Quantile–based short–range QPF evaluation over Switzerland

J Jenkner[*]

Institute for Atmospheric and Climate Science & ETH Zurich

Zurich, Switzerland

C Frei

Federal Office of Meteorology and Climatology MeteoSwiss

Zurich, Switzerland

C Schwierz

Institute for Atmospheric Science & University of Leeds

Leeds, UK

June 30, 2008

[*] Johannes.Jenkner@env.ethz.ch

Abstract

Quantitative precipitation forecasts (QPF) are often verified using categorical statistics. The traditionally used 2×2 contingency table is modified here by applying sample quantiles instead of fixed amplitude thresholds. This calibration is based on the underlying precipitation distribution and has beneficial implications for categorical statistics. The quantile difference and the debiased Peirce skill score split the total error into the independent components of bias and pixel overlap. It is shown that they provide a complete verification set with the ability to assess the full range of rainfall intensities. The technique enables skill to be estimated without spurious influences from the marginal totals and the problem of hedging is eluded. To exemplify the practical potential of the quantile–based contingencies, the method is applied to 6.5 years of operational rainfall forecasts from the Swiss Federal Office of Meteorology and Climatology (MeteoSwiss). Daily accumulations of the COSMO model at 7 km grid resolution are compared to a high–quality gridded observational record of spatially interpolated rain gauge data. The quantile–based scores are applied to single grid points and to predefined regions. A high–resolution error climatology is built up and reviewed in terms of potential error sources in the model. The seasonal QPF performance exhibits the most severe overestimation over the Northern Alps during winter, indicative of the impact of the model ice phase. The QPF performance related to model updates, such as the introduction of the prognostic precipitation scheme is also evaluated. It is demonstrated that the overlap of rain pixels continuously increases for subsequent versions of the COSMO model. All over the period, a strong geographical gradient of the rainfall matching is evident with a much higher Peirce skill score on the Alpine south side.

Zusammenfassung

Zur Verifikation von Quantitativen Niederschlagsvorhersagen (QNV) werden häufig kategorische Fehlermasse verwendet. Die traditionellerweise benutzte 2×2 Kontingenztafel wird hier durch die Anwendung von Quantilen anstelle von festen Amplitudenschwellwerten modifiziert. Diese Kalibrierung orientiert sich an der zugrundeliegenden Niederschlagsverteilung und beeinflusst die kategorischen Fehlermasse vorteilhaft. Die Quantilsdifferenz und der angepasste Peirce Skill Score teilen den Gesamtfehler in die unabhängigen Komponenten des Bias und der Gitterpunktsüberlappung auf. Es wird gezeigt, dass sie eine vollständige Verifikationsbasis bilden, die den gesamten Bereich an Niederschlagsintensitäten abdecken kann. Die Methodik erlaubt es, den Skill des Modells ohne störende Einflüsse der Randverteilungen zu bestimmen und umgeht die Problematik des "Hedging". Um die praktische Einsetzbarkeit von quantilsbasierten Fehlermassen aufzuzeigen, wird die Methode auf 6.5 Jahre an Niederschlagsvorhersagen vom Schweizer Bundesamt für Meteorologie und Klimatologie (MeteoSchweiz) angewendet. Tägliche Summen aus dem COSMO–Modell mit einer horizontalen Auflösung von 7 km werden mit einem hochwertigen gegitterten Beobachtungsdatensatz verglichen. Die quantilsbasierten Fehlermasse werden auf einzelne Gitterpunkte und auf festgelegte Gebiete angewendet. Eine hochaufgelöste Fehlerklimatologie wird erstellt und im Hinblick auf mögliche Fehlerquellen des Modells untersucht. Die jahreszeitlich gemittelten QNV–Fehler zeigen die grösste Überschätzung über den nördlichen Alpen im Winter, was auf den Einfluss des Eisschemas hinweist. Die QNV–Fehler während verschiedener Phasen der operationellen COSMO–Modellentwicklung, wie z.B. der Einführung des prognostischen Niederschlagsschemas, werden ebenfalls quantifiziert. Dabei wird gezeigt, dass sich die Gitterpunktsüberlappung der Niederschlagsgebiete kontinuierlich verbessert hat. Über die ganze Periode fällt ein grosser geographischer Fehlergradient mit einem viel höheren Peirce Skill

Score auf der Alpensüdseite auf.

1 Introduction

Precipitation forecasts are of societal, economic, and social interest and decision ma-king often relies on accurate rainfall predictions. Hence, research activity currently is large to improve Quantitative Precipitation Forecasting (QPF) and weather centers continuously evaluate their operational high–resolution limited–area models (LAM) to trace error sources.

QPF is particularly challenging over complex terrain. The Mesoscale Alpine Programme (MAP, Benoit et al., 2002) provided many new insights into the dynamics and challenges of predicting orographic precipitation (e.g. Richard et al., 2007, Rotunno and Houze, 2007, and references therein). A variety of factors determine the formation of heavy precipitation and influence QPF quality. They include: (a) accuracy of the synoptic–scale upper–level triggers of precipitation and their correct mesoscale interaction with the underlying orography (e.g. Fehlmann and Quadri, 2000, Martius et al., 2006), (b) representation of the orography (Accadia et al., 2005) and model resolution (Buzzi et al., 2004, Zängl, 2007), (c) the low-level moisture field (Martius et al., 2006, Mahoney and Lackmann, 2007) and the boundary layer structure (Rotunno and Houze, 2007), (d) enhancement of precipitation by microphysical processes (e.g. Zeng et al., 2001, Pujol et al., 2005) and turbulence (Houze and Medina, 2005).

Also the formulation and numerics of the Numerical Weather Prediction (NWP) system itself can influence QPF. Kaufmann et al. (2003) evaluated the former hydrostatic LAM of the COSMO consortium[1] and found that the model error is exceedingly sensitive to the activation of convection in parameterized precipitation and the sloping topography in resolved precipitation. Kållberg and Montani (2006) compared a non–hydrostatic versus a hydrostatic model which also differed in the data-assimilation schemes and numerics. Above all, they

[1] web site: www.cosmo-model.org

found more intense precipitation extremes in the non-hydrostatic formulation due to dissimilarities of the convection schemes. The prognostic treatment of precipitation markedly improved the wet (dry) bias on the windward (downwind) side of orography which is typically noted for QPF over complex terrain (e.g. Elementi et al., 2005).

From a more generic point of view, Hohenegger and Schär (2007) investigated the dynamics of error growth. They used a range of different initial perturbation procedures of a high-resolution ensemble and found a rapid radiation of the initial uncertainties throughout the computational domain and a further amplification over moist convectively unstable regions (compare also Hohenegger et al., 2006).

As is clear from the above, the problem of identifying sources of QPF errors is highly complex. An additional drawback arises from the fact that most of our knowledge today is based on case studies or on relatively short–term periods (up to 1–2 years) of investigation. However, QPF quality tends to be more case–dependent than model–dependent (Richard et al., 2003). A consistent quantification and rigorous investigation of QPF over a longer–term period is highly desirable to trace QPF errors and identify the main model shortcomings. In view of this, novel verification techniques for precipitation forecasts are currently being developed. Most notably, spatial approaches are able to consider different areal error aspects. An example is the intensity–scale technique (Casati et al., 2004, Mittermaier, 2006) which diagnoses skill as a function of "precipitation rate intensity" (Casati et al., 2004) and spatial scale of the error. Another example is the object–based quality measure SAL (Wernli et al., 2008) which assesses the coherence in structure, amplitude and location of precipitation objects. As a matter of fact, the findings of these and comparable methods are related to predefined areas, which leads to a smoothing of precipitation and local error features. Hence the attribution of individual verification aspects to specific locations or sites is hardly possible. Pertinent information is therefore lost over mountainous terrain where the spatial variability of atmospheric parameters usually is large (e.g. Frei and Schär, 1998, Schmidli et al., 2002).

In this context, the traditional grid–point verification still provides an expedient alternative. As explained by Murphy and Winkler (1987) or McBride and Ebert (1999), categorical statistics are traditionally used for dichotomous verification purposes such as the validation of precipitation. After the selection of a threshold value it is tested whether model and observations exceed the limit. In this sense, all standard methods rest upon a conventional 2x2 contingency table (Tab. 1), consisting of the number of hits H, misses M, false alarms F and correct negatives Z. From these four entries several error measures (or scores) such as frequency bias, probability of detection (POD), false alarm ratio, threat score, equitable threat score, Peirce skill score (PSS, Peirce, 1884), odds ratio skill score and others can be derived.

A desired property of categorical measures is equitability as defined by Gandin and Murphy (1992). The definition implies that random and constant forecasts possess unvarying expectation values irrespective of varying marginal totals. A necessary prerequisite for equitability is the ability to single out the joint distribution, defined by forecasts and observations together, and to reject the marginal distributions, defined separately by forecasts and observations. In this sense, only equitable scores ensure a fair comparison of samples with different characteristics. A simple example for a non–equitable score is the POD which exhibits higher values for random forecasts of frequent events than for those of rare events. Non–equitable measures can upgrade forecasts which are not consistent i.e. which do not correspond to the forecaster's judgment (see Murphy, 1993, for explanation). Consequently, both model developers and operational forecasters are tempted by hedging (Murphy and Epstein, 1967) and falsely adjust the number of forecasted events to achieve a better scoring.

So far, opinions about equitability substantially diverge in the literature. As proven by Gandin and Murphy (1992), the PSS, which is equivalent to the Hanssen–Kuipers discriminant (Hanssen and Kuipers, 1965), constitutes the unique solution (i.e. the only one possible) which satisfies the strict conditions for equitability, namely an invariable rating of both random and constant forecasts. In fact, the uniqueness of the PSS still applies to conditions with an unequal penalty for misses and false alarms (Manzato, 2005). However, the uniqueness is lost, if the conditions for equitability are slightly relaxed, i.e. constant forecasts are allowed to score unlike random forecasts (Marzban and Lakshmanan, 1999). In theory, scores can be equitable or nearly equitable (in the sense that random/constant forecasts do not vary or only slightly vary) and provide a verification free from disturbing mathematical antagonisms. In practice, the usefulness of measures essentially varies in different constructed and real cases (e.g. Marzban, 1998, Hamill and Juras, 2006). On the one hand, it is not feasible to compare forecasts for different event frequencies or base rates (e.g. Mason, 2003). The so–called base–rate error vitiates the value of finite–accuracy forecasts for rare events (Matthews, 1996). Even though most scores are deeply affected by this dilemma, Woodcock (1976) and Mason (1989) find the PSS to be unaffected. Indeed, Thornes and Stephenson (2001) only grant the odds ratio skill score and the PSS to cope with small base rates. On the other hand, common scores depend on the bias which Hilliker (2004) exemplifies by means of the threat score. The PSS exhibits the detriment of approaching the POD in rare–event situations. Since the POD correlates with the bias, the PSS does as well. Mesinger (2007) complains that the impact of the bias customarily is estimated in a subjective manner, because objective approaches are still missing. He introduces a method to debias the threat score and likewise the equitable threat score. To this end, he converts the standard contingency table to a setting with a unit bias and recommends it to use for an objective verification of the placing of precipitation systems. Although Mesinger's approach alleviates some of the basic problems of verifying biased distributions, a drawback remains, in that assumptions need to be made which are not derivable from the underlying distributions alone.

The purpose of our study is twofold. At first, we introduce a refined grid–point verification measure which fulfills the requirements of equitablility without making any additional assumptions. Thereby, we make use of the definition of a quantile which is equivalent to the terms of percentile and fractile (Wilks, 2006). Then we apply this measure to a long–term climatology of operational high–resolution precipitation forecasts over complex terrain. The resulting error climatology allows for an extensive model diagnosis to identify possible error sources. It is long enough to investigate subsets of the data base, such as seasonal error variations and chronological error developments caused by different operational model versions.

The structure of the paper is the following: Some background information about the observational analysis, the verified COSMO model and the geographic setting is provided in section 2. The refined verification methodology is derived and discussed in section 3. Then some observed precipitation characteristics are described in section 4, before we turn to present the verification results. Section 5 pinpoints seasonal error variations and section 6 highlights characteristics of different model versions. Finally results are discussed and synthesized in sections 7 and 8.

2 Verification data, model and domain

2.1 Observational analysis

The reference dataset used for model evaluation in this study is a gridded mesoscale analysis for Switzerland, which is derived from rain gauge observations by spatial interpolation. The underlying observation network encompasses typically 450 stations, corresponding to an average station distance of 10–15 km (Konzelmann and Weingartner, 2007). The Alpine in–situ observations are among the densest world–wide in high–altitude topography (Frei and Schär, 1998). Despite slight variations in the reading times across the network, the analyses can be considered as representing 24-hour totals from 06:00 until 06:00 UTC (Frei and Schär, 1998).

