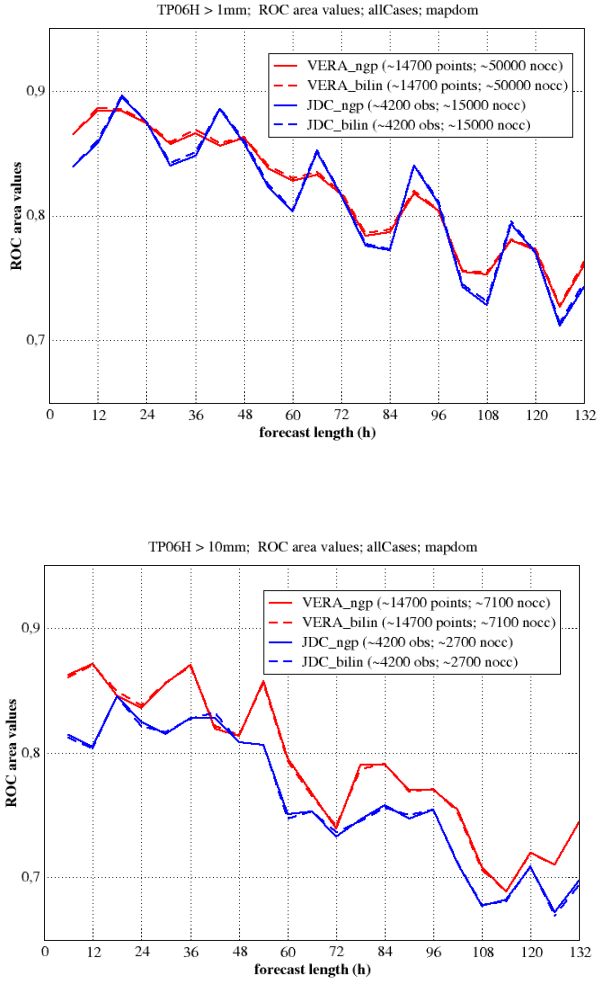
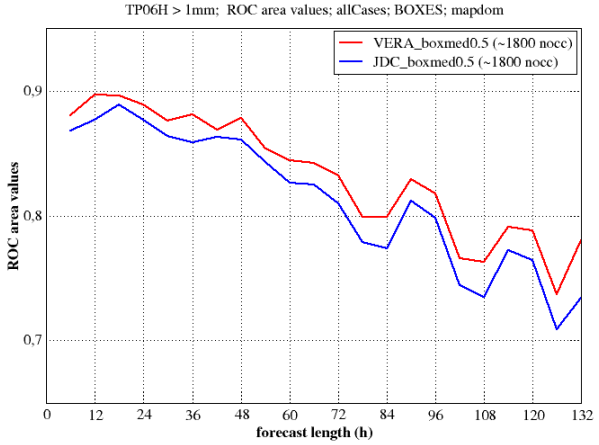
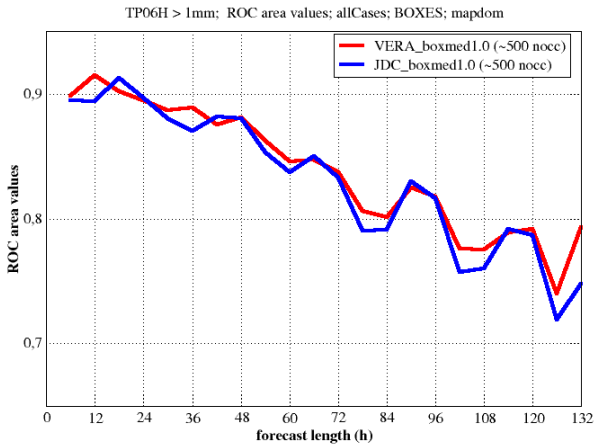



Figure 48: ROC area scores for two interpolation methods using two different reference datasets (JDC and VERA) for precipitation threshold 1mm/6h (top) and 10mm/6h (bottom).

In Fig. 49, a comparison is given between the probabilistic scores in DIST boxes of different size calculated using different reference datasets (JDC and VERA). DIST method is described in Sec. 3.3. It is seen that slightly higher skill is observed when COSMO-LEPS is verified against VERA gridded analysis. The skill increases with increasing box size. There is increasingly less dependence of the score on the verification network for larger boxes.



