16 Overview of methods for the verification of quantitative precipitation forecasts

Andrea Rossa[1], Pertti Nurmi[2], Elizabeth Ebert[3]

[1]Centro Meteorologico di Teolo, ARPA Veneto, Italy
[2] Meteorological Research, Finnish Meteorological Institute, Finland
[3]Centre for Australian Weather and Climate Research, Bureau of Meteorology, Australia

Table of contents

16.1 Introduction

In the area of hydrological risk management, both Quantitative Precipitation Estimates (QPE) and Quantitative Precipitation Forecasts (QFE) are key in quantifying the potential for flooding, especially on the short time scales, i.e., for relatively small river and urban catchments. In such a context, forecasting can be viewed as the attempt to reduce the uncertainty of the future state of the hydrometeorological system and so

anticipate mitigating actions. Authorities, however, often are still reluctant to devise and invest in such actions based on forecasts when their quality is unknown. In other words, for forecasts to be useful and effective the forecast quality and forecast uncertainty must be quantified.

Much effort has been and is being invested in the quest of working with imperfect precipitation observations and forecasts. A number of initiatives are underway, such as the Hydrologic Ensemble Prediction EXperiment (HEPEX, Schaake et al. 2007) which is an international project established by the hydrological and meteorological communities. The mission of HEPEX is to demonstrate how to produce reliable hydrological ensemble predictions that can be used with confidence by emergency management and water resources sectors to make decisions that have important consequences for economy, public health and safety. The COST 731 Action (Rossa et al. 2005) is a European initiative which deals with the quantification of forecast uncertainty in hydrometeorological forecast systems. It is linked to the MAP D-PHASE initiative (www.map.meteoswiss.ch), a WWRP Forecast Demonstration Project (FDP), which is to provide evidence of the progress meteorological and hydrological modeling has achieved over the last decade or so. A characteristic of an FDP is that strict evaluation protocols are established to demonstrate and document such progress. Indeed, many atmospheric and hydrological forecast systems participate in this effort. The atmospheric part includes nowcasting based on radar, very high resolution next-generation numerical weather prediction (NWP) models, operational models, as well as a number of limited area ensemble prediction systems.

In all of this verification, and verification of precipitation forecasts in particular, is fundamental! It is safe to say that the more detailed the forecasts the more complex the corresponding verification task. For example, verification of geostrophic flow can be viewed as relatively simple when compared to verification of turbulent flow. Precipitation is a stochastic quantity and exhibits fractal properties down to very small scales (e.g., Zawadzki 1973). It is difficult to observe, simulate and to verify. Furthermore, much more efforts have been invested in the development of forecasting techniques than in verification methodologies. This may be connected to the fact that the traditional approaches to verification of gridded forecasts were developed on relatively low resolution global NWP models to check the consistency of upper air fields against model analyses. Stanski et al. (1989) provide a thorough compilation of the statistics involved in NWP verification,

while Wilks (2006) is an excellent text and reference book for statistical methods in the atmospheric sciences, covering forecast verification.

However, with increasing resolution of the limited area models, verification of weather elements against observations has become a more complex problem. For example, while for medium-range forecasting typically daily rainfall accumulations are verified, the higher resolution meso-scale models are expected to have skill also in shorter time scales. Their performance is tested for shorter accumulation periods where for instance the timing and location of a frontal passage is essential and the traditional verification methods are not necessarily sufficient. Small positioning errors in the forecasts may result in the so-called 'double penalty': the verification measure tends to penalize rather than reward the model's capability to provide some sort of information on small scale features (see Sect. 16.3.2).

These issues are accentuated when it comes to verifying high resolution QPFs. The necessity to evaluate and justify the advantages of the ever higher resolution over the computationally less expensive coarser resolution NWP in terms of QPF quality has stimulated radically different verification approaches for spatial forecast fields over the last decade or so. These methods go well beyond point-to-point pair verification and borrow ideas from fields such as image and signal processing. The main lines of extension to judge whether or not a precipitation forecast for a given time and location is correct is to ask the question whether the main *characteristics* of fields are captured in the simulation. In other words, conditions for right and wrong are relaxed from 'at a given point and time' in several ways. For example, in the class of neighborhood methods the condition of correct location is successively relaxed to yield an effective scale-dependent measure of forecast goodness (Ebert 2008). Harris et al. (2001) investigate whether the characteristic scales of rainfall fields are successfully reproduced, without necessarily requiring correspondence in location, while Ebert and McBride (2000) look for corresponding rain objects and decompose the measure for quality in components for matching location, amount and structure. Davis et al. (2006) take the description of precipitation objects one step further but still require object matching between the forecast and the observations.

For hydrological applications the localization of precipitation is important on the scale of the considered catchment, so that it is useful to perform QPF verification on river basin averages (e.g., Oberto et al. 2006). Wernli et al. (2008) combine the idea of verifying precipitation within a predefined area, say a medium to large river catchment, in which not just the average rainfall amount is evaluated but also the

average capability of the model to predict location and structure of the rainfall field, measures that do not require object correspondence.

Datasets on which these methodologies are applied can span several years in order to try to document improvements in forecast quality. Improvements have been reported for parameters like the pressure or the temperature, but not for QPF (Hense et al. 2003). Performing verification over a full year will effectively mix a number of different flow regimes which, in theory, can present different challenges to a modeling system. Also, the verification results can be biased towards the most frequent regime, e.g. days with no intense weather. It is, therefore, quite common practice to differentiate verification for the four seasons, while it is far less common to perform a systematic separation of distinct flow regimes in which a forecast system may have different challenges to get realistic QPF.

The diversity of approach emerging from these examples, which are detailed further in Sects. 16.3.3 and 16.3.4, document the efforts of the scientific and operational community to find adequate measures to describe forecast quality of high resolution QPF. However, such a variety holds the risk that verification results become difficult to compare. There have been several efforts to harmonize verification activities in the recent past. ECMWF, for example, compiled a set of recommendations for their member states (Nurmi 2003), while the Working Group of Numerical Experimentation (WGNE) provided a survey of verification methods of weather elements and severe weather events (Bougeault 2002) and recommendations for the verification and intercomparison of QPFs from operational NWP models (WGNE 2004). There is an ongoing exercise in which the more recent verification techniques are to be compared on a set of common cases (ICP 2007).

Probabilistic QPF is a promising avenue of improvement for high resolution rainfall prediction (e.g., Mittermaier 2007). The main ideas behind probabilistic forecasting are based on the imperfect knowledge of initial conditions and key parameters in parameterization schemes of mainly moist processes. Ensemble forecasting, i.e., forecasts starting from slightly differing initial conditions, is an established technique for estimating forecast uncertainty of the global models in the medium range. It has become increasingly popular also for high resolution limited area models in shorter time ranges, as well as in nowcasting. The radar community has started recently to produce probabilistic QPEs based on the error characteristics of radar measurements (Germann et al. 2006). Probabilistic forecasting is adding considerable complexity to the verification problem in that 'right and wrong' no longer have a strict sense when it comes to a single forecast observation pair. Verification

needs to take the frequency of occurrence of events into account. These issues are, however, beyond the scope of this Chapter and will not be discussed.

This contribution aims at providing an overview on the standard techniques used in QPF verification and on recent, more sophisticated approaches, in order to provide a panorama of the tools available. The choice of technique for QPF verification may well be purpose-dependent, be it in hydrological applications for one or more catchments, in road weather forecasting for distinct stretches, or for model development where identifying specific model weaknesses is the necessary first step for improvement. It is, therefore, a specific goal of this writing to provide some sort of recommendations or guidelines to the collection of methods. For the sake of convenience, many of the illustrations are taken from the COSMO model (Steppeler et al. 2003), but the applied methods are by no means tied to this particular model. They are not even specific to NWP but can be applied to other comparisons of precipitation fields, e.g., QPE from different sensors (e.g., Ebert et al. 2007). An additional Chapter on QPF verification is presented by Tartaglione et al. (Chap. 17 in this book).

Section 16.3.2 reviews traditional verification scores and illustrates their limits for high resolution QPF verification. Section 16.3.3 deals with scale-dependent verification, while Sect. 16.3.4 with object-oriented approaches. Stratification of data sets to isolate model behavior in specific flow situations is dealt with in Sect. 16.3.5 before some recommendations are given in Sect. 16.3.6 as to the relative merits of the various techniques which have been discussed.

16.2 Traditional verification of QPF and limitations for high resolution verification

The strategy for any forecast verification application includes certain rational steps: choosing and matching a set of forecast/observation pairs, defining the technique to compare them, aggregating (pooling) and/or stratifying the forecast/observation pairs in appropriate data samples, applying the relevant verification statistics and, ultimately, interpreting the scores, not forgetting to analyze the statistical significance of the gained results. The latter is unfortunately quite often neglected both in verification studies as well as in operationally run forecasting systems.

Deterministic QPFs can be formulated and taken as either *categorical events* or *continuous variables* and verified correspondingly

utilizing respective verification approaches and measures. Verifying QPFs as categorical events is clearly more common. The categorical approach involves issues like whether or not it rained during a given time period (rather than at a given instant) or, alternatively, whether the rainfall amount exceeded a given threshold. Verifying rainfall amount as a continuous variable brings about certain caveats because the rainfall amount is not a normally distributed quantity. Very large rainfall amounts may be produced by a forecasting system and, then again, in some cases very little or no rain. Many of the verification scores for continuous variables, especially those involving squared errors, are very sensitive to large errors. Consequently, categorical verification scores provide generally more meaningful information of the quality of the forecasting systems (or skill of the human forecasters) producing QPFs.