The spatial analysis of rain-gauge observations is conducted with a modified version of the SYMAP algorithm, a distance and direction weighting scheme by Shepard (1984) (see also Willmott et al., 1985). In deviation from the traditional scheme, our procedure encompasses an antecedent climatological scaling of station observations and subsequent re-scaling of the gridded anomalies with a high-resolution climatology (Schwarb et al., 2001). The procedure is similar to that of Widmann and Bretherton (2000) and is applied to reduce systematic errors due to biases in the distribution of stations with height. In our procedure, the SYMAP algorithm is applied with a different distance weighting scheme. The purpose is to represent, in the analysis, regional area mean values rather than point values (see Frei and Schär, 1998). The adopted analysis method is similar to that applied in Frei et al. (2006), except that the analysis is undertaken originally on a 2 km grid and is subsequently aggregated to the grid of the forecasting model.

It should be noted that the verification dataset is affected by systematic biases in the rain-gauge measurements. In Switzerland rain-gauge under-catch is expected to range from about 4% at low altitudes in summer to more than 40% above 1500 mMSL in winter (Sevruk, 1985) and has to be considered while rating the rainfall bias later on. Moreover, there may be systematic inconsistencies between the model and verification grids as a result of differences in effective resolution. From the observation network we expect an effective resolution of the verification dataset in the order of 15 km in the flatlands and 15–25 km in the mountains. This is probably close but slightly coarser than the model resolution when taking two nominal grid pixels (i.e. 2x7 km = 14 km) as the effective model resolution.

2.2 Model description

The error climatology derived in this study is based on the non–hydrostatic model of the COSMO consortium in operation at the Swiss Federal Office of Meteorology and Climatology (MeteoSwiss). The model is the Swiss counterpart of the former German Lokal Modell which has been described by Steppeler et al. (2003). It has been developed for the purpose of high–resolution weather forecasts with a preferable representation of the meso–γ scale. A detailed description currently can be looked up on `www.cosmo-model.org`. The horizontal mesh size of the model is 7km and the vertical resolution constitutes 45 height–based hybrid levels. The grid structure is based on an Arakawa C setup with Lorenz vertical grid staggering. The basic equations are solved in a fully elastic manner with a dry reference state at rest. Advection is treated by a split explicit scheme based on a filtered leapfrog time integration (Asselin, 1972, Skamarock and Klemp, 1992) with a main time step of 40s. Moist convection is treated by the mass flux convergence scheme of Tiedtke (1989). The warm–rain regime refers to a bulk water–continuity model proposed by Kessler (1969) whereas the ice–cloud regime is based on an extended saturation adjustment technique

(Lord et al., 1984).

The preoperational phase of the model at MeteoSwiss started in July 2000 and the model became operational in April 2001. The geographical domain comprises central Europe and receives boundary conditions from the GME model of the German Meteorological Service and later from the Integrated Forecast System (IFS) of the ECMWF. The updating of the boundaries is treated with an adjusted Davies relaxation scheme (Davies, 1976). In the present study daily precipitation accumulations of the period between July 2000 and December 2006 are analyzed over Switzerland. Investigations are confined to operational 00 UTC model runs, from which daily sums are constructed using lead times between 6 and 30 hours.

Since both preoperational and operational forecasts are evaluated, there are various model updates within the considered period. Altogether there are roughly 60 changes of the model setup including bug corrections. Most of them are not expected to have a significant impact on daily rainfall fields. However, few upgrades are considered to influence QPF quality decisively. We refer to the most important changes in section 6.

2.3 Geographic aspects

Altogether 859 grid points of the COSMO model are evaluated within the borders of Switzerland ($\sim$41000 km^2). To highlight regional disparities with respect to QPF quality, the whole area is subdivided into six orographically distinct parts (Fig.1) using contour lines of the model topography as well as other landmarks. The area of the Jura comprises the Swiss part of the Jura mountains as well as some adjacent pixels in the canton Jura. The low elevations in the Middleland reach from Lake Geneva in the southwest to Lake Constance in the northeast. The hilly relief only varies slightly in height here. The northern

Alpine crest as well as the approximate canton borders between the Valais, the Ticino and the Grisons split up the Alps into four mountainous domains. The Valais and the Grisons can be regarded as central Alpine domains, whereas the Northern Alps and the Ticino lie on opposing sides of the Alpine crest. Overall highest elevations (around 3000 mMSL in the model topography) are found in the middle of the northern Alpine crest, along the southern border of the Valais and in the south of the Grisons. They are marked with an X in Fig.1. Individual regions contain the following numbers of COSMO grid points: Jura 55, Middleland 271, Northern Alps 208, Valais 114, Ticino 71, Grisons 140.

3 Methodology of verification

The standard 2x2 contingency table (Tab. 1) for dichotomous events (relating to a predefined precipitation threshold) contains the following entries: hits H, misses M, false alarms F and correct negatives Z. The four numbers are used to compute several different categorical scores (see section 1 for examples). Varying the threshold from low (representing weak precipitation rates) to high values (representing strong precipitation rates) renders possible a comprehensive investigation of different intensities.

However, there are profound shortcomings, if standard categorical statistics are applied directly. Firstly, a predetermined amplitude threshold cuts the precipitation distributions at an unknown location, i.e. it is not obvious a priori whether the threshold value represents light or heavy rainfall within the considered sample. In an extreme case, single cells of the contingency table can become zero. Then some scores cannot be computed (due to a division by zero) and statements about model behavior are hard to make. Secondly, the distributions under comparison usually differ considerably with respect to their range of values. Customary scores do not fulfill the requirements of equitability (Gandin and Murphy, 1992) or fail to be firm with respect to hedging (Stephenson, 2000). Thirdly, the joint distribution always comprises three degrees of freedom, as the four entries are only linked to the sample size. Three scores are required to display all verification characteristics. Stephenson (2000) proposes the triplet of odds ratio skill score, PSS and frequency bias. According to his comments, it is possible to draw complementary information out of the considered datasets, if concurrent scores are applied simultaneously. But it remains ambiguous, how to attribute individual verification aspects to these measures which are not all independent from each other. Fourthly, we argue that it is not possible to integrate amplitude–based scores over a range of intensities. In principle, it is not meaningful to average scores for different thresholds, because it is not obvious how many data points fall within a certain range of thresholds.

In this study, a refined version of categorical statistics is proposed to address the above problems and avoid equivocal verification results. The contingency table is defined by means of frequency thresholds instead of amplitude thresholds. To this end, sample quantiles are computed for both data records independently. Thus, the two datasets are compared using the same relative cut–off (according to the definition of a quantile) within each distribution, thereby automatically omitting the bias. The full derivation is outlined in appendix A. By using quantiles, only one degree of freedom is left within the contingency table (Appendix A). It is hence sufficient to acquire a single entry of the contingency table to determine the joint distribution.

Regardless of the contingencies, the bias can simply be assessed by the absolute or relative quantile difference QD or QD′ (Appendix B). Contrary to the approach of Ferro et al. (2005), quantile differences are not transformed here, but reveal differences of rainfall distributions both in location and scale. For instance, an underforecasting of light rainfall can be easily distinguished from an overforecasting of heavy rainfall. The overall offset, covering the entire distribution, can be reproduced by a weighted integral over the absolute value of all quantile differences (Appendix B, Eq. 4). In this way, amplitude errors are summarized and the dissimilarity of the considered distributions are quantified. As noted earlier, the PSS (equivalent to the true skill statistics and the Hanssen–Kuipers discriminant) is able to measure skill without being perturbed by the base rate (e.g. Woodcock, 1976, Mason, 1989). The sensitivity to hedging increases with the rarity of events (Stephenson, 2000), as the PSS converges towards the POD. However, quantile–based contingency tables overcome possible biases. If the number of predicted events are modified in the model output or later in the issued forecast, the definition of a quantile requires that they are compared to the same number of observed events. Consequently, quantile–based scores circumvent the problem of hedging. In respect thereof, the PSS is best suited to display the skill in our balanced setup and able to quantify the overlap, or the permutation, of bias–adjusted datasets (Appendix C). In the case of equal marginal totals (Appendix A, Eq. 1), the PSS merges into a pure ranking of misses (Appendix C, Eq. 6). If the proportion of misses with respect to their random expectation remains fixed, all forecasts receive the same constant rating. Given that a certain ratio fulfills the definition of a constant forecast, it is permissible to compare PSS values for different quantile probabilities or base rates directly. Similarly to the QD, the weighted integral of the PSS over the whole range of quantiles (Appendix C, Eq. 9) summarizes permutation errors over the entire range of intensities and the overall matching characteristics can be condensed in a single number.

Following the above derivation, we are left with two independent measures QD/QD′ and PSS which concertedly characterize the overall forecast error, namely bias and permutation/overlap/matching error. Accordingly, the sample uncertainty also subdivides into the two parts (Appendix D). From the course of the confidence intervals in Fig. 12, it can be gleaned that the quantile-based PSS remains much better defined for rare events than the conventional amplitude–based PSS. Quantile probabilities inherently are not affected by amplitude uncertainties, but their transformation to precipitation amounts suffers from ambiguities. We can achieve a high confidence for the PSS value of a certain quantile, but still hold a low confidence for the precipitation estimation of the quantile. Some applications require a link to rainfall amounts whereas others only require a link to frequencies which is automatically provided by quantiles.

4 Climatology of observations

Before we turn to present quantile–based error characteristics, it is helpful to consider observed frequency distributions with respect to seasonal and regional variations. Note first from Fig. 2 that the vast majority of days (∼80%) within the 6.5 year period experience precipitation of less than 5mm/day. Hence, the following section focuses on the upper part of the distributions with higher rainfall amounts.

During winter, the lowest precipitation amounts occur over Switzerland (Fig. 2a). Convection is suppressed by snow cover and a highly stable boundary layer. Consequently, precipitation events in uniform airmasses are sparse and accumulations are dominated by frontal passages. Weak and moderate rainfall events (<15 mm/day) are rarest in the Grison and strong events (>15 mm/day) are rarest in the Middleland. Modest rainfall (∼15 mm/day) is recorded most frequently in northern parts of Switzerland. On average, there are 10 days per winter with accumulations of more than 10 mm in the Middleland and the Northern Alps. If at all, heavy precipitation occurs very sparely in southern parts of Switzerland. Comparatively high amounts are recorded over the Jura mountains, the Northern Alps and at the western edge of the Valais. These maxima can be attributed to winterly weather regimes with westerlies and northwesterlies (see also Plaut and Simonnet, 2001).

During equinoctial seasons, regional distributions are more or less similar to the north of the Alps, but differ considerably to the south (Fig. 2b and Fig. 2d). To the north, low and medium intensities (<30 mm/day) are slightly more frequent during spring, whereas high intensities (>30 mm/day) are slightly more frequent during fall. 10 mm are exceeded on 10 to 16 days during a season. Since northerly wind directions are more dominant at the beginning of the year (Plaut and Simonnet, 2001), total amounts over the Northern Alps during spring exceed those during fall. However, northerly flow mostly brings about low amounts and rarely entails strong events. In contrast to the rest of Switzerland, frequencies in southeastern areas considerably differ during spring and fall with double accumulations in fall. Without exception, rainfall frequencies of all intensities are higher late in the year. In the Grisons, 10 mm per day on average is exceeded on 8 days per spring and 9 to 10 days per fall. In the Ticino, the frequency of having more than 10 mm even increases from 11 days during spring to 15 days during fall. Stratospheric intrusions have their climatological maximum both during spring and fall, but their precipitation efficiency is higher between September and November when they transport warmer and moister air towards the Alpine south side (Lin et al., 2001, Martius et al., 2006). Related events can locally deliver more than 150 mm/day in the Ticino. The Valais roughly receives equal rainfall amounts from the north and the south. Light and medium events (<30 mm/day) occur slightly more often during spring, whereas strong events (>30 mm/day) occur much more often during fall. 10 mm is reached on average on 11 days per spring and 9 days per fall. 50 mm is reached on a single day per fall, but even more seldomly per spring.

During summer, highest rainfall is recorded over the Northern Alps (Fig. 2c). 10 mm per day on average is exceeded on 19 days over the northern Alpine ridges and foothills which is 6 to 8 days more often than elsewhere. The Northern Alps receive precipitation most frequently, followed by the Ticino and the rest of Switzerland. However, the Ticino receives the majority of strong events (>25 mm/day), followed by the Northern Alps and the rest of Switzerland. Even though the domains of the Jura and the Middleland contain prominent thunderstorm tracks (Houze et al., 1993), the impact on accumulations is weak. Higher daily accumulations are rarest here by far.