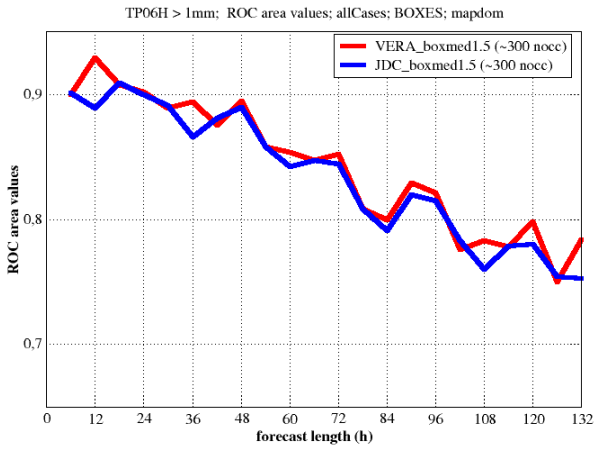




Figure 49: ROC area scores in DIST boxes of different size calculated using different reference datasets, JDC (blue) and VERA(red).

### 5.5 Conclusions

- For nearest grid point versus bilinear interpolation methods, the similar COSMO-LEPS forecast skill is observed using either gridded analysis or sparse observations (VERA or JDC) for verification network;
- for average precipitation in DIST boxes, there are similar scores for verification against gridded analysis or sparse obs for larger and larger boxes;
- as long as we "throw" everything in a box and compare average values (similar results considering the max values), the verification network does not make too much difference.

### 5.6 Future work

- To consider observation uncertainty: work with ensembles of VERA analysis and quantify scores variability;
- to work on higher-resolution ensembles (COSMO-E reruns).

## 6 References


Barrett, A. I., Gray, S. L., Kirshbaum, D. J., Roberts, N. M., Schultz, D. M. and Fairman, J. G., 2015: Synoptic versus orographic control on stationary convective banding. *Q.J.R. Meteorol. Soc.*, **141**, 1101–1113

Zied, B. B. and Theis, S. E., 2014: Spatial techniques applied to precipitation ensemble forecasts: from verification results to probabilistic products. *Meteorol. Appl.*, **21**, 4, 922–929

Bowler N., Pierce C., Seed A., 2006: STEPS: A probabilistic precipitation forecasting scheme which merges an extrapolation nowcast with downscaled NWP. *Q.J.R. Meteorol. Soc.*, **132**, 2127–2155

Casati B., Ross, G. and Stephenson, D. B., 2004: A New intensity-scale verification approach for the verification of spatial precipitation forecasts. *Meteorol. Appl.*, **11**, 141–154

Daubechies, I., 1992: Ten Lectures on Wavelets. *Society for Industrial and Applied Mathematics*, Philadelphia, PA, USA

Davis, C., Brown, B., and Bullock, R., 2006: Object-based verification of precipitation forecasts. Part I: Methodology and application to mesoscale rain areas. *Mon. Wea. Rev.*, **134**, 1772–1784

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

Ebert, E. and McBride, J., 2000: Verification of precipitation in weather systems: Determination of systematic errors. *J. Hydrol.*, **239**, 179–202

Ebert, E., 2008: Fuzzy verification of high resolution gridded forecasts: A review and proposed framework. *Meteor. Appl.*, **15**, 51–64, doi:10.1002/met.25

Gallus, W. A. Jr., 2010: Application of Object-Based Verification Techniques to Ensemble Precipitation Forecasts. *Weather and Forecasting*, **25**, 144–158

Gilleland E., 2017: SpatialVx, https://cran.r-project.org/package=SpatialVx, R package Version 0.6-1

Gilleland E., Ahijevych, D. A., Brown, B. G. and Ebert, E. E., 2010: Verifying Forecasts Spatially. *Bull. Amer. Meteor. Soc.*, **91**, 1365–1373, doi: http://dx.doi.org/10.1175/2010BAMS2819.1

Gilleland,E., Ahijevych, D., Brown, B. G., Casati, B. and Ebert, E. E., 2009: Intercomparison of Spatial Forecast Verification Methods. *Weather and Forecasting*, **24**, 5, 1416–1430