16.2.1 Common scores

There are a number of recent textbooks (Wilks 2006; Jolliffe and Stephenson 2003) and papers (Nurmi 2003; Bougeault 2002; Wilson 2001) as well as the website of the WMO/WWRP Joint Working Group on Verification (www.bom.gov.au/bmrc/wefor/staff/eee/verif/verif_web_page.html), which detail the traditional precipitation verification methods and give an exhaustive account of their features. Reference is made to these publications rather than elaborating on these attributes here. However, a general definition and a short overview of the most common scores for the verification of categorical QPFs will follow, accompanied by brief comments of their pros and cons. Occasional references are made to current literature where one can embrace a deeper understanding of the behavior of these measures. Some additional, more recent scores are also introduced. Although these cannot be considered 'traditional' they are covered because they fit the framework properly.

Categorical events - Forecasts of the exceedance of precipitation thresholds

The joint distribution of binary (yes/no) forecasts and associated observed events and non-events is unambiguously defined by the four elements of a 2x2 contingency table: *hits, false alarms, misses* and *correct rejections*. Categorical statistics are applied to evaluate these binary events which in our case is the accumulated rainfall amount during a given time period exceeding a specified threshold. The most

popular event is rainfall exceeding a threshold taken as rain versus no-rain. This threshold varies from country to country but is generally between 0.1 mm and 0.3 mm of accumulated precipitation during a 24-hour (or a 12-hour) period. These different definitions may have huge effects on the verification results (and their interpretation) since, as shown later, many of the categorical forecast verification measures are highly dependent on the observed frequency (or the *base rate*) of the event.

The seemingly simple definition of the binary event and the subsequent contingency distribution and its associated marginal distributions of forecasts and observations accommodate quite amazing complexity and there exist a large number of measures to tackle this ambiguity. Most of these scores have historical credentials as long as the history of forecast verification, dating back to the late 19th century. Consequently, they have been 're-invented' and renamed many times during later times.

The *Frequency Bias Index* (FBI) compares, as a ratio, the frequency of forecasts with the frequency of actual occurrences of the event. It ranges from zero to infinity and the optimal value for an unbiased forecasting system is one. The frequency bias is not a measure of accuracy as it does not provide information on the magnitude of forecast errors.

Probably the simplest and most intuitive performance measure providing some information on the accuracy of categorical forecasts is the *Proportion Correct* (PC) which gives the fraction of all correct forecasts (i.e., of the event and the non-event). This simplistic measure is easily very misleading since it rewards correct 'yes' and 'no' forecasts equally and is strongly influenced by the more common category which is normally the more uninteresting non-event. A prime educational example of the interpretation of PC and its consequences is the often cited, legendary Finley case (Finley 1884; Murphy 1996; see also e.g., Wilks 2006, pp. 267-268).

The *Probability Of Detection* (POD) measures the fraction of observed events that were correctly forecast, whereas the *False Alarm Ratio (FAR)* measures the fraction of forecast events that were observed to be non-events. In some literature, POD is called the *hit rate*, having as its complement the *miss rate* which gives the relative number of missed events. POD and FAR must always be examined together as neither of them is really adequate on its own. POD is sensitive to hits only and does not take into account false alarms, whereas FAR is sensitive to false alarms but takes no account of misses. Both of them can be artificially improved by producing excessive 'yes' forecasts (in

the case of POD) or 'no' forecasts (to improve FAR). Such bogus human forecasting behavior is often called *hedging*. FAR is very sensitive to the climatological frequency of the precipitation event (Fig. 1, left panel), which is a property quite common to many of the traditional verification measures.

While FAR is a measure of false alarms given the forecasts, *False Alarm Rate (F)* is a kindred measure which measures the false alarms given the observed non-events. It is also called the *Probability Of False Detection* (POFD). F is almost exclusively associated with the verification of probabilistic QPFs by combining it with the hit rate to produce the so-called *Relative Operating Characteristic* (ROC) diagram or curve. The ROC measures the ability of the forecast to discriminate between observed events and non-events and is commonly used in the verification of probabilistic forecasts (for more on ROC diagrams see Jolliffe and Stephenson 2003).

A popular, historical measure for verifying categorical forecasts results from simply subtracting F from POD. This skill score has many 'inventors' and therefore many names, like *True Skill Statistics* (TSS), *Peirce Skill Score* (PSS) and *Hanssen-Kuipers Skill Score* (KSS). Idealistically, it measures the skill of a forecasting system to distinguish the 'yes' cases from the 'no' cases. It also measures the maximum possible relative economic value attainable by a forecast system, based on a Cost-Loss model (Richardson 2000). For rare events (e.g., heavy precipitation) the frequency of correct rejections is typically very high, leading to a very low F and, consequently, the score asymptotes to POD.

Another commonly used performance measure, especially for rare events is the *Threat Score* (TS), also known as the *Critical Success Index* (CSI). It is defined as hits divided by the sum of hits, false alarms and misses. Because TS takes into account both false alarms and misses it can be considered as a simple measure that tries to remove from consideration correct forecasts of the (simple) non-events. However, TS is known to be sensitive (again) to the local climatology of precipitation (Fig. 1, center panel). To overcome this feature the otherwise similar *Equitable Threat Score* (ETS) aims at removing the effects of hits that occur purely due to random chance. (Fig. 1, right panel)

One of the most widely used skill scores is the *Heidke Skill Score* (HSS). Its reference accuracy measure is Proportion Correct which is adjusted to eliminate forecasts which would be correct due to random chance. The HSS is related to ETS via a direct relationship and therefore does not provide any additional information.

The *Odds Ratio* (OR) measures the forecasting system's probability (odds) to score a hit (H) as opposed to the probability of

making a false alarm (F). It produces typically high numeric values because a no-skill system would equal one and a perfect system yields a score of infinity. A transformed *Odds Ratio Skill Score* (ORSS) is scaled to have values in the range [-1, +1] to be comparable with other verification scores of categorical events. The Odds Ratio cannot be considered a traditional verification measure and has been applied very scarcely in meteorological (QPF) verification. Nevertheless, it is advocated to possess several attractive properties (Göber et al. 2004; Stephenson 2000).

Stephenson et al. (2007) have proposed a new score, the *Extreme Dependency Score* (EDS), specifically for the verification of rare events like heavy QPF. The score is reported to be insensitive to the base rate, is not dependent on the potential frequency bias of the forecasts and will not encourage hedging.

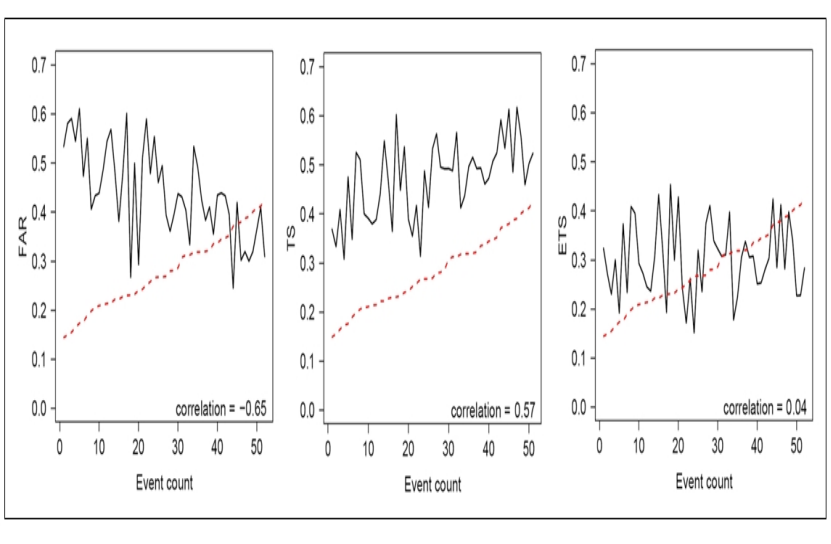


Fig. 1. Correspondence between a categorical verification measure (continuous line) and the frequency of observed rain events (dashed line) at a given observing station, for FAR (left), TS (center) and ETS (right). The rain event is defined as rainfall exceeding 0.3 mm during 24 hours and the vertical axis shows the relative frequency of such events arranged in ascending order

Continuous variables - Forecasts of time-integrated accumulated precipitation

As already discussed in the previous Section, deterministic QPFs are often formulated as *continuous variables*, but perhaps more often

verified as categorical events. Verification of continuous QPFs as such commonly involve statistics on how much the absolute forecast values depart from the corresponding observations, as well as the computation of relative (skill) measures against reference forecasts like climatology and persistence. What follows is a very brief description of the most common (traditional) verification methods applicable for the verification of continuous QPFs. In general, the intermittent and non-Gaussian distribution of precipitation strongly affects these measures which are generally sensitive to large errors.

The *mean value (arithmetic mean)* is always very useful to put forecast errors (see below) into their perspective. To define variability in rainfall the *sample variance* and the *sample standard deviation* are often used. The latter is conveniently in the same units as the original precipitation, being the square root of the previous.

The *Mean Error* (ME), or *bias*, is simply the arithmetic average of the difference between forecasts and observations. Like the frequency bias in the case of categorical QPF events, it is not an accuracy measure and does not produce information on the magnitude of forecast errors. The *Mean Absolute Error* (MAE) compensates for positive and negative forecast errors and is a scalar measure of forecast accuracy. The ME and the MAE viewed together provide useful information on the general behavior of forecast errors.

The *Mean Square Error* (MSE) is the average squared difference between forecasts and observations. Taking a square root of MSE produces the *Root Mean Square Error* (RMSE) which has the same units as the original entity. Due to the second power of these scores they are much more sensitive to large forecast errors than the MAE, which may be quite harmful in the presence of outliers in the dataset. The *correlation coefficient* (r) measures the degree of linear association between forecast and observed values, independent of absolute or conditional biases. This score is very sensitive to large errors and benefits from the square root transformation of precipitation amounts. The fear for high penalties when applying squared verification measures may easily lead a human forecaster to conservative forecasting (i.e., hedging).