If the spatial rainfall distributions within the predefined regions are considered (Schär et al., 1998, Frei et al., 2006), some local characteristics can be pointed out. Generally, largest amounts are found on the windward side and smallest amounts on the lee side of towering mountains or crests. Specific positions depend on the prevailing flow directions which release precipitation (Plaut and Simonnet, 2001). Thus, relatively small amounts are found on the Alpine south side during winter and on the Alpine north side during fall. Outstanding locations with lowest values are valleys and basins in the interior of the mountains. The average remains below 3 mm per day here. The overall maximum is found in a confined area around the highest summits within the northern Alpine crest (marked in Fig. 1). Including dry days, the seasonal mean constitutes 9 mm per day during summer here, whereas other maxima further to the northeast are 2 mm lower. The strong events during fall entail a seasonal mean of about 8 mm per day over the western part of the Ticino.

5 Seasonal error climatology

In the following, the focus is on seasonal error discrepancies within the complete 6.5 year period. Quantile differences and PSS values are computed for all seasons independently. First of all, maps for 90% quantiles are discussed revealing gridbox–scale error variations. The corresponding thresholds are exceeded every tenth day and expected to represent general QPF errors of reasonably strong events. Then the grid–point based scores are aggregated, i.e. applied for regions as a whole. Verification measures are computed for the compound data (grid points × days) of our six predefined domains (Fig.1) and discussed with respect to the whole range of intensities. Finally, the regional model performance is reviewed with the aid of integral error values summarizing the overall performance.

5.1 Grid–point based verification of 90% quantiles

Throughout all seasons, a strong wet bias is evident over the Northern Alps (Fig.3). During the course of the year, the 90% quantiles roughly account between 10 mm/day (winter) and 25 mm/day (summer) in the observations. However, they are more than 20 mm/day or locally 30 mm/day higher in the model forecasts. This bias is confined to few grid points during spring and fall but affects much more grid points during winter and summer. In addition to the Northern Alps, some parts of the Valais and the Grisons are also significantly overforecast. Even though the observed 90% quantiles mostly remain below 10 mm/day during winter, spring and fall, the overestimation in the model partly amounts up to 20 mm/day here. Thus, the relative amplitude errors are largest by far in these areas. In the Ticino, the bias is similar to the Northern Alps during spring and summer, but it almost vanishes during winter and only is noticeable at few grid points during fall. A closer look reveals that the bias is linked directly to the hillside of the topography. In the northern and central parts of the Alps, northwest aligned slopes are strongly overforecast whereas southeast aligned slopes are weakly or moderately underforecast. In the Ticino, the strongest overprediction resides over southerly oriented slopes, but only appears clearly during spring and summer and at specific grid points during fall (Fig. 3d). The deep valleys in the interior of the Alps and their direct surroundings are clearly underforecast. The strongest dry bias is found directly to the south of the northern Alpine crest. It is most pronounced during summer with peak values around 20 mm/day (Fig. 3c).

Regardless of the bias, seasonal matching characteristics vary substantially within Switzerland (Fig. 4). Generally, the regional PSS pattern is noisiest during summer and smoothest during fall. During winter, the matching clearly is linked to the topography. Western slopes display much higher PSS values than their eastern counterparts at that time. In particular, there is a distinct spatial PSS minimum (Fig. 4a) to the southeast of the highest point in the northern Alpine crest (marked in Fig. 1). Noteworthy, the best pattern overlap is found in the Ticino throughout the year. Both during winter and fall, the PSS locally reaches values over 0.8 in the Ticino corresponding to a POD over 0.82 (see Appendix C, Eq. 8 for this calculation). All over Switzerland, the poorest matching is detected during summer (Fig. 4c). The PSS barely passes 0.6 in the Ticino and some places elsewhere and only varies between 0.2 and 0.4 in the Middleland, the Valais and the Grisons. Note, that a PSS of 0.2 only implies a POD of 0.28 (see Appendix C, Eq. 8). During equinoctial seasons, the matching is different on both sides of the Alps. The pattern overlap in the Ticino is poorer during spring compared to fall, even though it is still superior to other regions. In contrast, it is a little improved in the Valais and the Grisons during spring compared to fall. On the Alpine north side, there are distinct sectors with PSS values around 0.35 and 0.55 during spring, whereas the matching is surprisingly uniform with values around 0.5 during fall. The clearest and most consistent regional separation is found late in the year. During fall, the pattern overlap is superior in the Ticino, inferior in the Valais and in the Grisons and middle–rate further to the north. Simultaneously, a prominent PSS gradient is present to the north of the Ticino and over the northern Alpine crest.

5.2 Regional verification of quantile courses

To detect possible error variations for different intensities, it is necessary to consider the whole range of quantiles. Therefore, our verification measures are applied to entire regions. The respective quantiles of the forecasts are given in Fig. 5 as reference and those of the observations are given in Fig. 2 for comparison of the 90% and 95% quantiles. Most notably, the Northern Alps and the Ticino stand out, as they mostly display much higher modeled quantiles than elsewhere. This behavior only partly applies to the observed quantiles which are considerably lower over the Northern Alps for 90% and 95%. This discrepancy is consistent with our earlier finding that the Northern Alps are highly overforecast nearby the 90% quantile (Fig. 3). The seasonal variations of high regional quantiles corresponds relatively well between model and observations. Figure 6 displays the quantile–based measures QD and PSS for different quantiles. Since the 50% quantiles mostly fall below 0.5 mm per day, dry quantiles beneath are omitted in the graphs. Concerning the bias, there is the general tendency that weak intensities are slightly overforecast, medium intensities are slightly under– or overforecast and strong intensities are strongly overforecast. However, distinct regions are exceptions from this archetype. On the one hand, the following domains are severely underforecast for almost all quantiles: the Jura during winter, spring and fall and the Valais during summer. On the other hand, the following domains are greatly overforecast for all quantiles: the Northern Alps all over the year and the Ticino during spring and summer. Interestingly, heavy precipitation in the Ticino is underforecast during winter and fall. Although the matching is superior for the same cases (Figs. 6d and 4), the extremely high observed precipitation amounts during fall (Fig. 4d) cannot be reached by the model.

Concerning the matching, the Ticino mostly holds the highest PSS values compared to the rest of Switzerland. Only low quantiles (nearby 50%) partly match better in other regions. As opposed to the south, the PSS curve is very similar in all northern regions, where the distributions of the Jura, the Middleland and the Northern Alps are close. Only during summer and for low quantiles, PSS values over the Northern Alps diverge slightly from the regions further to the west. The skill for weak intensities is slightly reduced and for strong intensities it is slightly enhanced here. In comparison to northern territories, the PSS is up to 0.1 worse in the central Alps. Except for stronger intensities during summer, the pattern overlap always is most critical in the Valais and the Grisons.

5.3 Rating of integral values

To survey the overall model performance with respect to individual seasons and domains, we discuss weighted error integrals as a condensed view of Fig. 6 (see Appendix B, Eq. 4 and Appendix C, Eq. 9; Tabs. 2 and 3 for further details). The integral bias ($\overline{\mathrm{QD}'}$) is highest during winter. $\overline{\mathrm{QD}'}$ constitutes 0.45 over the Northern Alps which is equivalent to a deviation of over 50%. This means in this case that all intensities are overestimated on average by half of their value. At the same time, $\overline{\mathrm{QD}'}$ amounts to 0.27 and 0.33 over the Jura and over the Grisons, respectively. This is equivalent to a deviation of over 30%. On the Alpine south side, the relative bias is highest during spring. The Ticino entails a $\overline{\mathrm{QD}'}$ of 0.36 between March and May which implies a deviation of over 40%. Lowest model biases are observed over the Middleland during spring and over the Grisons during summer. Interestingly, $\overline{\mathrm{QD}'}$ only accounts for 0.04 over the Grisons during summer, meaning that the rainfall distributions only differ by 4% throughout the domain. The integral matching score ($\overline{\mathrm{PSS}}$) generally is lowest during summer. Throughout Switzerland, the integral value is 0.06 lower between June and August than in other months. All over the year, the PSS is highest in the Ticino. Especially during fall, the integral value is almost 0.2 higher here than for the rest of Switzerland. Other outstanding regions are the Jura and the Middleland during spring and to a lesser extent the Northern Alps during winter. These areas exhibit a good matching and rainfall shifts are comparatively small here. In contrast, the overall worst matching is seen in the Jura, the Middleland, the Valais and the Grisons during summer as well as to a slightly smaller extent in the Valais and the Grisons during fall. The lowest value of 0.34 for the Jura and the Valais during summer corresponds to roughly 75% more mismatching grid points than for the largest value of 0.63 for the Ticino during winter.

6 Differentiation of model designs

There are three decisive model updates within the preoperational and operational phase of the COSMO model which possibly affect QPF to a large extent. First of all, a continuous assimilation cycle (Stauffer and Seaman, 1994) replaced a pure interpolation of initial values from the driving model. Wind, pressure, temperature and humidity have been nudged towards surface and upper–air observations since then. Later, there was a changeover of the driving model at the boundaries. The COSMO model was driven by the GME in the early stages, before the IFS has been used in exchange. The switch–over allowed the COSMO model to benefit from high–quality features of the ECMWF model such as the 4D Var analysis. Finally, the prognostic precipitation scheme substituted the diagnostic treatment with the assumption of column equilibrium for precipitation particles. Hence, hydrometeors have been advected by the ambient flow with different sedimentation velocities for snow and rain.

We apply our refined measures to quantify the impacts that the model updates had on QPF performance. The four adjacent periods separated by these three updates are evaluated separately in the following. The first phase (named IGD) comprises all days between 01/07/2000 and 30/10/2001. It features the interpolation for the initialization, the GME at the boundaries and the diagnostic precipitation. The second phase (NGD) ranges from 31/10/2001 until 15/09/2003 and features the nudging assimilation. The third phase (NID) uses the IFS instead of the GME boundary fields and covers all days between 16/09/2003 and 15/11/2004. The latest model setup (NIP) runs additionally with the prognostic precipitation and covers the period between 16/11/2004 and 31/12/2006. Note that the seasonal composition of the four slices is very similar. Even though the interannual variability of the errors is strong, we consciously consider all seasons together to retain reasonable sample sizes. In terms of overall amounts, IGD entails highest rainfall values in comparison to the phases afterwards. In particular, there were persistent rainfall episodes from October until November 2000 and from March until April 2001. The former were active on the Alpine south side and the latter mostly on the Alpine north side.

6.1 Grid–point based verification of 90% quantiles

During IGD, error patterns are significantly noisier than afterwards (Figs. 7a and 8a). Reasons are found during the first four months of the preoperational phase, when the model ran without a filtered orography (also pointed out in the outlook of Kaufmann et al., 2003). Daily precipitation fields were very spotty at that time and did not succeed in capturing area–wide rainfall. However, principal strengths and weaknesses of the COSMO model are already evident in the primary model setup. The overestimation over the northern Alpine foothills, the underestimation in the interior of the Alps, the good matching in the Ticino as well as the poor matching in deep valleys are obvious during IGD. The dry bias along the low elevations of the Middleland and especially inside the Alps affects a much larger area than afterwards. More than 40% of all Alpine grid points over 1500 mMSL are underforecast with an offset worse than 5 mm/day. However, observed 90% quantiles are roughly 5 mm/day higher than later on, meaning that the relative offset is comparatively low. The pattern overlap is most deficient at the western edge of the Valais and over the Jura mountains. At some grid points, the PSS only constitutes about 0.15 which is equivalent to a POD around 0.24 (see Appendix C, Eq. 8). The next two phases NGD and NID are unlike IGD but similar to each other. The percentage of severely underforecasted grid points over 1500 mMSL only drops down from 20% during NGD to 14% during NID (Figs. 7b,c). The only outstanding difference is the temporally diverging matching in the south (Figs. 8b,c). During NGD, the pattern overlap is average both in the Valais and the Ticino in comparison to other phases. During NID, it is inferior in the Valais and superior in the Ticino. Reasons for this disparate trend are not obvious and need to be investigated in further detail. The matching with the IFS at the boundaries is considerably better in the Jura and the Middleland than before. In contrast to the central Alps, the PSS exclusively remains over 0.3 here.