Gofa F., Boucouvala, D., Louka, P. and Flocas, H. A., 2017: Spatial verification approaches as a tool to evaluate the performance of high resolution precipitation forecasts. *Atmospheric Research*, DOI:10.1016/j.atmosres.2017.09.021

Gorgas, T. and Dorninger, M., 2012: Concepts for a pattern-oriented analysis ensemble based on observational uncertainties. *Q.J.R. Meteorol. Soc.*, **138**, 664, 769–784

Johnson, A. and Wang, X., 2012: Verification and Calibration of Neighbourhood and Object-Based ProbabilisticPrecipitation Forecasts from a Multimodel Convection-Allowing Ensemble, *Monthly and Weather Review*, **140**, 3054–3077

Lawson, J. R. and Gallus, W. A., 2016: Adapting the SAL Method to evaluate reflectivity forecasts of summer precipitation in the Central United States. *Atmos. Sci. Let.*, **17**, 524–530

Duc, L., Saito, K. and Seko, H., 2013: Spatial-temporal fractions verification for high-resolution ensemble forecasts, *Tellus A*, **65**, 18171, http://dx.doi.org/10.3402/tellusa.v65i0.18171

Marsigli, C., Montani, A. and Paccangnella, T., 2008: A spatial verification method applied to the evaluation of high-resolution ensemble forecasts. *Met. Apps.*, **15**, 125–143, doi: 10.1002/met.65

Montani, A., Cesari, D., Marsigli, C. and Paccagnella, T., 2011: Seven years of activity in the field of mesoscale ensemble forecasting by the COSMO-LEPS system: main achievements and open challenges, *Tellus A*, **63**, 3, 605–624, DOI: 10.1111/j.1600-0870.2010.00499.x

Muravev A. V., Kiktev, D. B. and Smirnov, A. V., 2017: Operational precipitation nowcasting system based on radar data and verification results for the warm period of the year (May-September 2017), *Hydrometeorological Research and Forecasting*, **1**, 367 (In Russian)

Radanovics, S. (a), 2017: Using the SAL technique for Ensemble forecast Verification 7th International Verification Workshop, Berlin- Presentation

Radanovics, S. (b), 2017: Personal Communication

Radanovics, S., Vidal, J. P. and Sauquet, E., 2015: Probabilistic SAL: Evaluating the spatial properties of probabilistic precipitation simulations, 15th EMS Annual Meeting-Presentation

Steinacker, R., Ratheiser, M., Bica, B., Chimani, B., Dorninger, M., Gepp, W., Lotteraner, C., Schneider, S. and Tschannett, S., 2006: A mesoscale data analysis and downscaling method over complex terrain. *Monthly Weather Review*, **134**, 2758–2771

Eckert, P., 2009: COSMO Priority Project INTERP: Final Report. COSMO Technical Report No. 16, http://www.cosmo-model.org/content/model/documentation/techReports/default.htm

Weniger, M. and Friederichs, P., 2016: Using the SAL technique for spatial verification of cloud processes. *Atmospheric and Oceanic Physics*, 2091–2108

Wernli , H., Paulat, M., Hagen, M. and Frei, C., 2008: SAL - a novel quality measure for the verification of quantitative precipitation forecasts. *Mon. Wea. Rev.*, **136**, 4470–4487

Wernli, H., Hofmann, C. and Zimmer, M., 2009: Spatial Forecast Verification Methods Intercomparison Project: Application of the SAL Technique. *Wea. Forecasting*, **24**, 1472–1484

Weusthoff, T., Ament, F., Arpagaus, M., Rotach, M. W., 2010: Assessing the benefits of convective-permitting models by neighbourhood verification: examples form MAP D-PHASE. *Mon. Wea. Rev.*, **138**, 3418–3433

**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.

No. 10: March 2010.

No. 11: April 2011.

No. 12: April 2012.

No. 13: April 2013.

No. 14: April 2014.

No. 15: July 2015.

No. 16: July 2016.

No. 17: July 2017.

No. 18: November 2018.

