Meteorol Atmos Phys 99, 155–167 (2008)
DOI 10.1007/s00703-007-0261-8
Printed in The Netherlands

Meteorology and Atmospheric Physics

[1] IAC, ETH Zurich, Zurich, Switzerland and IAU, Johann Wolfgang Goethe-University, Frankfurt a.M., Germany
[2] IMG, University of Vienna, Vienna, Austria

On upscaling of rain-gauge data for evaluating numerical weather forecasts

B. Ahrens[1] and **A. Beck**[2]

With 9 Figures

Received April 4, 2006; revised December 28, 2006; accepted January 30, 2007
Published online January 28, 2008

Summary

One of the main objectives of numerical weather prediction models is reliable forecasting of heavy rain events. This paper discusses problems and strategies of evaluation of daily rain forecasting with operationally available rain station data. The focus is on spatial upscaling of rain station data to the grid of the direct model output. We show limitations of regression – or smoothing – based upscaling like as done, for example, by Kriging analysis and promote probabilistic upscaling by ensembles of stochastic simulations conditioned to the available observations. These ensembles easily provide uncertainties of daily evaluation and unbiased estimates for second moment comparison statistics.

As an evaluation exercise we assess the quality of daily forecasts for Austria (total area: 84,000 km^2) with the limited-area model ALADIN (horizontal grid-spacing 10 km). A quasi–operational set-up is compared to a physically enhanced but less well tested and tuned set-up. It is shown that the evaluation uncertainty is large, but with a full year of forecasts available it is possible to conclude that the physically enhanced set-up simulates too much rain and significantly more than the operational version with only small differences in simulated patterns and variability.

1. Introduction

Nowadays, limited area numerical weather prediction models provide meteorological forecasts with horizontal grid spacing of only a few kilometers and grid spacing will decrease further in the coming years caused by progress in high-performance computing (Schär, 2001; Benoit et al, 2002). High-resolution precipitation forecasts are of primary interest. For example, in flood forecasting systems precipitation detail is a crucial input parameter, especially in mountainous watersheds.

Precipitation forecasts have to be evaluated and errors have to be quantified by comparison with meteorological observations. In the evaluation several decisions have to be made. First, it has to be decided what is to be evaluated: (a) direct output of the numerical weather prediction model, (b) data post-processed by some statistical adaptation like perfect prog, model output statistics, or Kalman filtering, or (c) the end product delivered to the end user after rating and eventual modification by a human forecaster? Here, direct model output shall be evaluated. Thus simulated precipitation fields will be considered in evaluation with values given for grid elements with several kilometers in diameter defined by the models numerical grid.

The second decision concerns the selection of a set of appropriate statistics for quantification of the comparison. This shall not be the issue of this paper. The interested reader is referred to, for example, Murphy and Winkler (1987), Wilks

(1995) and Wilson (2001). For the evaluation exercise presented here, we apply a small set of simple continuous statistics.

Our focus is on the third important decision: Which observational reference is appropriate? Rain station data is commonly preferred to remote sensing data, in particular radar data, because of the large observational uncertainties associated with precipitation products derived from remote sensing data (e.g., Young et al, 1999; Ciach et al, 2000; Adler et al, 2001).

Often done in an operational framework are comparisons of precipitation forecasts valid for grid areas with several kilometers in diameter (i.e., of millions of square meters) directly with rain station data. Each station measures precipitation amount at comparably small, point-like areas of less than one square meter only. Such a comparison can be implemented by simple means, but this area-to-point evaluation is criticized and it is proposed to perform some upscaling of the station data up to the forecast grid resolution (Tustison et al, 2001; Cherubini et al, 2002; Ahrens, 2007). Upscaling, that is data gridding by interpolation and change of support by averaging, allows area-to-area evaluation. Alternatively, downscaling of direct model output to the station sites could be applied and point-to-point evaluation could be performed. This paper prefers upscaling since (a) downscaling, which is basically the post-processing step in the forecast chain, adds uncertainty that is not attributable either to the model or to the downscaling step in evaluation, and (b) upscaling also adds uncertainty but less than downscaling as long as station density is sufficient and additionally improves spatial coverage of the evaluation. The upscaling uncertainty and its impact on evaluation uncertainty is the main issue of this paper.