The change of the precipitation scheme implicates the most conspicuous forecast improvements. Very clearly, the matching improved all over the country (Fig. 8d). During NIP, the PSS varies between 0.3 and 0.8 without exception. Above all, the pattern overlap in the main Alpine valleys improves significantly. Former PSS values around 0.3 rise to values around 0.5 which implicates that the hits H are 1.5 times more frequent than before. The poorest matching still is seen throughout the Grisons and over some parts of the northern Alpine ridges. Aggregated over the Valais, the PSS for the 90% quantiles rises from 0.33 during NID to 0.54 during NIP. Aggregated over Switzerland, it rises from 0.47 to 0.56. In contrast to the pattern overlap, the bias over the Alps displays an ambivalent trend during NIP. Precipitation now is advected to areas which have been underforecast before and maxima of the underestimation are now diminished. The dry bias rarely falls short of 5 mm/day and does not fall short of 10 mm/day any more. However, maxima of overestimation are broadened simultaneously and partly extend to the lee inside the Alps. Peak values of 30 mm/day only drop down to 20 mm/day. Therefore the areal wet bias strongly worsens over some Alpine slopes and in particular in the center of the Grisons. Aggregated over the Grisons, the overcharge of the 90% quantiles rises from 0 mm/day during NID to 2.9 mm/day during NIP. Aggregated over Switzerland, it changes from −0.1 mm/day to 1.6 mm/day.

6.2 Regional verification of quantile courses

Developments of regional error characteristics for a range of rainfall intensities can be gleaned from Fig. 10. The respective quantiles of the forecasts are given in Fig. 9 as reference. Most notably, modeled quantiles over the Jura and the Ticino are highest during IGD, whereas they are standard afterwards. Astonishingly, the regional bias in the first model setup IGD is remarkably small, especially for low intensities (Fig. 10a), even though related quantiles display higher rainfall amounts than afterwards (Fig. 9a). However, it has to be kept in mind that the aggregated quantile difference does not display local rainfall shifts within one region. Thus, a simultaneous overestimation and underestimation within one domain cancel each other out. Given the noisiness of the IGD field, this might explain the small QD values. During NGD, regional amplitude errors increase considerably, above all for lower intensities. Simultaneously, the PSS rises in all regions considerably. The improvements are largest in the northern regions. The PSS in the Jura, the Middleland and the Northern Alps is enhanced by almost 0.1 for small quantiles (∼70%). The alteration stands out least in the Ticino where the performance already has been very good. Improvements between NGD and NID are not so clear. In the central Alps, the matching is insufficient for most quantiles. Above all, the PSS is lowered in the Valais for quantiles between 75% and 95%. In contrast, the overlap is excellent for almost all quantiles in the Ticino. In regions to the north, the matching is slightly worse than before for quantiles below 75%, but it is slightly improved for quantiles above. Considering NIP, improvements for all quantiles are remarkable. The values of the PSS only slightly drop off towards high quantiles. The performance for 60%, 70% and 80% quantiles is very similar now. However, an increase of the overestimation is also obvious for all quantiles. Merely the flat Middleland remains almost bias–free. The amplitude error in the Ticino remains, but it is drastically increased to the north. In particular, low and medium intensities are now significantly overforecast in all Alpine areas. For example, the 70% quantile offset increased in the Northern Alps from 0.6 mm/day to 1.8 mm/day. Note, that the 70% quantiles roughly correspond to 3 mm/day in the observations.

6.3 Rating of integral values

The integrated scores recapitulate the QPF behavior explained above. The integral performance strongly depends on regional characteristics and QPF improvements deeply vary among the domains. The average quantile difference

(Tab. 4) clearly increases over the Jura between IGD and NGD. $\overline{\mathrm{QD}'}$ rises from 0.07 to 0.29 implying that relative deviations increased from 7% to almost 35% along the rainfall distributions. At the same time, amplitude errors level off over the Middleland and diversely changed over the Alps. However, the most conspicuous worsening of the bias is obvious during NIP. All over Switzerland $\overline{\mathrm{QD}'}$ rises from 0.09 to 0.22 implying that deviations grow from 9% to 25%. None of the Alpine regions has its lowest $\overline{\mathrm{QD}'}$ value in the last period with the most sophisticated model version. The change is most critical over the Ticino and the Grisons where $\overline{\mathrm{QD}'}$ rises from 0.14 to 0.30 and from 0.07 to 0.37, respectively. The bias only improves over the Jura and the Middleland at the time. Interestingly, $\overline{\mathrm{QD}'}$ continuously increases within the four phases for the Northern Alps. The value rises from 0.16 during IGD to 0.38 during NIP. Concerning $\overline{\mathrm{PSS}}$ (Tab. 5), QPF performance is best for the latest model version. Except for the Ticino, where the pattern overlap already has been superior before, all regions display their highest $\overline{\mathrm{PSS}}$ during NIP. Thus, the great improvement of advecting precipitation by the ambient flow is brought out clearly by the summary measure. Considering all phases, the matching is upgraded most in the Jura and the Valais where $\overline{\mathrm{PSS}}$ rises by almost 0.2 between IGD and NIP.

7 Interpretation of verification results

Some aspects are worthwhile noting in the interpretation of our results. Firstly, the observational precipitation analysis is not perfect. In particular, it exhibits a negative bias (too low values) due to systematic measurement errors. In the Alpine region, these biases are less than 12% from spring to fall, but can reach several tens of percent in winter for exposed stations above 1500 mMSL (Sevruk, 1985, Richter, 1995). While these biases may possibly explain some part of the apparent quantile overestimation in winter, the overall characteristics of the model errors remain valid. The magnitude of the model bias is substantially larger than expected measurement biases. Note that only about 10% of the rain-gauge stations are at elevations above 1500 mMSL even in inner Alpine regions (see e.g. Fig. 6 in Frei and Schär, 1998) and hence the bias in the observational analysis is much smaller than that at exposed stations. Moreover, our primary focus is on high quantiles (i.e. intense rainfall or heavy snowfall) for which the measurement bias is considerably smaller than for light events.

Generally, our verification results agree with those found by Elementi et al. (2005) in case studies. First and foremost, the severe overestimation at the Alpine fringe is confirmed. The COSMO model versions under consideration seem to have problems to represent the correct flow and moisture field around orography. The windward side receives too much and the lee side too little precipitation. This behavior is most pronounced with the diagnostic precipitation scheme, but still holds for the prognostic scheme. Further investigations (not shown) have proven that the overestimation mostly stems from the resolved part of the total rainfall (see Kaufmann et al., 2003, for comparison) which holds especially for very strong events. Recent tests with the COSMO model revealed that a three–step Runge–Kutta time integration scheme (Wicker and Skamarock, 2002) partly rectifies the overestimation on the windward side of a mountain (D. Leuenberger (MeteoSwiss), pers. comm.). However, it remains ambiguous, how inconsistencies of the used leapfrog scheme affect the formation of precipitation and cause the offsets on the windward side.

Concerning the bias, the overestimation is most pronounced over higher elevations during winter. Even though the results are slightly falsified by wintertime measurement errors, snowfall seems to be more overstated than rainfall. This can be attributed to problems with the cloud ice scheme for the winters 2003/2004 and 2004/2005 (F. Schubiger (MeteoSwiss), pers. comm.).

Interestingly, the common bias pattern does not always hold for the Alpine south side. In opposition to the findings discussed above, there is a slight dry bias over the Ticino during winter and fall. Obviously, a drying of the water cycle is superimposed to the usual overestimation of precipitation on the Alpine fringe. Comparisons of COSMO forecasts and COSMO analyses revealed that forecasts over the Po valley tend to be significantly dryer than in the corresponding assimilation cycle. Apparently, the water balance in the COSMO model is predisposed to leak over the northwest of Italy. However, the cause of this is not clear and requires further investigation.

A positive correlation between the quantile difference and the PSS can be confirmed statistically for the diagnostic precipitation scheme. In other words, the PSS usually is higher for positive quantile differences than for negative ones. The reason is the absent horizontal transport of hydrometeors which produces both an underestimation and a poor matching on the lee side of mountains. By introducing the prognostic precipitation scheme, this correlation vanishes. The bias no longer interferes significantly with the matching and a main error source which affected both verification components seems to be removed. Note that this behavior only is proven so clearly, because all intensities are related to quantiles/frequencies and not to amplitudes.

At the same time, a significant worsening of the wet bias is evident during the latest model phase. In particular over the Ticino and the Grisons, the relative offset of the distributions in comaparison increased drastically. Most notably, southerly flow now entails a much higher wet bias than before. Possible explanations relate to unmentioned model updates like a change of the cloud ice scheme, the introduction of prognostic turbulent kinetic energy or the switch of the IFS boundary fields to higher resolutions. Again, implications of verification results do not transfer directly to model diagnosis and the original error source remains to be investigated.

One of the most outstanding findings is the high quality of the pixel overlap on the Alpine south side. As shown in section 5, the pattern overlap in the Ticino is much better than elsewhere. The matching disparity to other regions is present throughout the year and throughout all model versions. It is most manifest for strong intensities and most pronounced during winter and fall (Figs. 6a,d). Explanations are found in the special geographical position of southern Switzerland in connection to the prevailing synoptic flow patterns. Meridionally aligned stratospheric intrusions determine the large–scale predictability of heavy precipitation (Fehlmann and Quadri, 2000) and primarily support the observed enhanced pixel overlap during fall. The relief of the Ticino is uniformly aligned to the south and lee effects play a minor role for a southerly flow. The high predictability also is supported by idealized model simulations of Gheusi and Davies (2004). They found that precipitation distributions near the Lago Maggiore are comparatively insensitive to the changes in direction and speed of the incident flow from south to southwest i.e. from 10 to 30 m/s. In addition to its favorable exposure to the south, the Ticino is shielded by the Alpine crest for northerly flow directions. Embedded shallow fronts provide a potential for misforecasts on the windward side, but are obstructed by the Alpine crest and do not affect the Ticino (E. Zala (MeteoSwiss), pers. comm.). Their frequency maximum on the windward side during the cold season (Jenkner et al., 2008) supports this explanation for an enhanced wintertime pixel overlap in the Ticino.

Concerning the seasonal cycle, it is obvious that the pixel overlap is highest during winter (in some regions also during fall) and lowest during summer. In this regard, model performance strongly anti–correlates with the amount of convective precipitation or likewise with the boundary layer height. The dependency emerges clearly in our results and is worthwhile mentioning, even though convection schemes are already well–known to present difficulties to QPF (e.g. Elementi et al., 2005). The interaction of the boundary layer and the free atmosphere is well–developed during the convective season and challenges the interplay of parameterized and resolved processes in today's models. The final outcome is a degradation of the local skill in convective situations.

8 Summary

In the present study a novel treatment of the traditional categorical verification has been presented. By using frequency thresholds instead of amplitude thresholds, deterministic verification applications strongly benefit from inherent advantages. We recommend to use the absolute or relative quantile difference to describe the bias and the Peirce skill score to describe the pattern overlap. In this way, the total error is split up into an amplitude part and a matching part. If multiple quantiles are taken into account, spectral performance can be assessed. This setup allows for a meaningful juxtaposition of different value ranges and renders possible a meaningful evaluation of individual intensities. In this context, our distribution–oriented approach makes a decisive step towards equitable scores, as defined by Gandin and Murphy (1992). In our opinion, it is more meaningful for model developers to relate verification results to characteristic numbers of the model output (e.g. quantiles) than to a priori fixed limits. The main advantages and challenges of the refined approach are itemized here for recapitulation:

- The degrees of freedom within the contingency table reduce to one. The information content can be displayed by a single score.

- Owing to the use of quantiles, the conducted debiasing has a physical validity and only relates to the characteristics of the distributions under comparison. A commensurate calibration cannot be achieved so easily by other approaches.

- The whole range of precipitation intensities can be assessed by looping over quantiles. Different value ranges are related to each other in an expedient and consistent way.
- The Peirce skill score strongly simplifies with the use of quantiles and is no longer susceptible to hedging.
- Synoptic biases or gridding errors do not influence the matching error score, if they are applied to homogenous subsets (in our case orographically distinct areas) with a consistent bias behavior.
- Scores can be integrated over all quantiles while weighting them accordingly. Total performance is condensed meaningfully in a single number.
- As a slight drawback, quantiles universally are less intuitive than fixed thresholds. A verification which is issued to the public mostly requires values to be linked to proper rainfall amounts.

The methodology has been applied to 6.5 year of preoperational and operational forecasts from the Swiss implementation of the COSMO model. Seasons and model versions have been evaluated separately. 90% quantiles have been investigated in detail and quantile courses have been discussed for six predefined regions. The most important results are recapitulated in the following:

- Regional QPF performance strongly is determined by local orography in connection with the impinging flow direction. Distinct areas with a good and poor pattern overlap can be identified. The fine–scale error structure emerges most clearly during winter.

- The COSMO model exhibits a persistent overestimation over the Alpine foreland and a transient underestimation over interior valleys (primarily during winter and summer and with the diagnostic precipitation scheme). The Alpine south side partly is an exception with comparatively low seasonal amplitude errors during winter and fall.