Many of these accuracy measures, especially the MAE and the MSE, are commonly used to construct a *skill score* that measures the fractional (percentage) improvement of the forecast system over a reference forecast. The reference estimate is preferably persistence for forecasts with a lead time of c. 24 hours or less and climatology for longer range forecasts.

16.2.2 The double penalty issue

The traditional point-matching categorical and continuous verification measures are quite intuitive, easy to perceive and, above all, they have been used for many decades. There is no urge to cease applying them as long as their pros and cons and occasionally notorious behavior is known, understood and acknowledged. Whatever their pitfalls the traditional standard verification measures still do return optimum scores for, hypothetically, optimum forecasts, regardless of the underlying properties, like the resolution, of the models that produce these forecasts. A further aspect that favors preserving existing common verification methods is the lengthy time lag before new innovations in forecast verification research are mature enough to be accepted by the community and applicable for common use.

The verification endeavor has become more and more demanding and from a scientific perspective increasingly rewarding, during recent years with the continuously enhanced resolution of the NWP models, resulting effectively also in the detail in which a human forecaster depicts the weather. Today it is not uncommon to have detailed local, site-specific QPFs for several days ahead as compared to earlier times when forecasts were formulated rather as area and time-averaged entities. The most obvious and meaningful way to produce time/space focused precipitation forecasts would be using a probabilistic approach but the verification of probabilistic QPFs (PQPF) is not covered in this Chapter. Nevertheless, as long as NWP models do produce categorical QPFs, their quality needs to be evaluated from this perspective.

Let us consider, for example, a model forecast low pressure pattern having a phase error of half a wavelength and another model having not forecast the pattern at all. The former model would be punished twice, for not having the low where it is supposed to be and, secondly, for having the low where it is not supposed to be (*double penalty*). The latter model, however, would get penalized for only not having forecast the pattern.

It is quite common that high resolution, meso-scale, forecast models produce forecasts with seemingly realistic small scale (precipitation) patterns but with amplitude and gradients which may be somewhat misplaced. In the case of convective precipitation and/or narrow frontal rainbands such misplacements are hardly surprising but may show up as quite dramatic results when verified with common verification measures. The timing and space errors will result in a much larger RMSE than for the smoother lower resolution model forecast.

There are seldom, if ever, trivial cases in the real atmosphere. Figure 2 shows a comparison between a global and a regional NWP model in a case with well-defined precipitation patterns. The regional model shows some explicit small-scale structures with a reasonably realistic amplitude, albeit somewhat misplaced, when compared to a radar-based quantitative precipitation estimate (QPE), taken that the radar-based analysis is realistic. On the other hand, there are features in the global model (indicated by ovals) which are almost totally missing from the regional model. It would be quite hard to interpret intuitively or visually (applying *'eyeball verification'*) and even with objective verification measures which one of the forecasts is the better. As a matter of fact, one would need to first define the purpose of the forecast (end-user, application, etc) and of the verification.

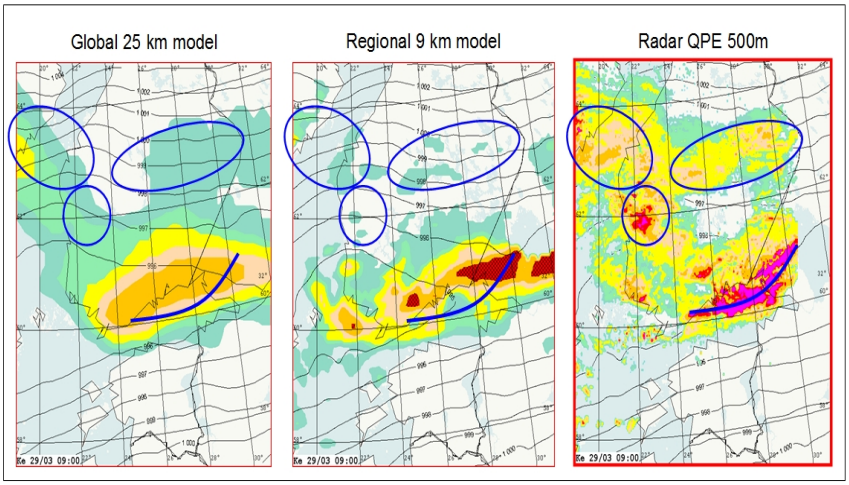


Fig. 2. A comparison of two NWP models operating at different spatial resolutions and the corresponding radar-based QPE. Some of the main features are indicated by the ovals and the arched curves

Table 1 compares three NWP models operating at different resolutions. The QPF verification is done against three different 'observed truths', one based on rain gauge data, one on radar-based QPE and the third on merging these two data sources (the merging method is not relevant here). The results are somewhat mixed and incoherent. The highest scores (underlined numbers) are mostly gained for the coarse-scale model and the lowest ones (numbers with shaded backgrounds) for the fine-scale model, hinting at double penalty reminiscent behavior. However, the scores reflect also quite strongly the observation (or

analysis) type which has been applied as the 'truth' behind the verification. The important role of observations in verification is elaborated further in the Chap. 17, by Tartaglione et al., later in this book. The example here is presented merely to emphasize the complexity of verification. Nevertheless, it is advisable to use combined gauge-radar precipitation analyses whenever possible when verifying QPFs at high temporal and spatial resolutions.

Table 1. Verification statistics for three NWP models operating at different spatial resolutions and applying three different analysis types

	Global 25 km model			Local 22 km model			Local 9 km model		
	Gauge	Merge	Radar	Gauge	Merge	Radar	Gauge	Merge	Radar
Bias	1.44	1.45	.59	1.73	1.66	.73	.59	.59	.25
MAE	.33	.32	.47	.48	.48	.69	.31	.33	.67
RMSE	.57	.53	.93	.78	.77	1.14	.63	.66	1.25
r	.62	.70	.72	.40	.46	.42	.42	.45	.32
POD	.74	.74	.51	.65	.59	.39	.50	.42	.20
FAR	.49	.49	.14	.63	.64	.47	.15	.28	.19
KSS	.66	.66	.48	.52	.46	.28	.49	.41	.19
ETS	.38	.38	.39	.25	.22	.18	.43	.33	.15

The double penalty may be interpreted in terms of the categorical precipitation verification terminology: a forecast is penalized twice, for not getting the precipitation at the correct location (*miss*) and forecasting the precipitation at the wrong location (*false alarm*). This is illustrated schematically in Fig. 3 (a) where a high resolution forecast (left) would attain dramatically worse scores than its low resolution competitor (right) although the shape and amplitude appear perfect on the high resolution output.

The differences in scores are exclusively due to the misplacement of the entities. Applying a spatial translation and matching of the forecast pattern of the high resolution system with the observed field would result in the schematic shown as Fig. 3 (b). Such an exercise would result in perfect scores in this simplistically naive example. Object matching verification techniques are elaborated further in Sect. 16.4.

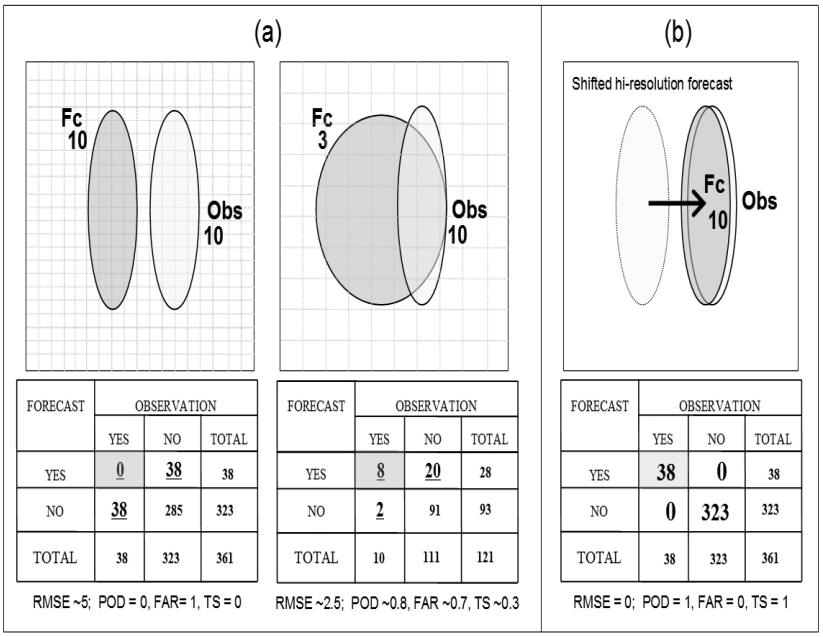


Fig. 3. A schematic of the double penalty effect on a high resolution forecast (a; left) compared to a low resolution forecast (a; right) and after applying the spatial matching technique such as that of Ebert and McBride (2000) (b)

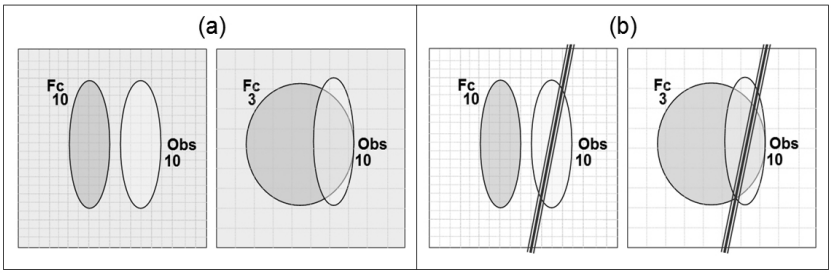


Fig. 4. A schematic of two different hypothetical forecast/verification applications, a hydrological catchment, indicated by the shading of the square domain (a) and a highway stretch, shown by the vertical line through the domain (b)

It is required in forecast verification, likewise in weather forecasting, that the target (end-) users and the purpose of verification/forecasting are known beforehand. This issue is briefly underlined using our previous example. The forecast/verification area of interest might be a distinct hydrological catchment area (indicated as the

shaded rectangular area in Fig.4 (a)), or the focus of interest might be along a highway stretch (represented by the thick vertical line in Fig. 4 (b)). The quality of the forecasts would be evaluated quite differently for these two applications.