Upscaling of station data is typically done by some smoothing technique like inverse distance weighted based interpolation (e.g., the monitoring product of the Global Precipitation Climatology Centre (cf. http://gpcc.dwd.de) or the Alpine analysis by Frei and Schär, 1998) or Kriging based methods (e.g., Creutin and Obled, 1982; Rubel and Hantel, 2001). These smoothing based fields are named analyses in the following. For example, a recent analysis of precipitation for the European Alpes by Frei and Häller (2001) has a time resolution of 24 h and a spatial grid-resolution of about 25 km with regionally even lower effective resolution depending on the available surface station network (i.e., accuracy at grid-scale is regionally reduced and should be improved by spatial averaging of the analysis). This type of analysis has been successfully applied in evaluation at the 100 km-scale (see, e.g., Ahrens et al, 1998; Ferretti et al, 2000; Frei et al, 2003). At higher-resolutions the analysis uncertainties increase and the impact of these uncertainties on evaluation have to be dealt with.

The smoothing characteristics of typical upscaling approaches are an additional challenge. Smoothing deteriorates applicability in comparisons with higher-moment statistics. Higher-moment comparison statistics are nonlinear functionals on the spatial fields and their estimates are biased in case of spatial fields with under- or overestimated spatial variability (Aldworth and Cressie, 2003).

An alternative upscaling approach is based on stochastic simulation of an ensemble of precipitation fields with conditioning on the available station data. The idea is to simulate stochastically field realizations that ''honor'' the observed data, their point values, their areal mean, and their covariance structure (Journel, 1974; Chilès, 1999). Therefore, the spatial variability is represented more realistically in stochastic realizations of precipitation fields than in the analysis. Then the forecast can be compared with an ensemble of simulated fields and an ensemble of values is generated for the statistical parameters considered. The ensemble mean field is an analysis (and thus smoother than any ensemble member) and, if first-moment statistics is used, the comparison with the ensemble mean field yields the same comparison accuracy as the mean of the ensemble of statistics values. Additionally, the spread in the ensemble of statistics values provides a precision measure without troublesome estimation and interpretation of the analysis variance. However, the mean of higher-moment comparison statistics is not the same as the biased estimate from forecast evaluation with a smoothing analysis.

This paper compares spatial upscaling of rain-gauge data by Kriging analysis with upscaling by stochastic simulation in evaluation of daily precipitation forecasts by the limited-area model ALADIN in two set-ups with 10 km grid spacing.

The ALADIN model, the evaluation period and type of forecasts, as well as the available station data are introduced in the next section. Section 3 discusses the applied evaluation approaches and subsequent sections discuss the respective results. Finally, some concluding remarks will be given.

2. Precipitation data

For illustrational purposes we investigate the year 1999 in Austria. The considered time resolution is daily, the spatial pixel support is about $10 \times 10\,\mathrm{km}^2$, and the extent of the evaluation area is $84{,}000\,\mathrm{km}^2$.

Observational rain data is available from two sets of rain station data. The first set is a dense network of about 900 stations as provided by the Hydrographisches Zentralbüro, Vienna (delivery date: Feb. 2005) with daily resolution. This set is named HZB in the following. The second data set is provided by the Austrian national weather agency ZAMG consisting of about 120 stations