- Overall matching characteristics are worst during summer. All over Switzerland, the pattern overlap is roughly 6–7% worse during summer than during the rest of the year. Seasonal dispartities are even higher over the Jura and the Ticino. Altogether, the matching is worst over the Jura and the Valais during summer.

- The skill is much higher on the Alpine south side than on the north side. On average, the pattern overlap is roughly 15% better over the Ticino than elsewhere.

- The matching continously improves within the validated period and can be clearly attributed to updates of the COSMO model. All over Switzerland, the pattern overlap is 12% better in the latest considered phase than in the first one.

- The overall bias is worst by far in the latest considered phase. The relative amplitude error constitutes about 9% between 2000 and 2004, but it rises to 25% between 2005 and 2006.

The steep orography in Switzerland imposes a severe constraint on QPF performance. The discussed QPF errors presented here exhibit a strong dependency on weather patterns with specific flow features in conjunction with the complex terrain. Thus, it is instructive to study model performance with respect to different synoptic situations. Such a study will be conducted with a consistent model version in a follow–up paper.

Acknowlegdement The authors cordially thank Daniel Leuenberger for providing excellent technical support and together with Emanuele Zala and Francis Schubiger for giving helpful comments concerning the interpretation of results. Additional comments provided by Pertti Nurmi are acknowledged as well. Concerning the funding, we are indepted to the Swiss National Science Foundation who supported our project financially.

A Quantile–oriented binary contingency tables

The entries of the standard 2x2 contingency table (Tab. 1) are defined by means of sample quantiles instead of a preassigned threshold value. Thus, the marginal distributions (H+F, Z+M, etc.) are fixed automatically. If the same quantile probability is chosen for the model and the observations ($p \equiv p_{mod} = p_{obs}$), the event frequency is the same in both datasets. Two additional interrelations are incorporated in the conceptual formulation:

$$\mathrm{M} = \mathrm{F} \qquad \mathrm{M} + \mathrm{Z} = p\mathrm{N} \tag{1}$$

Firstly, the misses M equal the false alarms F and the bias automatically is removed from subsequent considerations. Secondly, the quantile probability p by definition sets the base rate to $1 - p$ and divides the sample into $(1 - p)$N events and pN non-events. Thus, all four entries H, M, F and Z additionally are linked to p itself. Since the sample size N=H+M+F+Z is given, the four incidences are connected by a total of three constraints. Only one degree of freedom is left and uniquely describes the joint distribution.

The determination of p imposes stringent restrictions on the valid range of the four entries in the contingency table. Depending on p, the four counts only vary within a limited span (Fig. 11). If the quantile probability is below 0.5, there are always some hits H by definition, as exceeding events cover more than half of the sample. If the quantile probability is above 0.5, there are always some correct negatives Z by definition, as exceeding events cover less than half of the sample. The misses M and false alarms F consistently are limited at the top. They are restricted either by the number of non-events ($p < 0.5$) or by the number of events ($p > 0.5$). Just in case of the median ($p = 0.5$), all four counts hold the same range of values. Then the setting is balanced and the number of hits H by definition equals the number of correct negatives Z.

A random forecast divides the joint distribution into a region with skill and one without skill. The borderline is $(1-p)^2\mathrm{N}$ for the hits, $(p-p^2)\mathrm{N}$ for the misses or false alarms and $p^2\mathrm{N}$ for the correct negatives (Fig. 11). Thus, a valuable forecast always exhibits less than $(p-p^2)\mathrm{N}$ misses or false alarms respectively. In case of the median ($p = 0.5$), the border consistently resides at N/4 which is in the center of valid ranges. It gradually approaches a margin for p converging towards 0 or 1.

B Quantile differences

Certain sample quantiles are useful in an exploratory rainfall verification. Even though the quantile–quantile plot is useful in small datasets, it is beneficial to map quantile differences in gridded samples (Ferro et al., 2005). In our context, the quantile difference directly exhibits a distinctive amplitude error (i.e. bias between the datasets):

$$\mathrm{QD}(p) = q_{mod}(p) - q_{obs}(p) \tag{2}$$

As an option, QD can also be standardized by individual quantiles. Similar to the definition of the amplitude component of the SAL measure (Wernli et al., 2008), the scaling of the error preferably is done by the arithmetic mean of the observation and the forecast:

$$\mathrm{QD}'(p) = \frac{2\,\mathrm{QD}(p)}{q_{obs}(p) + q_{mod}(p)} \tag{3}$$

As a matter of fact, QD describes the absolute quantile difference whereas QD$'$ describes the relative quantile deviation. Note that QD$'$ varies between -2 and 2. A positive QD or QD$'$ corresponds to a wet bias i.e. an overestimation of precipitation. A negative QD or QD$'$ stands for a dry bias i.e. an underestimation of precipitation. As the case may be, it is more convenient to display QD or QD$'$.

However, it is advisable to use the relative quantile deviation to express the overall performance over all precipitation magnitudes. The weighted average over all quantiles accounts for the collective rainfall deviation. The integral preferably is computed with the absolute value, because deviations with a varying sign cancel out in the average. Since the quantile difference originally is an additive quantity, we use the arithmetic mean of observed and modeled quantiles as weighting function:

$$\overline{\mathrm{QD}'} = \frac{1}{\int w(p)dp} \int_0^1 w(p)|\mathrm{QD}'(p)|\; dp \qquad w(p) = \frac{q_{obs}(p) + q_{mod}(p)}{2} \tag{4}$$

Note that $\overline{\mathrm{QD}'}$ only varies between 0 and 2. Amplitude deviations of 5%, 10%, 20%, 40% and 80% lead to $\overline{\mathrm{QD}'}$ values of 0.049, 0.095, 0.182, 0.333 and 0.571, respectively. The lower the value of $\overline{\mathrm{QD}'}$ the more alike are the compared distributions in terms of intensities and the smaller is the overall amplitude error.

C The Peirce skill score with quantiles

The Peirce skill score (Peirce, 1884) can be reformulated with an offset for selected quantile probabilities $dp = p_{mod} - p_{obs}$ with p_{mod} and p_{obs} denoting the quantile probabilities of the chosen amplitude threshold:

$$\mathrm{PSS}(p) = \frac{\mathrm{H}}{\mathrm{H}+\mathrm{M}} - \frac{\mathrm{F}}{\mathrm{F}+\mathrm{Z}} = 1 - \frac{\mathrm{M}}{(p_{obs} - p_{obs}^2)N} - \frac{dp}{p_{obs}} \tag{5}$$

Hence, a possible bias (dp) changes the PSS more intensely for large base rates $(1 - p_{obs})$ than for small ones. The problem of hedging results from the fact that people may change p_{mod} while p_{obs} remains fixed. However, it is eluded, if quantile probabilities coincide to each other ($p \equiv p_{mod} = p_{obs}$, see Appendix A for details) and a change of p_{mod} always implicates a change of p_{obs}. Several skill scores merge into each other in that situation. In particular, the Peirce skill score (Peirce, 1884) gets equal to the Heidke skill score (Heidke, 1926) and the Clayton's skill score (Clayton, 1934). Its formula further simplifies to:

$$\mathrm{PSS}(p) = 1 - \frac{\mathrm{M}}{\mathrm{M}_{rand}} \qquad \mathrm{M}_{rand} = (p - p^2)\mathrm{N} \tag{6}$$

The symmetry of the random misses M_{rand} (see Fig. 11b for visualization) ascertains the PSS to be invariant with respect to taking the complement of the events (see Stephenson, 2000, for details). Therefore it is equivalent to define the quantiles going upward or downward along the distribution. In addition, a swapping of forecasts and observations is allowed in our setup and the PSS fulfills the requirements for transpose symmetry (see Stephenson, 2000, for definition).

Positive PSS–values indicate skill compared to a shuffled sample, since a random forecast results in PSS$= 0$ at all times. Owing to the limited range of the four entries in the contingency table (see Appendix A for details), the PSS only varies between $1 - 1/(1-p)$ and 1 for $p \leq 0.5$ or accordingly between $1 - 1/p$ and 1 for $p \geq 0.5$. Different probabilities for a complete miss of all events are included in the concept and it is not possible to score much worse than coincidence in the case of rare events or rare non-events.

Using the definitions introduced by Gandin and Murphy (1992) the PSS can be computed in different ways corresponding to the four entries of the matrix product between the symmetric scoring matrix **S** and the symmetric performance matrix **P**:

$$\begin{gathered} \mathbf{SP} = \begin{pmatrix} 1/p & -1/p \\ -1/(1-p) & 1/(1-p) \end{pmatrix} \mathrm{PSS}(p) \\ \mathbf{S} = \begin{pmatrix} p/(1-p) & -1 \\ -1 & (1-p)/p \end{pmatrix} \qquad \mathbf{P} = \frac{1}{\mathrm{N}} \begin{pmatrix} \mathrm{H} & \mathrm{M} \\ \mathrm{M} & \mathrm{Z} \end{pmatrix} \end{gathered} \tag{7}$$

To interpret the PSS, it can be additionally transformed into the POD. However, the convenient characteristics of the PSS are then lost:

$$\mathrm{POD}(p) = 1 - p\,(1 - \mathrm{PSS}(p)) \tag{8}$$

Individual PSS values can be integrated to express the overall matching performance. The weighted average over all quantiles accounts for the collective skill. Since PSS is a multiplicative quantity, the geometric mean of observed and modeled quantiles is used as weighting function:

$$\overline{\mathrm{PSS}} = \frac{1}{\int w(p)dp} \int_0^1 w(p)\ \mathrm{PSS}(p)\ dp \qquad w(p) = \sqrt{q_{obs}(p) q_{mod}(p)} \tag{9}$$

The higher the value of $\overline{\mathrm{PSS}}$ the larger is the pixel overlap and the smaller is the overall permutation error.

D The uncertainty of results

Essentially, the estimation of the sample quantiles is sensitive to the available dataset (e.g. Conover, 1980). Since rainfall distributions usually are highly skewed, there are only some nonparametric methods to quantify the uncertainty of the computed quantiles. A convenient approach is described by Conover (1980). If a dataset is considered independent and identically distributed, the binomial distribution describes the probability that a single data point constitutes the targeted quantile. Confidence intervals are obtained directly, if border quantiles are taken from the following range $[r, s]$ in the order statistics:

$$r = \mathrm{N}p + y_{\alpha/2}\sqrt{Np(1-p)} \qquad s = \mathrm{N}p + y_{1-\alpha/2}\sqrt{\mathrm{N}p(1-p)} \tag{10}$$

Intrinsically, $y_{\alpha/2}$ and $y_{1-\alpha/2}$ denote quantiles of the binomial distribution and $\mathrm{N}p(1-p)$ represents its variance. If N is sufficiently large, the central limit theorem can be applied. Then $y_{\alpha/2}$ and $y_{1-\alpha/2}$ can be taken from the standard normal distribution.

Let us exemplify the proceeding of Conover (1980) by means of the highest evaluated quantile in the smallest and the largest regional sample. The Jura only encompasses N= 23485 data items during the period NID (427 days × 55 grid points, see section 6 for details). Observed and modeled 99% quantiles correspond to 38.4 mm and 27.2 mm respectively. Following the explained method, the confidence interval (95% significance level) ranges from $r = 23220$ to $s = 23280$ which implies [37.4 mm, 39.5 mm] and [25.5 mm, 28.7 mm]. In contrast, the Middleland encompasses N= 210296 data items during the period NIP (776 days × 271 grid points, see section 6 for details). Observed and modeled 99% quantiles correspond to 29.4 mm and 33.6 mm respectively. The confidence interval now ranges from $r = 208104$ to $s = 208283$ which implies

[29.1 mm, 29.8 mm] and [33.1 mm, 34.2 mm]. On account of the larger sample, the uncertainty is much smaller in the second case.

According to the error itself, the overall sample uncertainty subdivides into two different parts. One portion relates to the bias and another one relates to the pixel overlap. On the one hand, the estimation of quantiles directly affects the uncertainty of the quantile difference. Basically, variances add up to that of the difference. On the other hand, the determination of matching and non–matching pixels affects the uncertainty of the Peirce skill score. Hanssen and Kuipers (1965) derived the variance of the PSS by means of parametric assumptions. In our notation, it looks like:

$$var(\mathrm{PSS}(p)) = \frac{1}{\mathrm{N}} \left(\frac{1}{4p(1-p)} - \mathrm{PSS}(p)^2 \right) \tag{11}$$

Following equation 11, the uncertainty of the PSS is highest for extreme quantiles and PSS values close to zero. In contrast, it is lowest for quantiles around the median and PSS values close to 1.