16.3 Scale-dependent techniques

As just seen, as precipitation forecasts from models and nowcasts are made at increasingly higher spatial and temporal resolution, the ability of the forecast to achieve an exact match with the observations becomes more difficult owing to the double penalty issue. The question becomes, if poor skill is shown at fine scales then at what scales does the forecast skill become acceptable? Scale-dependent verification methods address this question by measuring the correspondence between the forecast and the observations on a variety of space and time scales.

16.3.1 Neighborhood methods

Neighborhood (sometimes called 'fuzzy') verification approaches reward closeness by relaxing the requirement for exact matches between forecasts and observations. Ebert (2008) describes a framework for neighborhood verification using multiple methods. Some of these methods compute standard verification metrics for deterministic forecasts using a broader definition of what constitutes a 'hit'. Barnes et al. (2007) propose a conceptual framework to take into account close calls when evaluating U.S. National Weather Service weather warnings. Other methods treat the forecasts and/or observations as probability distributions and use verification metrics suitable for probability forecasts. Implicit in each neighborhood method is a particular decision model concerning what constitutes a good forecast. For example, one decision model could be that a good forecast must predict at least one event near an observed event.

The key to this approach is the use of a spatial window or neighborhood surrounding the forecast and/or observed points. The treatment of the points within the window may include averaging (upscaling), thresholding, or generation of a probability density function, depending on the metric used. Some methods compare neighborhoods of forecasts with neighborhoods of observations, while others compare the forecast neighborhood with the observation in the center of the neighborhood. Starting with the finest scale (neighborhood

of one grid box) the size of the neighborhood is increased to provide verification results at multiple scales, thus allowing the user to determine at which scales the forecast has useful skill. Multi-dimensional windows can be used to represent closeness in space, time, intensity, and/or some other aspect.

Three of the most useful of the neighborhood techniques are described in this Section. They are demonstrated by verifying a high resolution (0.02º) forecast from the COSMO model against high-quality radar observations over Switzerland (Leuenberger 2005).

As seen in Fig. 5 the model predicted the rainfall structure quite well. However, the ETS (see Sect. 16.3.2) computed at grid scale for a 0.1 mm threshold was only 0.33. This illustrates the need for verification methods that give credit to 'close calls' and 'near misses'.

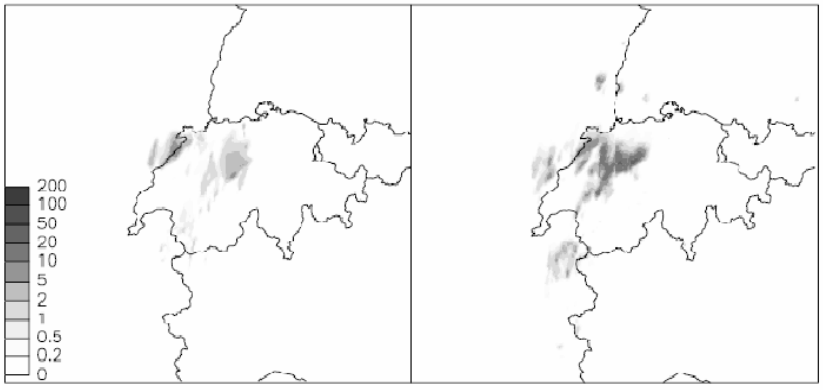


Fig. 5. 17h forecast from the COSMO model (left) and radar quantitative precipitation estimate (right) of hourly rainfall accumulation (mm) over Switzerland ending 17:00 UTC on 8 May 2003 (from Leuenberger 2005)

The most widely used neighborhood verification technique is upscaling, in which forecasts and observations are averaged to increasingly larger grid scales for comparison using a range of standard statistics (e.g., Zepeda-Arce et al. 2000; Cherubini et al. 2002; Yates et al. 2006). The implied decision model is that a good forecast has a similar mean rain amount as the observations. The upscaling verification of the COSMO forecast is shown in Fig. 6 in which the ETS is plotted as a function of spatial scale and rain intensity.

The verification scores generally improve with increasing scale and smaller rain thresholds, as expected. A relative peak in performance for the heavier rain rates is seen at a spatial scale of about 0.3 degrees, consistent with the good placement of the rain maximum.

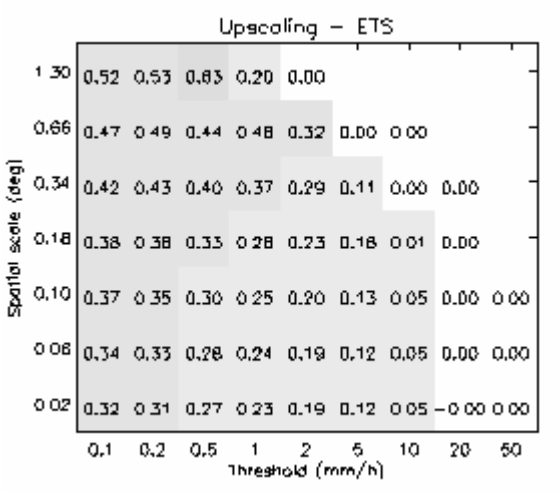


Fig. 6. Equitable threat score for the COSMO forecast shown in Fig. 5, as a function of spatial scale and rain threshold, when upscaling is used to average forecasts and observations to larger scales

Atger (2001) developed a multi-event contingency table method for comparing high resolution gridded rainfall forecasts to point observations. In this approach closeness is evaluated simultaneously in two or more 'dimensions' (spatial proximity, temporal proximity and similarity of rain intensity). A hit is counted whenever a forecast event is sufficiently close to an observed event. Multi-dimensional contingency tables are generated for varying thresholds, from which the hit rates can be plotted against the false alarm rates as points on a Relative Operating Characteristic (ROC) diagram.

The ROC in Fig. 7 suggests that the COSMO forecast successfully predicted rain close to where it was observed, both in terms of spatial location and intensity.

The fractions skill score (FSS) method of Roberts and Lean (2007) compares the forecast and observed fractional occurrences of rain exceeding a given threshold. The FSS is a version of the Brier Skill Score (Jolliffe and Stephenson 2003) in which the observed occurrence is the event fraction within the neighborhood and the reference forecast is the no-overlap forecast. Roberts and Lean showed that the target value of FSS above which the estimates are considered to have useful skill is given by $0.5+f_{obs}/2$, where f_{obs} is the frequency of observed events over the full domain. The FSS values for the COSMO forecast (Fig. 8) are greater for light thresholds and larger scales, with useful

skill displayed at spatial scales of 0.1 degree and larger for light rain, 0.2 to 0.7 degrees for moderate rain, and not at all for the heaviest rain rates.

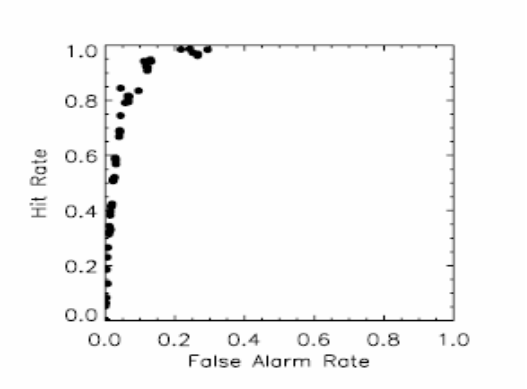


Fig. 7. Relative Operating Characteristic for the COSMO forecast shown in Fig. 5. Each point shows the hit rate and false alarm rate for a particular combination of spatial scale and rain intensity threshold

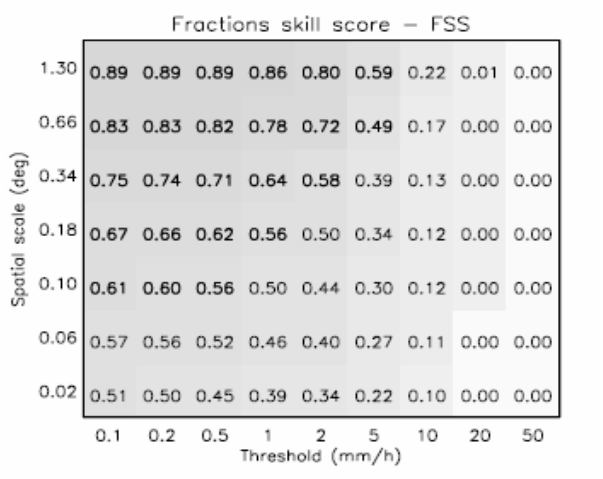


Fig. 8. Fractions skill score for the COSMO forecast shown in Fig. 5. FSS measures the similarity of the forecast and observed rain fractions for a variety of spatial scales and rain thresholds. The bold values indicate useful skill

16.3.2 Spatial decomposition methods

Another type of scale-dependent verification uses a spatial filter to decompose or separate the gridded forecasts and observations into different spatial scales and then computes the error separately for each scale. The scale-dependent errors sum to the total error. Scale decomposition allows errors associated with different phenomena to be isolated and identified. Several spatial filters have been proposed, including 2D Fourier transforms (Stamus et al. 1992), discrete cosine transforms (de Elia et al. 2002) and 2D discrete wavelet filters (Briggs and Levine 1997; Casati et al. 2004). Once the scale separation has been accomplished different continuous, categorical and probabilistic verification metrics may be applied.