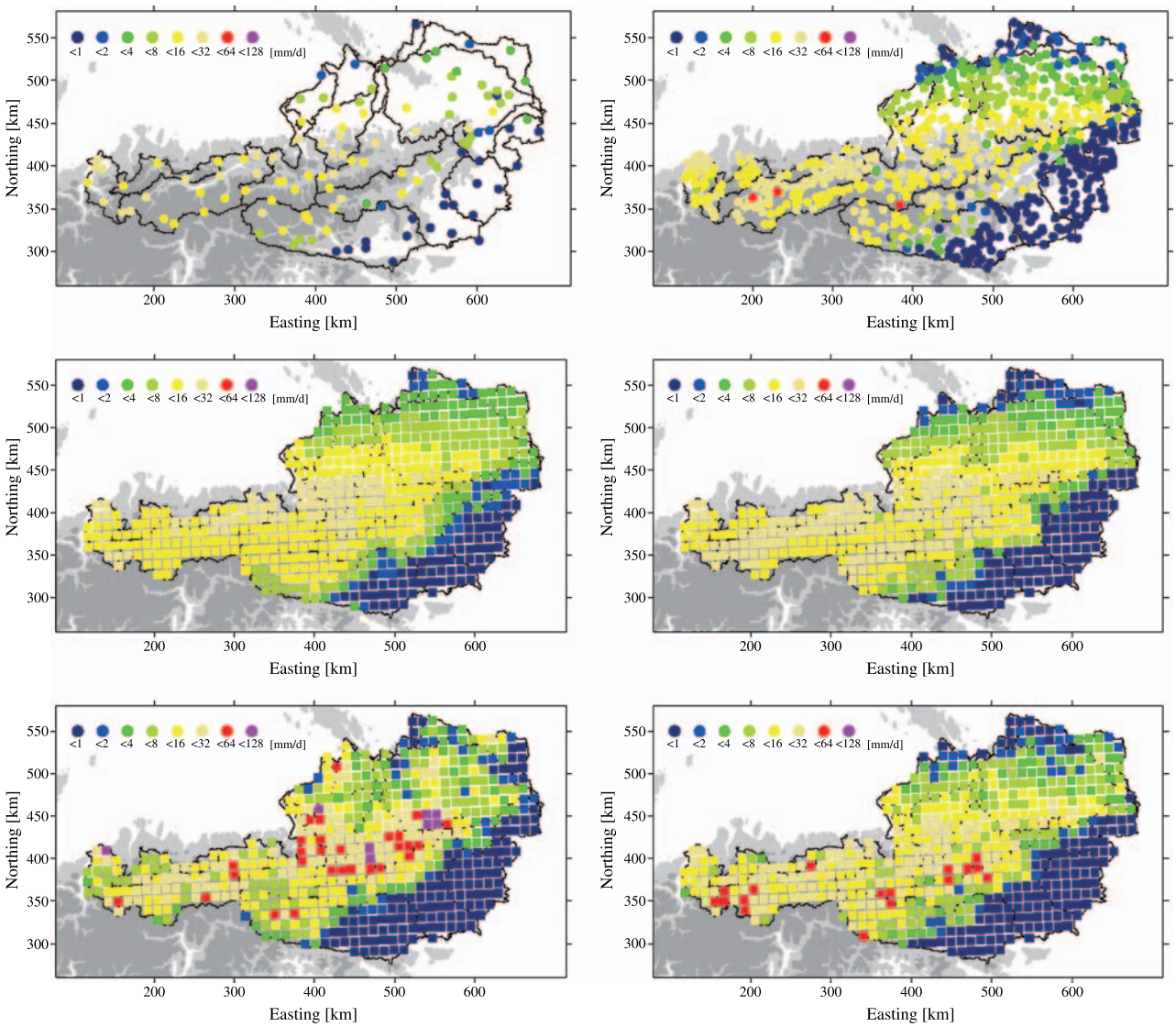

Fig. 1. Precipitation as observed by rain gauges (top row) and upscaled by Kriging analysis (middle row) or by condition stochastic simulations (bottom row) for 19 August 1999 in Austria. The left column illustrates information as provided by the TAWES station set and the right column as provided by the HZB set. The colored bullets in the top row show station positions and station observations. The colored boxes in the middle and bottom row show precipitation values calculated as representative values for the pixels of the ALADIN grid. The orography is indicated by grey shading (light-grey: elevations above 800 m MSL, and dark-grey: elevations above 1500 m MSL). The main Austrian watersheds are indicated by black isolines

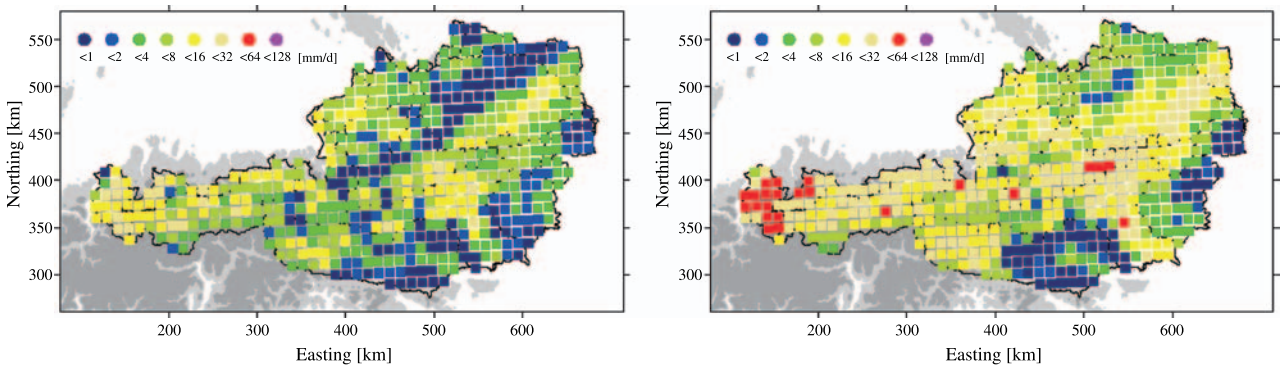

Fig. 2. ALADIN forecasts with set-ups A1 (left panel) and A2 (right panel) for 19 August 1999 in Austria

with 10-min resolution. This set, named TAWES, is independent from the HZB set and generated by automatic weather stations and available in near real time. Within this paper the daily time scale is applied. Thus, the TAWES data is accumulated to daily values. The spatial distributions of the two station sets are illustrated in Fig. 1 (top panels).