Strictly speaking, the spatial correlation among individual grid points is crucial and must be maintained while estimating uncertainties of regional scores. In contrast, individual days are approximately independent from each other. Thus, meaningful confidence intervals are obtained by fixing the spatial configuration and varying the temporal composition. Proper resampling methods are explained by Ferro et al. (2005). We applied a bootstrap with 500 repetitions and computed quantile–based confidence intervals. That means we resampled available days with replacement and used ordinary quantiles to determine confidence limits. Once again, we use the Jura during NID and the Middleland during NIP for illustration purposes (Fig. 12). The uncertainty generally is much smaller in the larger sample. Confidence intervals of the quantile difference tend to spread when moving to higher quantiles, because they are not scale–invariant. The uncertainty of the Peirce skill score only slightly increases for rare events which is in contrast to a conventional computation based on amplitude thresholds.

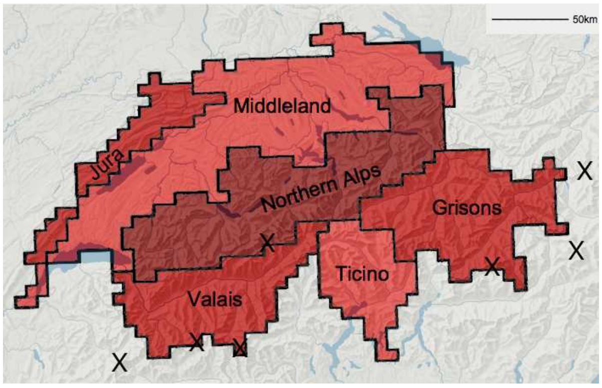

Figure 1: Subdivision of Switzerland into six orographically distinct areas: Jura, Middleland, Northern Alps, Valais, Ticino, Grisons. Square dividing lines indicate the margins of the COSMO grid points. Highest mountain formations are marked with an X.

	observed yes	observed no
predicted yes	H	F
predicted no	M	Z

Table 1: 2x2 Contingency table with hits H, misses M, false alarms F and correct negatives Z.

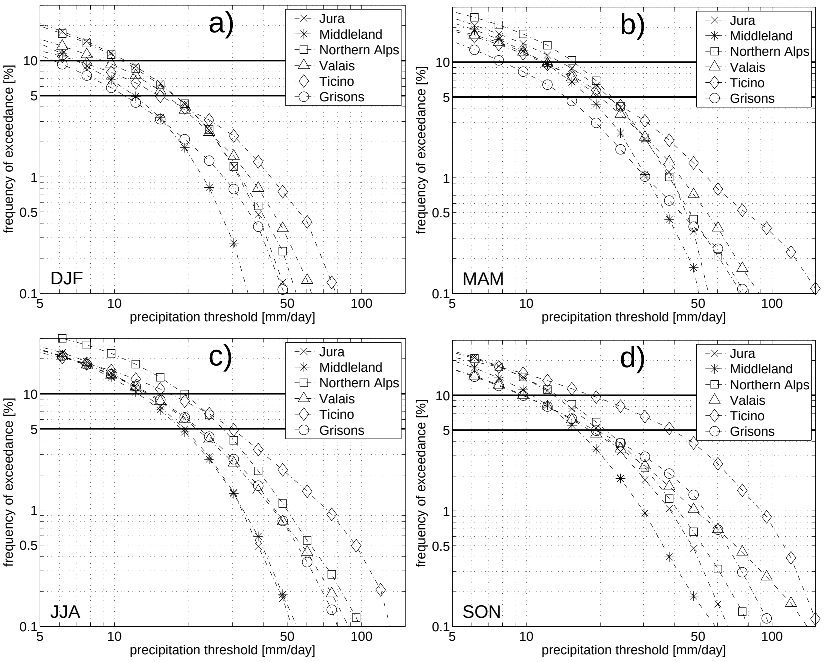

Figure 2: Observational sample probabilities of exceedance of daily rainfall sum in the six different sub–regions defined in Fig. 1: winter (a), spring (b), summer (c) and fall (d). A frequency of 1% corresponds to a marginal day during one season. Frequencies are tantamount to quantiles, as the latter define the limit which a predetermined proportion of the data falls below. For instance, the 90% and the 95% quantiles directly can be read off, if the frequencies of exceedance, namely 10% and 5% (solid horizontal lines), are converted to proper rainfall amounts.

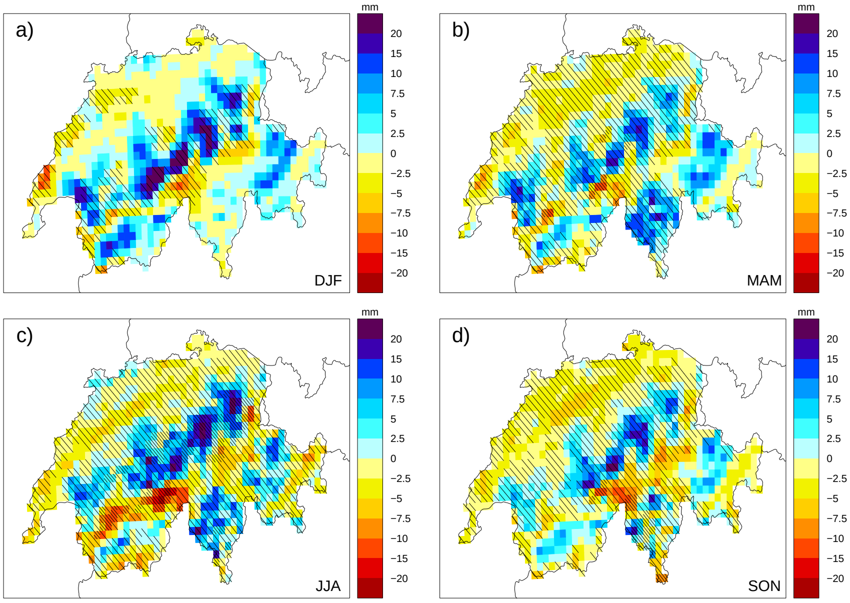

Figure 3: Difference of 90% quantiles (COSMO–observations) of daily precipitation sums [mm/day]: winter (a), spring (b), summer (c) and fall (d). Lightly hatched grid points indicate observed 90% quantiles over 10 mm/day and densely hatched grid points indicate observed 90% quantiles over 20 mm/day.

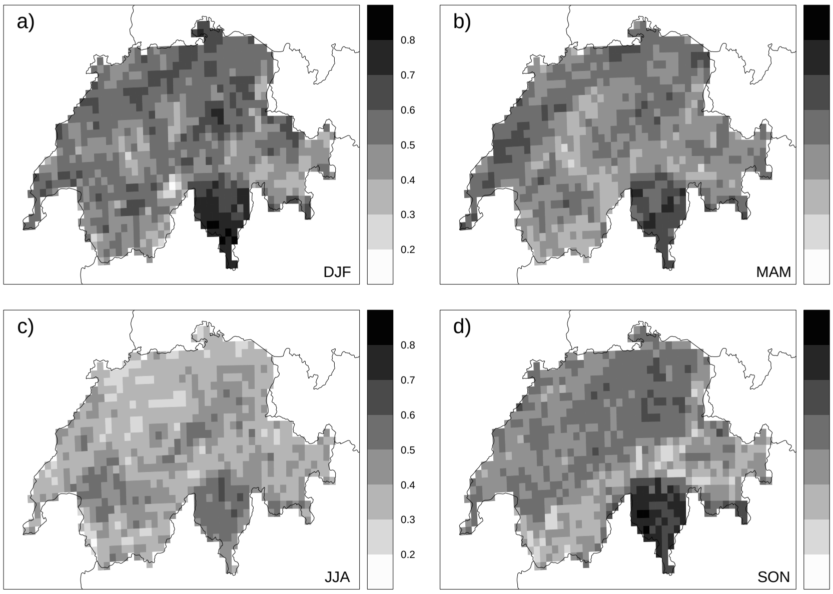

Figure 4: Peirce skill score for 90% quantiles of daily precipitation sums [1: perfect, 0: random forecast]: winter (a), spring (b), summer (c) and fall (d).

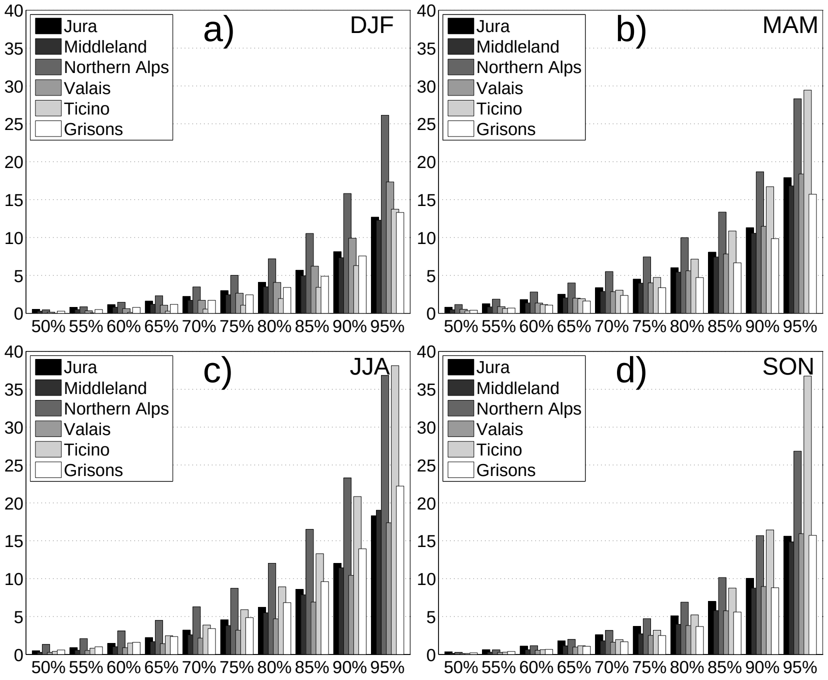

Figure 5: Modeled daily precipitation amounts [mm/day] for discrete quantile probabilities: winter (a), spring (b), summer (c) and fall (d).

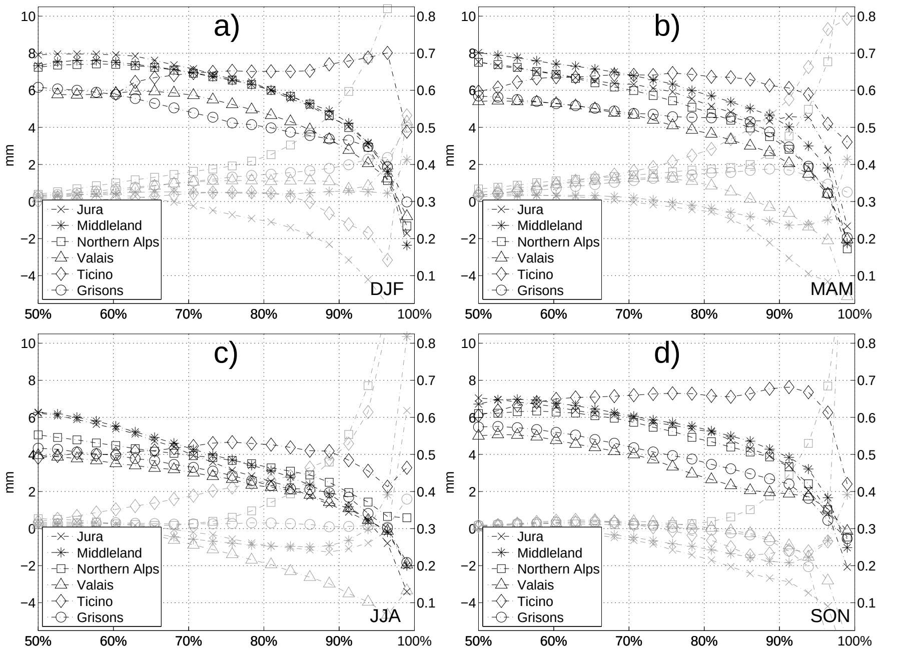

Figure 6: Area–average of the quantile difference (gray, left scale) [mm/day] and the Peirce skill score (black, right scale) for quantiles between 50% and 99%: winter (a), spring (b), summer (c) and fall (d).

	Jura	Middlel.	N. Alps	Valais	Ticino	Grisons	CH
winter	0.27	0.13	0.45	0.18	0.15	0.33	0.24
spring	0.17	0.08	0.25	0.16	0.36	0.22	0.13
summer	0.10	0.13	0.23	0.24	0.30	0.04	0.11
fall	0.22	0.13	0.19	0.12	0.11	0.22	0.06

Table 2: Weighted integral $\overline{\mathrm{QD}'}$ values for different domains and seasons. The scaling shows the relative amplitude deviation and is explained in appendix B [0.1: 10.5% deviation, 0.2: 22.2% deviation, 0.3: 35.3% deviation]. The darker the shading the higher are overall observed amounts.