In particular, the intensity-scale method of Casati et al. (2004) uses thresholding to convert the forecast and observations into binary images. Wavelet decomposition is applied to the binary error image and a skill score based on the mean squared error is computed for each scale. Figure 9 shows the application of the intensity-scale method to the COSMO forecast. The lowest skill is associated with small scales and high rainfall intensities while the greatest skill is found at large scales.

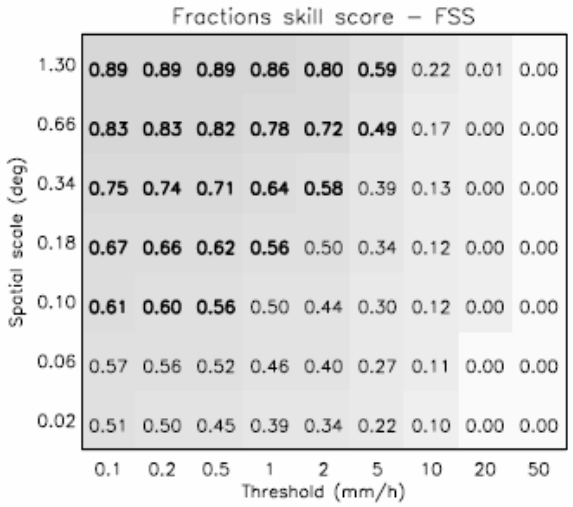


Fig. 9. Scale-dependent Heidke skill score for the COSMO forecast shown in Fig. 5, computed using the intensity-scale method of Casati et al. (2004)

If the aim of the verification is to compare the multi-scale statistical properties of the forecast rainfall to those of the observed rainfall, that is, to evaluate whether the forecast rain looks realistic,

regardless of the actual placement of the rain relative to the observations, then the multi-scale approach described by Harris et al. (2001) may be used. They compare the power spectrum, structure function and moment scaling analysis of high resolution model output with radar data and to evaluate which scales are well represented by the model.

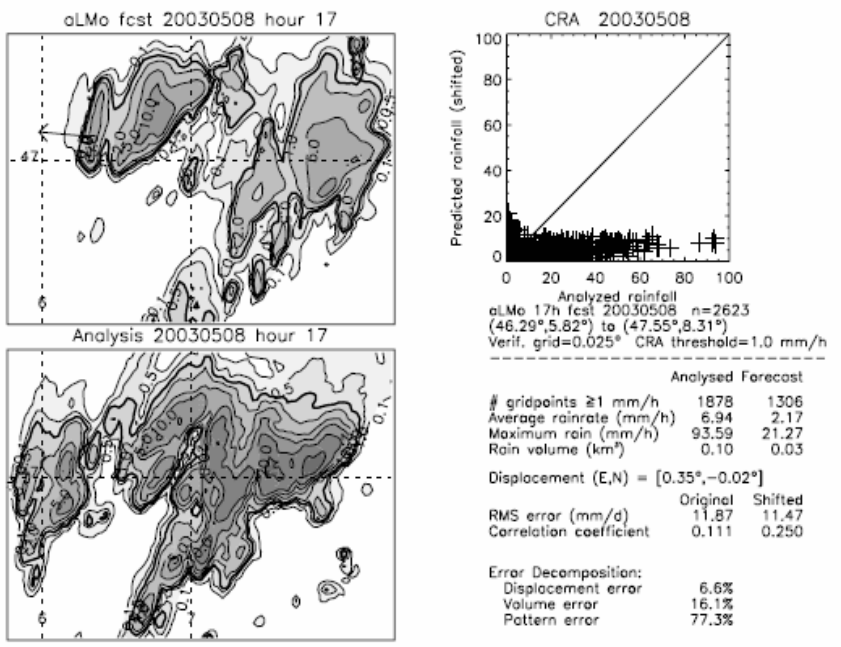


Fig.10. CRA verification of the COSMO forecast shown in Fig. 5. The red arrow in the upper left panel indicates that the best-fit of the forecast to the observations is made by translating the forecast approximately 30 km to the west

16.4 Object and entity-based techniques

The tendency of a human analyst, when presented with a rainfall map, is to focus on features of interest such as areas of heavy rain. Object- or entity-based verification techniques imitate this intuitive approach by identifying and comparing rain features in the forecast and observed fields, often using a pattern recognition methodology. By focusing on

the properties of larger objects, the fine-scale errors take on lesser importance. Like most scale-dependent verification techniques, the object-based techniques require observations to be on the same grid as the forecast.

One of the early object-based approaches was the contiguous rain area (CRA) technique of Ebert and McBride (2000), in which a rain threshold is applied to identify overlapping or nearby entities in the forecast and observed fields. The entities are matched by spatially translating the forecast field over the observed field until a best-fit criterion is met and the properties of the matched entities are then compared. The total error can be decomposed into contributions from location, volume and pattern error.

Tartaglione et al. (Chap. 17 in this book) apply this methodology to precipitation verification over Cyprus. To see how the CRA verification compares to the neighborhood and scale decomposition approaches, Fig. 10 shows results for the COSMO forecast. According to the error decomposition, the majority of the error was due to differences in fine scale pattern, with only about 7% being due to incorrect location of the rain area.

A more sophisticated pattern recognition algorithm has recently been developed by Davis et al. (2006). Now called Method for Object-based Diagnostic Evaluation (MODE), it uses a convolution threshold approach to first identify objects in forecast and observed fields. The properties of the objects (e.g., location, area, shape, orientation, texture, etc) are then input to a fuzzy logic algorithm that both merges nearby objects in a scene and matches them between the forecast and observations. Verification consists of quantifying the differences in the properties of matched forecast and observed objects. The user can assign different weights to these properties in the fuzzy merging/ matching algorithm in order to emphasize certain important aspects of the forecast, for example, rain location or maximum intensity.

An image processing approach that has recently applied to spatial verification is morphing. Instead of trying to directly match objects in the forecast and observed fields, morphing distorts the forecast field until it optimally matches the observations. The 2D fields of distortion vectors and bias of the phase-corrected forecast give information about the forecast error. Application of morphing to precipitation verification is made difficult by the fact that rain features may exist in the forecast but not in the observations and visa versa. Recently, Keil and Craig (2007) proposed a forecast quality measure (FQM) that combines information about the displacement and amplitude errors. The distortion vectors are computed using a pyramidal matching algorithm where

possible and, where no match can be found, an amplitude error is computed as the squared difference between the two fields. The FQM is the sum of the two normalized errors and reflects their subjective evaluation of forecast quality.

Cluster analysis also derives from the science of image processing and is a natural approach for associating pixels into objects in a high resolution rainfall grid, yet this strategy has only recently been used for verifying forecasts. In the verification method of Marzban and Sandgathe (2006, 2007) the forecast and observations are combined into a single field. K-means clustering is used to group pixels into k clusters based on their location and intensity and these clusters are further iteratively grouped using hierarchical agglomerative clustering. As the number of clusters is varied from k to 1, essentially increasing the spatial scale, the relative population of forecast and observed pixels in each cluster determines whether it is classified as a hit (forecast pixels between 20% and 80% of total), miss, or false alarm. These enable the calculation of categorical verification scores such as the threat score.

An object-based verification approach that assesses the structure of forecast rainfall in a pre-defined region such as a river basin is the Structure Amplitude Location (SAL) method of Wernli et al. (2008). As implied by the name, this approach compares the area mean structure, amplitude and location of threshold-defined precipitation objects in the forecast and observed fields, but does not attempt to match them. This approach is quite intuitive and computationally simple. Instead of a single number this method provides three: S, A and L (normalized structure, amplitude and location errors). An example of a SAL verification is given in Sect. 16.5.2.

16.5 Stratification

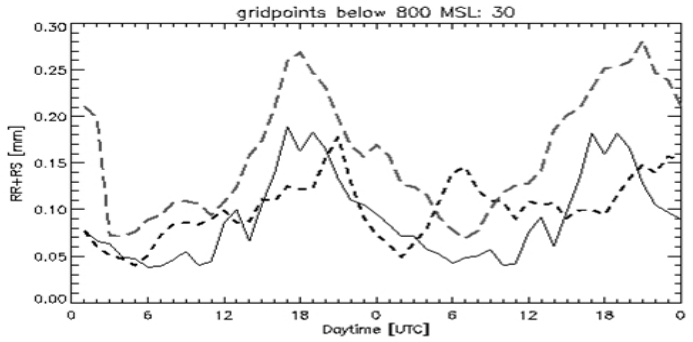
It is arguable whether QPF quality can be synthesized into one single number since one might be interested both in the forecast system's ability to predict the occurrence of rain, as well as how skillful it is in forecasting heavy rain. One might suspect that the performance in winter and in summer could be different, or that, for instance, model performance in anticyclonic conditions may differ from that in a vigorous northerly flow. These differences again may depend on the geographical location, especially with respect to the presence of a land-sea border or mountains. This kind of differentiated evaluation is achieved by appropriately stratifying the verification data set. If

stratification into relatively homogeneous subsamples is not performed then the verification results may be artificially high (Hamill and Juras 2007). For example, a model may appear to perform well when it is, in fact, only differentiating winter and summer regimes. This is hardly useful. Various examples are presented for which differentiation with respect to event intensity, seasons, time of the day, geographical regions and weather types were applied, in order to illustrate the potential of stratification to unmask systematic model errors.