***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 nonhydrostatic 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.*

No. 17: D. Leuenberger, M. Stoll and A. Roches (2010):
*Description of some convective indices implemented in the COSMO model.*

No. 18: Daniel Leuenberger (2010):
*Statistical analysis of high-resolution COSMO Ensemble forecasts in view of Data Assimilation.*

No. 19: A. Montani, D. Cesari, C. Marsigli, T. Paccagnella (2010):
*Seven years of activity in the field of mesoscale ensemble forecasting by the COSMO–LEPS system: main achievements and open challenges.*

No. 20: A. Roches, O. Fuhrer (2012):
*Tracer module in the COSMO model.*

No. 21: Michael Baldauf (2013):
*A new fast-waves solver for the Runge-Kutta dynamical core.*

No. 22: C. Marsigli, T. Diomede, A. Montani, T. Paccagnella, P. Louka, F. Gofa, A. Corigliano (2013):
*The CONSENS Priority Project.*

No. 23: M. Baldauf, O. Fuhrer, M. J. Kurowski, G. de Morsier, M. Müllner, Z. P. Piotrowski, B. Rosa, P. L. Vitagliano, D. Wójcik, M. Ziemiański (2013):
*The COSMO Priority Project 'Conservative Dynamical Core' Final Report.*

No. 24: A. K. Miltenberger, A. Roches, S. Pfahl, H. Wernli (2014):
*Online Trajectory Module in COSMO: a short user guide.*

No. 25: P. Khain, I. Carmona, A. Voudouri, E. Avgoustoglou, J.-M. Bettems, F. Grazzini (2015):
*The Proof of the Parameters Calibration Method: CALMO Progress Report.*

No. 26: D. Mironov, E. Machulskaya, B. Szintai, M. Raschendorfer, V. Perov, M. Chumakov, E. Avgoustoglou (2015):
*The COSMO Priority Project 'UTCS' Final Report.*

No. 27: J-M. Bettems (2015):
*The COSMO Priority Project 'COLOBOC': Final Report.*

No. 28: Ulrich Blahak (2016):
*RADAR_MIE_LM and RADAR_MIELIB - Calculation of Radar Reflectivity from Model Output.*

No. 29: M. Tsyrulnikov and D. Gayfulin (2016):
*A Stochastic Pattern Generator for ensemble applications.*

No. 30: D. Mironov and E. Machulskaya (2017):
*A Turbulence Kinetic Energy – Scalar Variance Turbulence Parameterization Scheme.*

No. 31: P. Khain, I. Carmona, A. Voudouri, E. Avgoustoglou, J.-M. Bettems, F. Grazzini, P. Kaufmann (2017):
*CALMO - Progress Report.*

No. 32: A. Voudouri, P. Khain, I. Carmona, E. Avgoustoglou, J.M. Bettems, F. Grazzini, O. Bellprat, P. Kaufmann and E. Bucchignani (2017):
*Calibration of COSMO Model, Priority Project CALMO Final report*

No. 33: N. Vela (2017):
*VAST 2.0 - User Manual.*

No. 34: C. Marsigli, D. Alferov, M. Arpagaus, E. Astakhova, R. Bonanno, G. Duniec, C. Gebhardt, W. Interewicz, N. Loglisci, A. Mazur, V. Maurer, A. Montani, A. Walser (2018):
*COsmo Towards Ensembles at the Km-scale IN Our countries (COTEKINO), Priority Project final report.*

No. 35: G. Rivin, I. Rozinkina, E. Astakhova, A. Montani, D. Alferov, M. Arpagaus, D. Blinov, A. Bundel, M. Chumakov, P. Eckert, A. Euripides, J. Förstner, J. Helmert, E. Kazakova, A. Kirsanov, V. Kopeikin, E. Kukanova, D. Majewski, C. Marsigli, G. de Morsier, A. Muravev, T. Paccagnella, U. Schättler, C. Schraff, M. Shatunova, A. Shcherbakov, P. Steiner, M. Zaichenko (2018):
*The COSMO Priority Project CORSO Final Report.*

No. 36: A. Raspanti, A. Celozzi, A. Troisi, A. Vocino, R. Bove, F. Batignani (2018):
*The COSMO Priority Project VERSUS2 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), Romania (NMA) and Russia (RHM). 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 editor:

Massimo Milelli
*Massimo.Milelli@arpa.piemonte.it*