	Jura	Middlel.	N. Alps	Valais	Ticino	Grisons	CH
winter	0.48	0.46	0.46	0.41	0.63	0.43	0.46
spring	0.49	0.47	0.43	0.38	0.56	0.41	0.45
summer	0.34	0.36	0.41	0.34	0.47	0.36	0.39
fall	0.43	0.46	0.43	0.38	0.60	0.38	0.46

Table 3: Weighted integral $\overline{\mathrm{PSS}}$ values for different domains and seasons. The darker the shading the higher are overall observed amounts.

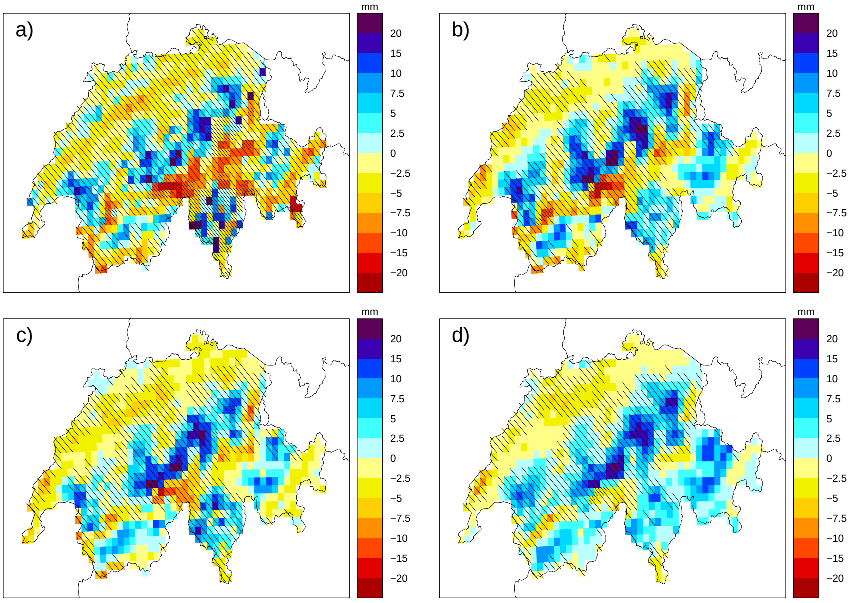

Figure 7: Difference of 90% quantiles (COSMO–observations) of daily precipitation sums [mm/day]: IGD (a), NGD (b), NID (c) and NIP (d). Lightly hatched grid points indicate observed 90% quantiles over 10 mm/day and densely hatched grid points indicate observed 90% quantiles over 20 mm/day.

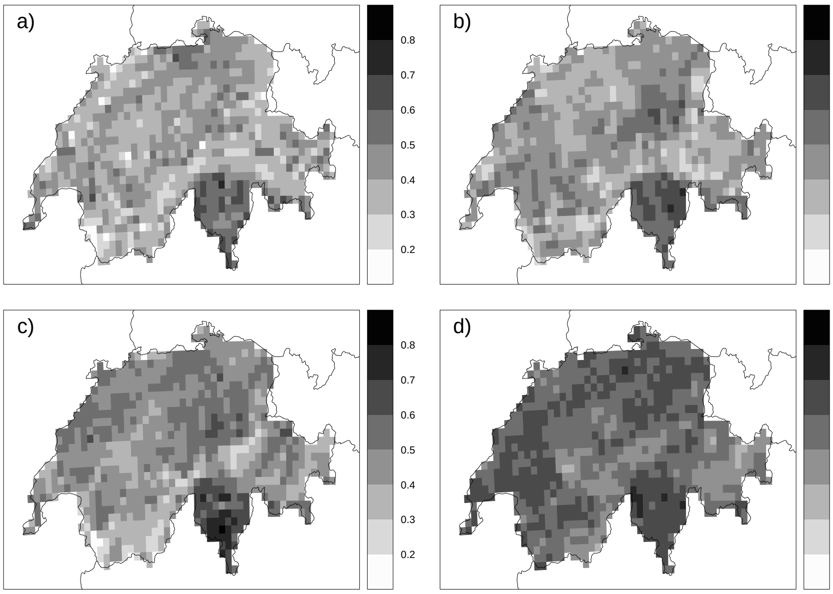

Figure 8: Peirce skill score for 90% quantiles of daily precipitation sums [1: perfect, 0: random forecast]: IGD (a), NGD (b), NID (c) and NIP (d).

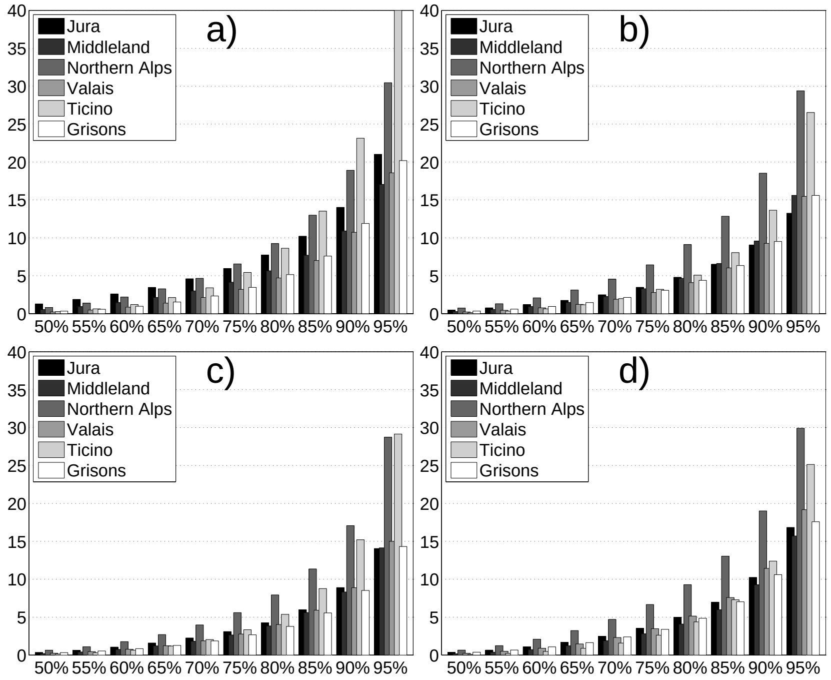

Figure 9: Modeled daily precipitation amounts for discrete quantile probabilities [mm/day]: IGD (a), NGD (b), NID (c) and NIP (d).

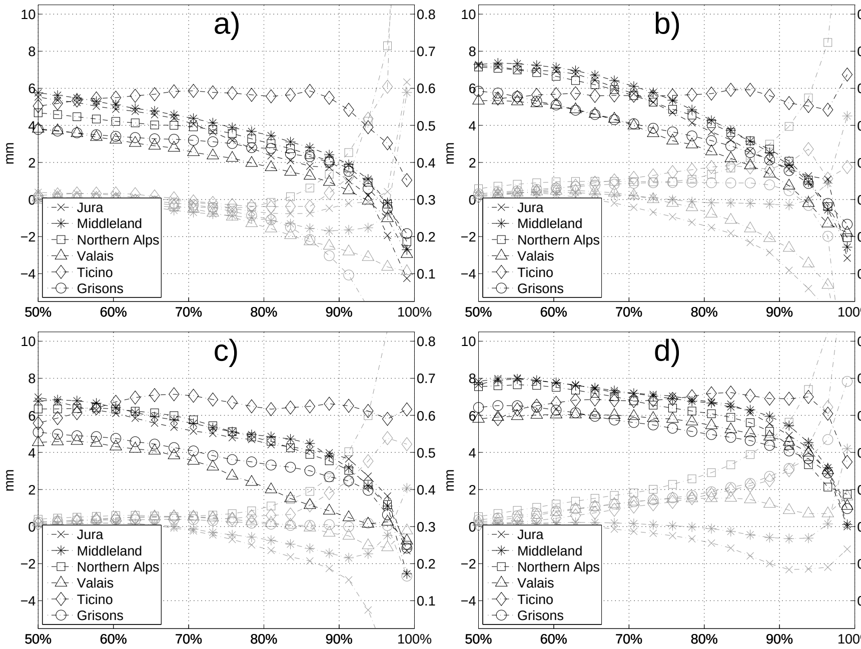

Figure 10: Area–average of the quantile difference (gray, left scale) [mm/day] and the Peirce skill score (black, right scale) for quantiles between 50% and 99%: IGD (a), NGD (b), NID (c) and NIP (d).

	Jura	Middlel.	N. Alps	Valais	Ticino	Grisons	CH
IGD	0.07	0.13	0.16	0.18	0.14	0.22	0.07
NGD	0.29	0.07	0.26	0.23	0.15	0.19	0.09
NID	0.28	0.12	0.28	0.08	0.14	0.07	0.09
NIP	0.13	0.07	0.38	0.17	0.30	0.37	0.22

Table 4: Weighted integral $\overline{\mathrm{QD}'}$ values for different domains and periods. The scaling shows the relative amplitude deviation and is explained in appendix B [0.1: 10.5% deviation, 0.2: 22.2% deviation, 0.3: 35.3% deviation]. The darker the shading the higher are overall observed amounts.

	Jura	Middlel.	N. Alps	Valais	Ticino	Grisons	CH
IGD	0.34	0.38	0.37	0.30	0.49	0.34	0.39
NGD	0.40	0.38	0.39	0.32	0.58	0.35	0.40
NID	0.44	0.42	0.44	0.34	0.61	0.39	0.44
NIP	0.52	0.52	0.51	0.49	0.59	0.48	0.51

Table 5: Weighted integral $\overline{\mathrm{PSS}}$ values for different domains and periods. The darker the shading the higher are overall observed amounts.

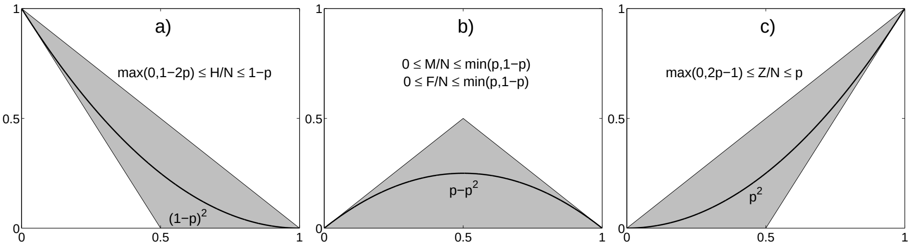

Figure 11: Possible entries (gray shaded) and random expectation values (thick black lines) of the four entries in the 2x2 contingency table: a) hits H, b) misses M or false alarms F, c) correct negatives Z. The x–axis displays the range of quantile probabilities p whereas the y–axis shows the proportion of the sample size H/N, M/N or F/N and Z/N.

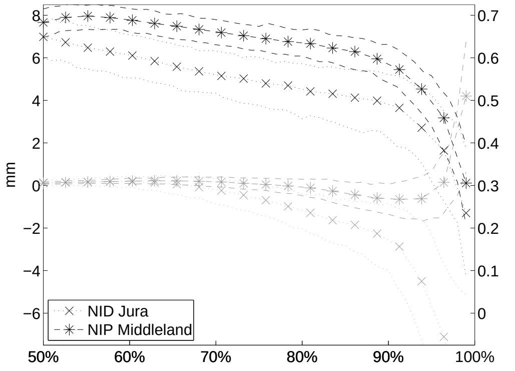

Figure 12: Examples for 95% confidence intervals (lines without markers) obtained from a bootstrap sample with 500 repetitions for the quantile difference (gray, left scale) [mm/day] and the Peirce skill score (black, right scale): The Jura during NID (16/09/2003 until 15/11/2004, 427 days) and the Middleland during NIP (16/11/2004 until 31/12/2006, 776 days) are displayed.

References

Accadia, C., S. Mariani, M. Casaioli, A. Lavagnini, and A. Speranza, 2005: Verification of precipitation forecasts from two limited-area models over Italy and comparison with ECMWF forecasts using a resampling technique. *Wea. Forecasting*, **20**, 276 – 300.

Asselin, R., 1972: Frequency filter for time integrations. *Mon. Wea. Rev.*, **100**, 487 – 490.

Benoit, R., C. Schär, P. Binder, S. Chamberland, H. C. Davies, M. Desgagne, C. Girard, C. Keil, N. Kouwen, D. Luthi, D. Maric, E. Muller, P. Pellerin, J. Schmidli, F. Schubiger, C. Schwierz, M. Sprenger, A. Walser, S. Willemse, W. Yu, and E. Zala, 2002: The real-time ultrafinescale forecast support during the special observing period of the MAP. *Bull. Amer. Meteor. Soc.*, **83**, 85 – +.