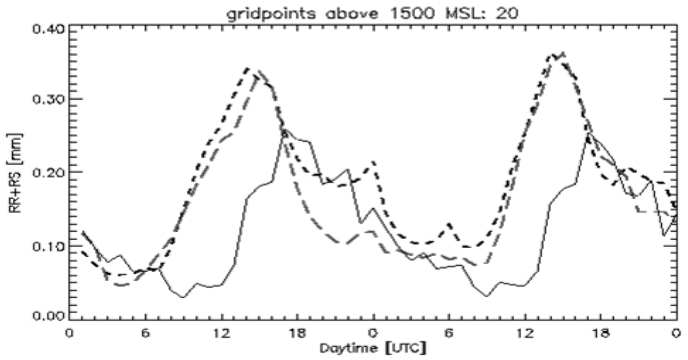
16.5.1 Seasonal, geographical and temporal stratification

Schubiger et al. (2006) present highlights of the comprehensive verification suite of COSMO, the operational NWP model of MeteoSwiss. Differentiation of rain intensity shows that occurrence of rain, or light rain, is generally overestimated, while COSMO tends to overestimate heavy rain over the mountains and underestimate it over the flatter Swiss Plateau. Averaging the diurnal cycle over a period of time effectively shows the forecast bias as a function of the hour of the day. Figure 11 shows such an averaged diurnal cycle for the months June and July 2006. In addition to singling out the hours of the day, model precipitation is verified separately for mountain stations (station height > 1500masl) and stations located over the Swiss Plateau (station height < 800masl). Given that a good part of the convective activity in the warm season consists of thermal convection in the mountains, this verification nicely isolates and reveals the problem that convection is triggered too early in the model, a misbehavior that was somewhat mitigated but not eliminated with a modified parameterization scheme for deep convection. More information on COSMO shortcomings were found looking at mountain and lowland stations separately (Schubiger et al. 2006).

Ebert et al. (2007) evaluate near-real-time satellite-derived QPE and NWP QPF on a global scale. They find that their performances are highly dependent on the rainfall regime and essentially opposed to each other, i.e., that satellite-derived QPE performs best in summer and at lower latitudes, whereas NWP has greatest skill in winter and at higher latitudes. Again, Ebert et al. (2003) report global NWP model QPF ETS values in the range of 0.4-0.5 in winter when synoptic weather is prevailing, while ETS values drop to 0.3 in summer when convective weather is predominant.



a)



b)

Fig. 11. Verification of the average diurnal cycle of precipitation (in mm) of COSMO forecasts for June and July 2006 for Swiss Plateau (< 800masl, panel a) and Mountain stations (>1500masl, panel b). Continuous lines denote observations, short-dashed lines operational and long-dashed lines in grey a modified parameterization scheme for cumulus convection which is able to somewhat mitigate the early onset of convective precipitation in the mountains (courtesy F. Schubiger and S. Dierer, MeteoSwiss)

16.5.2 Weather-type dependent stratification

Monthly, seasonal and annual statistical verifications are limited in that their performance is judged over the whole spectrum of weather types the atmosphere can produce. The danger is that they can mask differences in forecast quality when the data, even in terms of flow regimes, are not homogeneous. Further, they can

bias the results toward the most commonly sampled regime (for example days with no severe weather). A weather situation-dependent classification is another means by which stratification can be constructed. Rossa et al. (2003, 2004) have used the Schuepp Wetterlageneinteilung (Wanner et al. 1998) to perform a stratified COSMO QPF verification against QPE derived from the Swiss radar network (SRN) for years 2001 and 2002.

Zala and Leuenberger (2007) updated it for 2006 using a 'home made' classification into 11 classes comprising low flow configurations (cyclonic, anticyclonic and flat pressure distributions) and stronger flow configurations subdivided into the eight main wind directions.

Looking at the overall, unstratified, data set (Fig. 12) one is led to think that the COSMO 24h QPF accumulations (forecast range +6 to +30h) are quite decent. Bias values are smaller than 1 millimeter per day for large parts of the domain covered by the SRN, whereas the wet bias stays moderate even on some mountain peaks. There is a slight dry bias on the northern Swiss Plateau. However, looking at the various weather classes one can appreciate very significant differences in QPF quality in terms of precipitation bias. The most notable systematic behavior arises from the model's difficulty to partition orographic precipitation adequately between the up- and downwind side in that the upwind side generally receives too much and the lee side too little precipitation (Fig. 13a). This is especially true for the model version with instant fall-out of rain once this latter is formed (diagnostic precipitation scheme). This problem is somewhat mitigated, but not eliminated, with the introduction of the so-called prognostic precipitation scheme, which is capable of transporting formed raindrops with the wind. The most dramatic model error appears to occur in situations of southwesterly flow (Fig. 13b), when the model exhibits a widespread and quite marked dry bias over the Swiss Plateau and a portion of the northern foothills of the Alps, while retaining overestimation on the upwind side of part of the orography. In situations with northerly flow, including northwest and northeast, overestimation is substantial, while the dry bias over the Swiss Plateau is still there.

Jenkner et al. (2008a) construct a stratification based on the dynamic identification of distinct flow regimes. This is done by identifying upper-level streamers of potential vorticity (PV) and classifying them with respect to the orientation of their axis. As an example Fig. 14 shows the quantile-based Peirce skill score (PSS, 80% quantile, Jenkner et al. 2008b) of cases in which southwest-northeastward tilted PV streamers propagate past the Alps for two longitudinal ranges. Days with streamers upstream of the Alps (class

LC1_2, streamer axis between 10°W and 0°) are separated from days with streamers downstream of the Alps (class LC1_4, streamer axis between 10°E and 20°E). The former induce a southwesterly flow over Switzerland whereas the latter cause a northerly flow. The PSS identifies that the pixel-by-pixel matching of the COSMO re-forecast for LC1_2 (78 days) is low over the Swiss Plateau, while it is significantly higher for the class LC1_4 (51 days). The SAL verification (Fig. 15) adds considerable information revealing COSMO's tendency to underforecast precipitation in cases of approaching troughs, both in quantity and areal extension. After the trough axis has passed the Alps the model overforecasts precipitation with somewhat lesser error in terms of its structure.

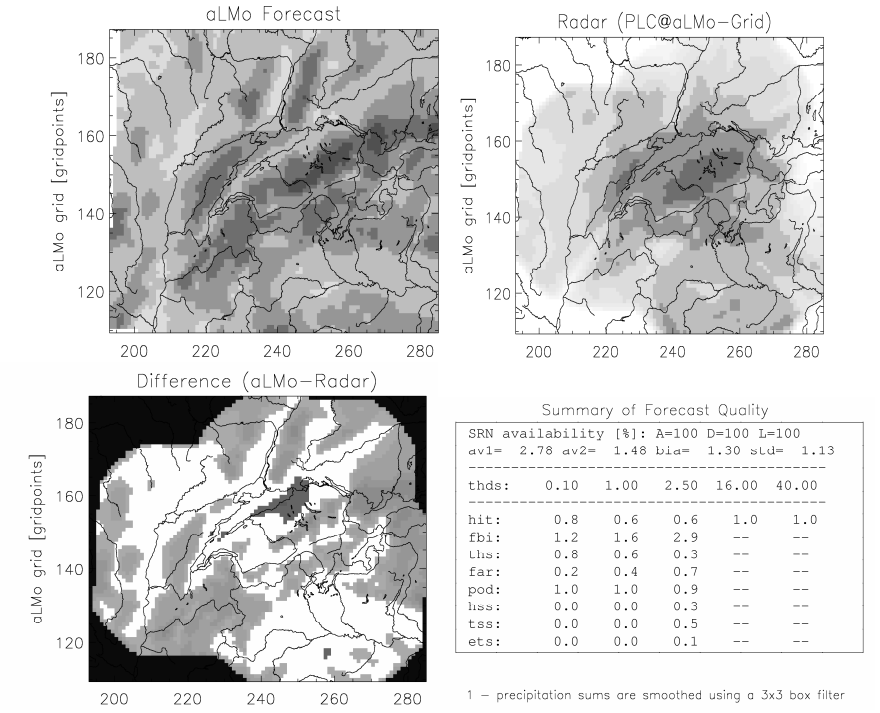


Fig. 12. Verification of operational COSMO precipitation forecasts (upper left panel, forecast range +06h - +30h) against the Swiss Radar Network (upper right panel) for the climatic year 2005. Shading denotes average daily precipitation in mm/24h in a log scale (0.1, 0.16, 0.25, 0.4, 0.63, 1.0, 1.6, 2.5, 4.0, 6.3, 10, 16, 25 etc, darkest shading 10-16mm/24h). Lower left panel denotes the average daily bias COSMO daily averaged QPF (white areas are within -1 and 1mm/24h, darker areas denote a dry bias, lighter areas a wet bias, steps from 1mm/24h up as in other plots), while the statistical scores are evaluated on all grid points