Buzzi, A., S. Davolio, M. D'isidoro, and P. Malguzzi, 2004: The impact of resolution and of MAP reanalysis on the simulations of heavy precipitation during MAP cases. *Meteorol. Z.*, **13**, 91 – 97.

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

Clayton, H., 1934: Rating weather forecasts. *Bull. Amer. Meteor. Soc.*, **15**, 279 – 283.

Conover, W., 1980: *Practical Nonparametric Statistics*, Wiley and Sons Ltd, chapter 3. 95 – 142.

Davies, H. C., 1976: Lateral boundary formulation for multilevel prediction models. *Quart. J. Roy. Meteor. Soc.*, **102**, 405 – 418.

Elementi, M., C. Marsigh, and T. Paccagnella, 2005: High resolution forecast of heavy precipitation with Lokal Modell: Analysis of two case studies in the Alpine area. *Nat. Hazards and Earth Syst. Sc.*, **5**, 593 – 602.

Fehlmann, R. and C. Quadri, 2000: Predictability issues of heavy alpine southside precipitation. *Met. Atm. Phys.*, **72**, 223 – 231.

Ferro, C., A. Hannachi, and D. Stephenson, 2005: Simple nonparametric techniques for exploring changing probability distributions of weather. *J. Climate*, **18**, 4344 – 4354.

Frei, C. and C. Schär, 1998: A precipitation climatology of the Alps from high-resolution rain-gauge observations. *Int. J. Climatol.*, **18**, 873 – 900.

Frei, C., R. Scholl, S. Fukutome, R. Schmidli, and P. Vidale, 2006: Future change of precipitation extremes in Europe: Intercomparison of scenarios from regional climate models. *J. of Geophys. Research-Atm.*, **111**, D06105.

Gandin, L. and A. Murphy, 1992: Equitable skill scores for categorical forecasts. *Mon. Wea. Rev.*, **120**, 361 – 370.

Gheusi, F. and H. C. Davies, 2004: Autumnal precipitation distribution on the southern flank of the Alps: A numerical-model study of the mechanisms. *Quart. J. Roy. Meteor. Soc.*, **130**, 2125 – 2152.

Hamill, T. and J. Juras, 2006: Measuring forecast skill: Is it real skill or is it the varying climatology? *Quart. J. Roy. Meteor. Soc.*, **132**, 2905 – 2923.

Hanssen, A. and W. Kuipers, 1965: On the relationship between the frequency of rain and various meteorological parameters. *Meded. Verh.*, **81**, 2 – 15.

Heidke, P., 1926: Berechnung des Erfolges und der Güte der Windstärkevorhersagen im Sturmwarnungsdienst. *Geogr. Ann.*, **8**, 301–349.

Hilliker, J., 2004: The sensitivity of the number of correctly forecasted events to the threat score: A practical application. *Wea. Forecasting*, **19**, 646 – 650.

Hohenegger, C., D. Luthi, and C. Schär, 2006: Predictability mysteries in cloud-resolving models. *Mon. Wea. Rev.*, **134**, 2095 – 2107.

Hohenegger, C. and C. Schär, 2007: Predictability and error growth dynamics in cloud-resolving models. *J. Atmos. Sci.*, **64**, 4467 – 4478.

Houze, R. and S. Medina, 2005: Turbulence as a mechanism for orographic precipitation enhancement. *Quart. J. Roy. Meteor. Soc.*, **62**, 3599 – 3623.

Houze, R., W. Schmid, R. Fovell, and H. Schiesser, 1993: Hailstorms in Switzerland - left movers, right movers, and false hooks. *Mon. Wea. Rev.*, **121**, 3345 – 3370.

Jenkner, J., M. Sprenger, I. Schwenk, C. Schwierz, S. Dierer, and D. Leuenberger, 2008: Detection and climatology of fronts in a high-resolution model reanalysis over the Alps, to be submitted.

Kållberg, P. and A. Montani, 2006: A case study carried out with two different NWP systems. *Nat. Hazards and Earth Syst. Sc.*, **6**, 755 – 760.

Kaufmann, P., F. Schubiger, and P. Binder, 2003: Precipitation forecasting by a mesoscale numerical weather prediction (NWP) model: Eight years of experience. *Hydr. and Earth Syst. Sci.*, **7**, 812 – 832.

Kessler, E., 1969: On the distribution and continuity of water substance in atmospheric circulations. *Meteor. Monog.*, **10**, pp 84.

Konzelmann, T. and R. Weingartner, 2007: *Niederschlagsmessnetze*, Landeshydrol. und Geol., Bern. Plate 2.1.

Lin, Y., S. Chiao, T. Wang, M. Kaplan, and R. Weglarz, 2001: Some common ingredients for heavy orographic rainfall. *Wea. Forecasting*, **16**, 633 – 660.

Lord, S., H. Willoughby, and J. Piotrowicz, 1984: Role of a parameterized ice-phase microphysics in an axisymmetric, nonhydrostatic tropical cyclone model. *J. Atmos. Sci.*, **41**, 2836 – 2848.

Mahoney, K. and G. Lackmann, 2007: The effect of upstream convection on downstream precipitation. *Wea. Forecasting*, **22**, 255 – 277.

Manzato, A., 2005: An odds ratio parameterization for ROC diagram and skill score indices. *Wea. Forecasting*, **20**, 918 – 930.

Martius, O., E. Zenklusen, C. Schwierz, and H. C. Davies, 2006: Episodes of Alpine heavy precipitation with an overlying elongated stratospheric intrusion: A climatology. *Int. J. Climatol.*, **26**, 1149 – 1164.

Marzban, C., 1998: Scalar measures of performance in rare-event situations. *Wea. Forecasting*, **13**, 753 – 763.

Marzban, C. and V. Lakshmanan, 1999: On the uniqueness of Gandin and Murphy's equitable performance measures. *Mon. Wea. Rev.*, **127**, 1134 – 1136.

Mason, I., 1989: Dependence of the critical success index on sample climate and threshold probability. *Aust. Met. Mag.*, **37**, 75 – 81.

— 2003: Binary Events. *Forecast Verification: A Practitioner's Guide in Atmospheric Science*, I. Jolliffe and D. Stephenson, eds., Wiley and Sons Ltd, 37 – 76.

Matthews, R., 1996: Base-rate errors and rain forecasts. *Nature*, **382**, 766 – 766.

McBride, J. and E. Ebert, 1999: Verification of quantitative precipitation forecasts from operational numerical weather prediction models over Australia. *Wea. Forecasting*, **15**, 103–121.

Mesinger, F., 2007: Bias adjusted precipitation threat scores, submitted to Wea. Forecasting.

Mittermaier, M., 2006: Using an intensity-scale technique to assess the added benefit of high-resolution model precipitation forecasts. *Atmos. Sci. Lett.*, **7**, 35 – 42.

Murphy, A., 1993: What is a good forecast - an essay on the nature of goodness in weather forecasting. *Wea. Forecasting*, **8**, 281 – 293.

Murphy, A. and E. Epstein, 1967: A note on probability forecasts and "hedging". *J. Appl. Meteor.*, **6**, 1002 – 1004.

Murphy, A. and R. Winkler, 1987: A general framework for forecast verification. *Mon. Wea. Rev.*, **115**, 1330 – 1338.

Peirce, C., 1884: The numerical measure of the success of predictions. *Science*, **4**, 453–454.

Plaut, G. and E. Simonnet, 2001: Large-scale circulation classification, weather regimes, and local climate over france, the Alps and western Europe. *Climate Reasearch*, **17**, 303 – 324.

Pujol, O., J. Georgis, M. Chong, and F. Roux, 2005: Dynamics and microphysics of orographic precipitation during MAP IOP3. *Quart. J. Roy. Meteor. Soc.*, **131**, 2795 – 2819.

Richard, E., A. Buzzi, and G. Zängl, 2007: Quantitative precipitation forecasting in the Alps: The advances achieved by the Mesoscale Alpine Programme. *Quart. J. Roy. Meteor. Soc.*, **133**, 831 – 846.

Richard, E., S. Cosma, R. Benoit, P. Binder, A. Buzzi, and P. Kaufmann, 2003: Intercomparison of mesoscale meteorological models for precipitation forecasting. *Hydr. Earth Syst. Sci.*, **7**, 799 – 811.

Richter, D., 1995: Ergebnisse methodischer Untersuchungen zur Korrektur des systematischen Messfehlers des Hellmann-Niederschlagsmessers. *Bericht des Deutschen Wetterdienstes*, Deutscher Wetterdienst DWD, volume 194, 93 pp.

Rotunno, R. and R. Houze, 2007: Lessons on orographic precipitation from the Mesoscale Alpine Programme. *Quart. J. Roy. Meteor. Soc.*, **133**, 811 – 830.

Schär, C., T. Davies, C. Frei, H. Wanner, M. Widmann, M. Wild, and H. C. Davies, 1998: Current Alpine climate. chapter 2. *A View from the Alps: Regional Perspectives on Climate Change*, P. Cebon, U. Dahinden, H. C. Davies, D. Imboden, and C. Jäger, eds., MIT press, 21 – 72.

Schmidli, J., C. Schmutz, C. Frei, H. Wanner, and C. Schär, 2002: Mesoscale precipitation variability in the region of the European Alps during the 20th century. *Int. J. Climatol.*, **22**, 1049 – 1074.

Schwarb, M., C. Daly, C. Frei, and C. Schär, 2001: Mean annual and seasonal precipitation in the European Alps 1971-1990. *Hydrological Atlas of Switzerland*, Landeshydrol. und Geol., Bern, plates 2.6, 2.7.

Sevruk, B., 1985: Systematischer Niederschlagsmessfehler in der Schweiz. *Der Niederschlag in der Schweiz*, B. Sevruk, ed., Geol. Schweiz. Hydrol., volume 31, 65–75.

Shepard, D., 1984: Computer mapping: The SYMAP interpolation algorithm. *Spatial Statistics and Models*, G. Gaile and C. Willmott, eds., Dordrecht, 133–145.

Skamarock, W. and J. Klemp, 1992: The stability of time-split numerical-methods for the hydrostatic asnd the nonhydrostatic elastic equations. *Mon. Wea. Rev.*, **120**, 2109 – 2127.

Stauffer, D. and N. Seaman, 1994: Multiscale 4-dimensional data assimilation. *J. Appl. Meteor.*, **33**, 416 – 434.

Stephenson, D., 2000: Use of the "odds ratio" for diagnosing forecast skill. *Wea. Forecasting*, **15**, 221–232.

Steppeler, J., G. Doms, U. Schättler, H. Bitzer, A. Gassmann, U. Damrath, and G. Gregoric, 2003: Meso-gamma scale forecasts using the nonhydrostatic model LM. *Met. Atm. Phys.*, **82**, 75 – 96.

Thornes, J. and D. Stephenson, 2001: How to judge the quality and value of weather forecast products. *Met. Apps.*, **8**, 307 –314.

Tiedtke, M., 1989: A comprehensive mass flux scheme for cumulus parameterization in large-scale models. *Mon. Wea. Rev.*, **117**, 1779 – 1800.

Wernli, H., M. Paulat, M. Hagen, and C. Frei, 2008: SAL - a novel quality measure for the verification of quantitative precipitation forecasts, submitted.

Wicker, L. and W. Skamarock, 2002: Time-splitting methods for elastic models using forward time schemes. *Mon. Wea. Rev.*, **130**, 2088 – 2097.

Widmann, M. and C. Bretherton, 2000: Validation of mesoscale precipitation in the NCEP reanalysis using a new gridcell dataset for the northwestern United States. *J. Climate*, **13**, 1936 – 1950.

Wilks, D., 2006: *Statistical Methods in the Atmospheric Sciences*, Elsevier inc., chapter 3. 23 – 70.

Willmott, C., C. Rowe, and W. Philpot, 1985: Small-scale climate maps - a sensitivity analysis of some common assumptions associated with grid-point interpolation and contouring. *American Cartographer*, **12**, 5 – 16.

Woodcock, F., 1976: Evaluation of Yes-No forecasts for scientific and administrative purposes. *Mon. Wea. Rev.*, **104**, 1209 – 1214.

Zängl, G., 2007: To what extent does increased model resolution improve simulated precipitation fields? A case study of two north-Alpine heavy-rainfall events. *Meteorol. Z.*, **16**, 571 – 580.

Zeng, Z., S. Yuter, R. Houze, and D. Kingsmill, 2001: Microphysics of the rapid development of heavy convective precipitation. *Mon. Wea. Rev.*, **129**, 1882 – 1904.