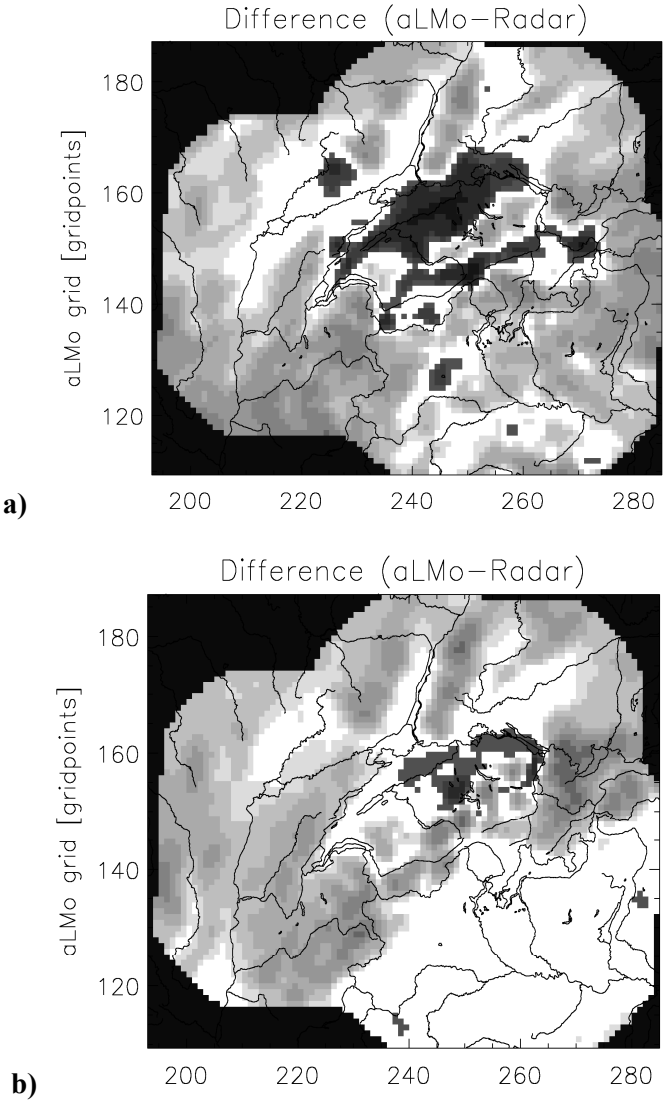


Fig. 13. As the bias field in Fig. 12, but weather classes 'southwest' (panel a, 50 days) and 'northwest' (panel b, 20 days)

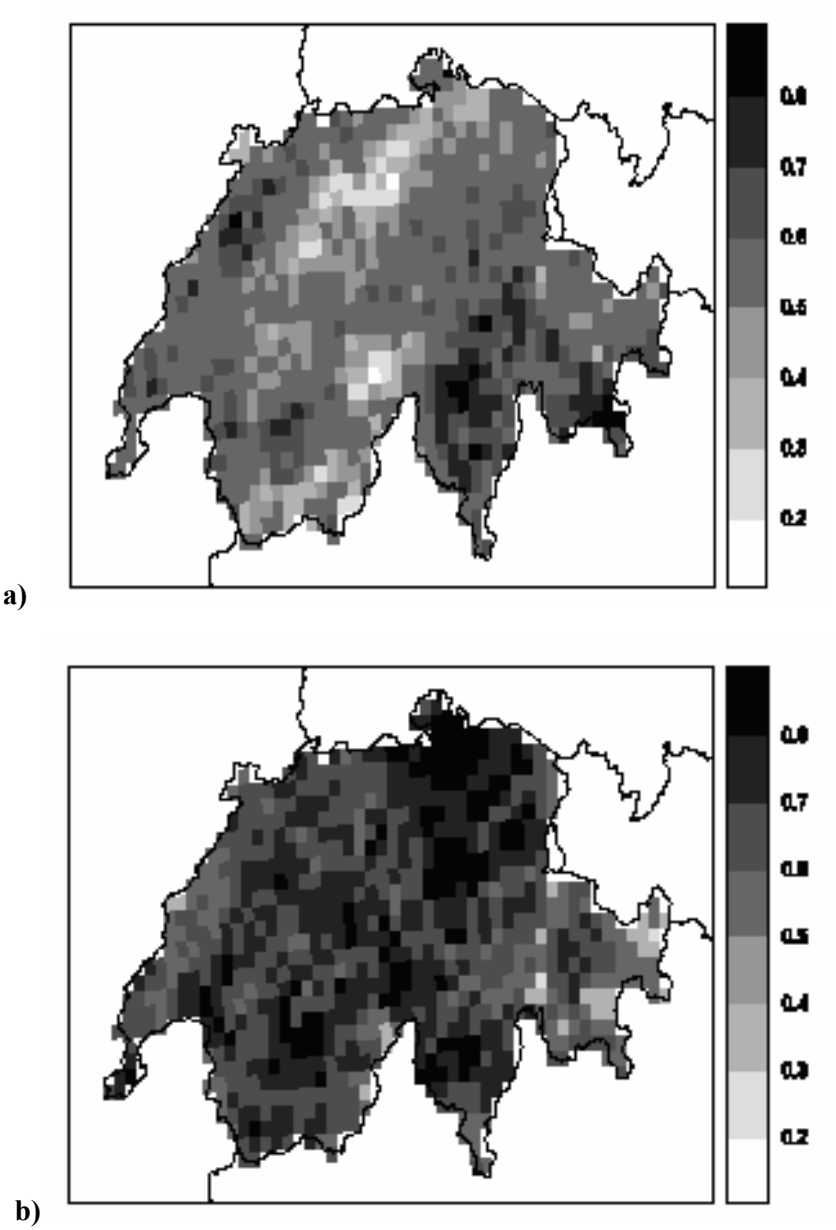
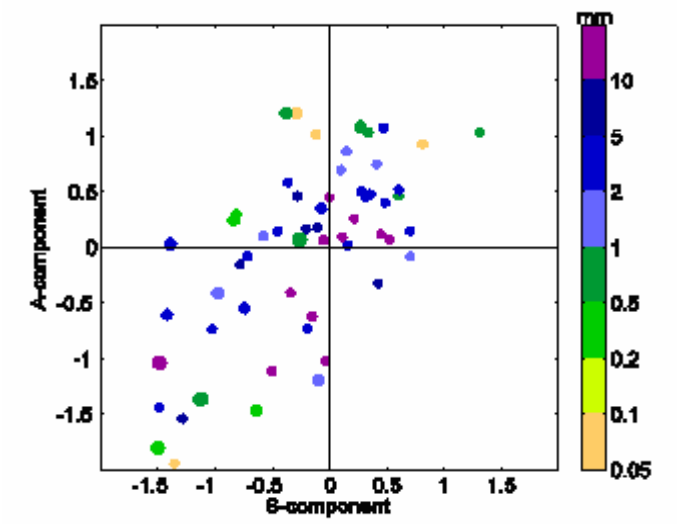
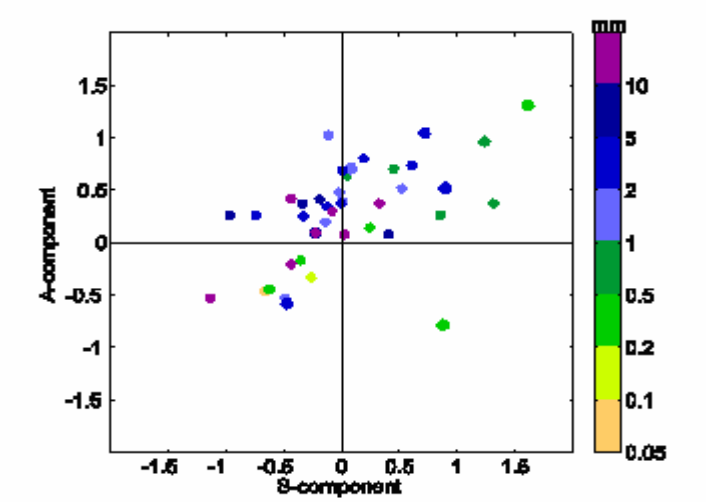


Fig. 14. Quantile-based Peirce Skill Score (PSS) for the 80% quantile in dependence of the flow classes LC1_2 (panel a, 78 days) and LC1_4 (panel b, 51 days, see text for explanation). The PSS measures how well the COSMO QPF match the observations for every individual pixel (courtesy J. Jenkner, ETH Zurich)



a)



b)

Fig. 15. SAL verification (Wernli et al. 2008) for the two flow regimes displayed in Fig. 14 for the area of Switzerland. The horizontal axis denotes how well the model matches the structure of the precipitation areas, the vertical axis how well it matches the rainfall amount, while the size of the dots denote average positioning errors. The grey scales denote the daily precipitation accumulations for the days attributed to this flow regime (values are 0.05, 0.1, 0.2, 0.5, 1, 2, 5 and 10 mm, courtesy J. Jenkner, ETH Zurich)

16.6 Which verification approach should I use?

The verification approach that one should choose will depend in part upon the observations available for verifying the forecasts. As shown in Sect. 16.4.2 and discussed in greater detail by Tartaglione et al. (Chap. 17 in this book), the nature and accuracy of the 'truth' data have a profound effect on the verification results. If one has only rain gauge data available then the choices for verification approaches are limited to the traditional metrics and a few of the neighborhood techniques. We advocate using merged gauge-radar QPE where possible to take advantage of the additional spatial information available from these analyses and help to 'prove' the improvements in the new higher resolution models.

The standard continuous and categorical verification statistics computed from point match-ups are well understood and have been used for many years. In most cases it is advisable to continue computing such statistics, especially if a long time series of verification results is available and one wants to compare the accuracy of a new forecast system to that of an older system. However, if the resolution of the new forecast has been increased then the double penalty problem may lead to poorer verification results, even if one intuitively feels that the forecast is better. A more diagnostic evaluation using spatial verification methods may be desirable, especially if verifying data are available on a grid, say from radar or gauge analyses.

The neighborhood verification approach is useful when the forecasts are made at high resolution and it is unreasonable to expect a good match with the observations at the finest resolution. For verifying model forecasts at scales the model may be expected to resolve, the methods that compare against neighborhoods of observations may be more useful. If the aim is to evaluate the accuracy of the forecast for any given point of interest, it is better to use methods that compare forecasts to the observation in the center of the neighborhood. Among the neighborhood methods described in Sect. 16.4.1, the upscaling method is appropriate for users who wish to know if the rain amount is correct, for example, hydrologists using NWP forecasts to predict catchment rainfall and model developers evaluating the water balance of a model. The multi-event contingency table method is especially good for evaluating high resolution model output that may lead to advice and warnings for specific locations. NWP model developers and users can use the fractions skill score to determine at which scales the model has useful skill.

Scale decomposition methods separate the errors by scale, unlike neighborhood methods that filter out smaller scales. The scale decomposition methods are good for investigating the source of forecast errors when they are caused by processes occurring on different scales (for example, cloud-scale processes or large-scale advective processes). When the goal is to know whether a model's precipitation field resembles observed rainfall in a structural sense, then computing the multi-scale statistical properties is a sensible way to proceed. The SAL method is philosophically similar, but applied to objects rather than pixels.

Object-based verification approaches represent rain features as objects and are, therefore, quite intuitive. Many of these techniques give practical information about forecast quality such as location and amplitude errors. These methods tend to be more complex than other methods and also involve the choice of one or more parameters (the threshold used to define objects, for example) to which the method may be quite sensitive. Object-based approaches work well for well-defined rain areas appearing in both the forecast and observations (e.g., meso-scale convective systems, frontal systems and daily rainfall accumulations) but they do not handle noisy rain fields very well.

Independently of the chosen approach, appropriate stratification of verification data sets can help to isolate specific problems in the QPF systems. Hydrologists concerned with river catchments in mountainous terrain, for example, may well be interested in knowing in what regions of the forecast domain a QPF exhibits systematic errors.

16.7 References

Atger F (2001) Verification of intense precipitation forecasts from single models and ensemble prediction systems. Nonlinear Proc Geoph 8:401-417

Barnes LR, Gruntfest EC, Hayden MH, Schultz DM, Benight C (2007) False alarms and close calls: A conceptual model of warning accuracy. Weather Forecast (accepted)

Bougeault P (2002) WGNE survey of verification methods for numerical prediction of weather elements and severe weather events. CAS/JSC WGNE Report No. 18, Appendix C. http://www.wmo.ch/web/wcrp/documents/wgne18rpt.pdf

Briggs WM, Levine RA (1997) Wavelets and field forecast verification. Mon Weather Rev 125:1329-1341

Casati B, Ross G, Stephenson DB (2004) A new intensity-scale approach for the verification of spatial precipitation forecasts. Meteorol Appl 11:141-154

Cherubini T, Ghelli A, Lalaurette F (2002) Verification of precipitation forecasts over the Alpine region using a high-density observing network. Weather Forecast 17:238-249

Davis C, Brown B, Bullock R (2006) Object-based verification of precipitation forecasts. Part I: Methods and application to mesoscale rain areas. Mon Weather Rev 134:1772-1784

Ebert EE (2008) Fuzzy verification of high resolution gridded forecasts: A review and proposed framework. Meteorol Appl (accepted)

Ebert EE, McBride JL (2000) Verification of precipitation in weather systems: Determination of systematic errors. J Hydrol 239:179-202

Ebert EE, Damrath U, Wergen W, Baldwin ME (2003) The WGNE assessment of short-term quantitative precipitation forecasts. B Am Meteorol Soc 84: 481-492

Ebert EE, Janowiak JE, Kidd C (2007) Comparison of Near-Real-Time Precipitation Estimates from Satellite Observations and Numerical Models. B Am Meteorol Soc 88:47–64

de Elia R, Laprise R, Denis B (2002) Forecasting skill limits of nested, limited-area models: A perfect-model approach. Mon Weather Rev 130:2006-2023

Finley JP (1884) Tornado predictions. American Meteorological Journal 1:85-88

Germann U, Berenguer M, Sempere-Torres D, Salvadè G (2006) Ensemble radar precipitation estimation - a new topic on the radar horizon. In: Proceedings 4th European Conference on Radar in Meteorology and Hydrology. 18-22 September 2006, Barcelona, Spain. http://www.erad2006.org

Göber M, Wilson CA, Milton SF, Stephenson DB (2004) Fairplay in the verification of operational quantitative precipitation forecasts. J Hydrol 288:225-236

Hamill TM, Juras J (2007) Measuring Forecast Forecast Skill: Is it Real Skill or is it the Varying Climatology?. Q J Roy Meteor Soc 132 (in press)

Harris D, Foufoula-Georgiou E, Droegemeier KK, Levit JJ (2001) Multiscale statistical properties of a high-resolution precipitation forecast. J Hydrometeorol 2:406-418

Hense A, Adrian G, Kottmeier Ch, Simmer C, Wulfmeyer V (2003) Priority Program of the German Research Foundation: Quantitative Precipitation Forecast. Research proposal available at http://www.meteo.uni-bonn.de/projekte/SPPMeteo/reports/SPPLeitAntrag_English.pdf

ICP (2007) Spatial Forecast Verification Methods Intercomparison Project (ICP). http://www.ral.ucar.edu/projects/icp/index.html

Jenkner J, Dierer S, Schwierz C (2008a) Conditional QPF verification using synoptic weather patterns - a 3-year hindcast climatology (in preparation)

Jenkner J, Frei C, Schwierz C (2008b) Quantile–based short–range QPF evaluation over Switzerland (to be submitted)

Jolliffe IT, Stephenson DB (2003) Forecast verification. A practitioner's guide in atmospheric science. Wiley and Sons Ltd, pp 240

Keil C, Craig GC (2007) A displacement-based error measure applied in a regional ensemble forecasting system. Mon Weather Rev (in press)

Leuenberger D (2005) High-resolution radar rainfall assimilation: exploratory studies with latent heat nudging. Ph.D. thesis, ETH Zurich, Switzerland, Nr. 15884. http://e-collection.ethbib.ethz.ch/cgi-bin/show.pl?type=diss&nr=15884

Marzban C, Sandgathe S (2006) Cluster analysis for verification of precipitation fields. Weather Forecast 21:824-838

Marzban C, Sandgathe S (2007) Cluster analysis for object-oriented verification of fields: A variation. Mon Weather Rev (in press)

Mittermaier MP (2007) Improving short-range high-resolution model precipitation forecast skill using time-lagged ensembles. Q J Roy Meteor Soc 133:1-19

Murphy AH (1996) The Finley affair: A signal event in the history of forecast verification. Weather Forecast 11:3-20

Nurmi P (2003) Recommendations on the verification of local weather forecasts. ECMWF Tech. Memo 430:18.t http://www.ecmwf.int/publications/library/ecpublications/_pdf/tm430.pdf

Oberto E, Turco M, Bertolotto P (2006) Latest results in the precipitation verification over Northern Italy. COSMO Newsletter 6:180-184

Richardson DS (2000) Skill and relative economic value of the ECMWF ensemble prediction system. Q J Roy Meteor Soc 126:649-667

Roberts NM, Lean HW (2007) Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. Mon Weather Rev (in press)

Rossa A, Arpagaus M, Zala E (2003) Weather situation-dependent stratification of precipitation and upper-air verification of the Alpine Model (aLMo). COSMO Newsletter 3:123-138

Rossa AM, Arpagaus M, Zala E (2004) Weather situation-dependent stratification of radar-based precipitation verification of the Alpine Model (aLMo). ERAD Publication Series 2:502-508

Rossa et al. (2005) The COST 731 Action MoU: Propagation of uncertainty in advanced meteo-hydrological forecast systems. http://www.cost.esf.org

Schaake JC, Hamill TM, Buizza R, Clarke M (2007) HEPEX, the Hydrological Ensemble Prediction Experiment. B Am Meteorol Soc (accepted)

Schubiger F, Kaufmann P, Walser A, Zala E (2006) Verification of the COSMO model in the year 2006, WG5 contribution of Switzerland. COSMO Newsletter 6:9

Stamus PA, Carr FH, Baumhefner DP (1992) Application of a scale-separation verification technique to regional forecast models. Mon Weather Rev 120:149-163

Stanski HR, Wilson LJ, Burrows WR (1989) Survey of common verification methods in meteorology. World Weather Watch Tech. Rept. No.8,

WMO/TD No.358, World Meteorological Organization, Geneva, Switzerland, pp 114

Stephenson DB (2000) Use of the 'odds ratio' for diagnosing forecast skill. Weather Forecast 15: 221-232

Stephenson DB, Casati B, Wilson CA (2007) The extreme dependency score: A new non-vanishing verification measure for the assessment of deterministic forecasts of rare binary events. Meteorol Appl (submitted)

Steppeler J, Doms G, Schättler U, Bitzer H-W, Gassmann A, Damrath U, Gregoric G (2003) Meso-gamma scale forecasts using the non-hydrostatic model LM. Meteorol Atmos Phys 82:75–96

Wanner H, Salvisberg E, Rickli R, Schuepp M (1998) 50 years of Alpine Weather Statistics (AWS), Meteorol Z N.F. 7:99-111

Wernli H, Paulat M, Hagen M, Frei C (2008) SAL - a novel quality measure for the verification of quantitative precipitation forecasts, Mon Weather Rev (submitted)

Wilks DS (2006) Statistical methods in the atmospheric sciences. an introduction. 2nd edn. Academic Press, San Diego, pp 627

Wilson C (2001) Review of current methods and tools for verification of numerical forecasts of precipitation. COST717 Working Group Report on Approaches to verification. http://www.smhi.se/cost717/

WWRP Joint Working Group on Verification website (2004) Forecast verification - Issues, methods and FAQ. http://www.bom.gov.au/bmrc/wefor/staff/eee/verif/verif_web_page.html

Yates E, Anquetin S, Ducrocq V, Creutin J-D, Ricard D, Chancibault K (2006) Point and areal validation of forecast precipitation fields. Meteorol Appl 13:1-20

Zala E, Leuenberger D (2007) Update on weather-situation dependent COSMO-7 verification against radar data. COSMO Newsletter 7:1-4

Zawadzki I (1973) Statistical properties of precipitation patterns. J Appl Meteorol 12:459-472

Zepeda-Arce J, Foufoula-Georgiou E, Droegemeier KK (2000) Space-time rainfall organization and its role in validating quantitative precipitation forecasts. J Geophys Res 105:10129-10146