This paper discusses the challenge of rain-gauge data upscaling in evaluation of forecast precipitation fields. Here, as an example, these forecast fields are simulated by the limited-area numerical weather prediction (NWP) model ALADIN (Aire Limitée Adaptation Dynamique dévelopement InterNational, see, e.g., Bubnova et al (1995); Ahrens et al (2003) and in the World Wide Web at http://www.cnrm.meteo.fr/aladin/). Here, ALADIN (version 25) is applied in two slightly different set-ups. One set-up is close to the set-up that is applied, for example, by the Austrian national weather service, but the initial and lateral boundary conditions for the limited-area model are derived from ECMWF ERA40 data (Uppala et al, 2005). The precipitation fields are derived from 30 h forecasts initialized daily at 00 UTC and discarding the leading 6 hours to account for model spin-up. The numerical horizontal grid-spacing is about 10 km and thus the precipitation values are given for $10 \times 10\,\mathrm{km}^2$ blocks. This first ALADIN set-up, named A1 in the following, is described in more detail in Beck et al (2004).

A second set-up, named A2, is applied here for discussing the impact of the upscaling strategy on comparative evaluation. Set-up A2 differs from A1 in two changes: (a) daily initialization of the atmospheric fields only and surface parameters that evolve freely in the year-long simulation besides a small relaxation against the ERA40 surface, and (b) application of a more sophisticated radiation parameterization based on Morcrette (1991) that is more expensive in computational resources than the default scheme following Geleyn and Hollingsworth (1979). This second set-up is principally advantageous and motivated by our goal to apply ALADIN in climate research, but it is never applied operationally and thus lacks the fine-tuning that has been done for the operational set-up. Figure 2 illustrates that the precipitation forecasts are sensitive to the changes between the set-ups.

In applications the direct model output should not be applied and some smoothing of the direct output is recommended (e.g., Grasso, 2000; Ahrens, 2003b). Here, the goal is the evaluation of changes in the model set-up and thus the evaluation of the direct model output. Therefore, the scale of comparison between precipitation forecasts and observation is the 10 km-scale.

3. Evaluation method

As motivated in the introduction the applied evaluation method is comparison of direct model output fields against upscaled rain-gauge observations. The upscaled fields are areally averaged onto the model grid. Therefore, forecast and reference fields are prepared for the same grid and an area-to-area comparison of grid elements respecting the grid scales can be performed. This is a substantial advantage over evaluation against station data (i.e., area-to-point comparison). The second potential advantage of upscaling is that station representativeness problems (clustering of

stations around larger cities or along valleys) can be compensated.

Here, we call upscaling involving data-fitting techniques (such as regression, polynomial and spline fitting, Kriging, etc.) an analysis and the estimated field is an analysis field. A common problem of most analysis schemes is error estimation (e.g., Kriging variances underestimate the analysis error in case of precipitation since the Kriging assumptions are not properly fulfilled). A second problem is that analysis fields are expected to underestimate the true field variance (e.g., the smoothing relationship of Kriging states that the analysis variance at any location is the data variance minus the Kriging variance). These problems have to be taken into account in an evaluation based on analyses.

The details of the analysis scheme are of minor importance here and as an example method the ordinary block Kriging with spherical variogram model is applied. Kriging variants are often proposed and applied in precipitation analysis (Creutin and Obled, 1982; Atkinson and Lloyd, 1998; Goovaerts, 1999; Beck and Ahrens, 2004). For the necessary variogram estimation we applied a sub-optimal but robust approach. From the daily data of the year 1999 we estimated from standardized observations a climatological variogram range to about 40 km with a sill of 1 $(\mathrm{mm/d})^2$ (by construction). For daily analyses the sill is rescaled with the observed data variance. Chosen averaging blocks are 10 km in diameter and thus pixel support of the analysis is $10 \times 10\,\mathrm{km}^2$ like of the NWP model forecasts. In case of HZB data analysis a local neighborhood of 64 stations and in case of TAWES data of 8 stations is considered in pixel interpolation. The mean station inter-distance is about 7 and 25 km for the HZB and TAWES data set, respectively. Therefore, in both cases stations in a pixel neighborhood of about $2500\,\mathrm{km}^2$ are considered in interpolation (of course, with decreasing influence with increasing distance). This illustrates the smoothing characteristics of Kriging.

Figure 1 shows the block Kriging results with pixel support of 10 km using TAWES (second row, left panel) or HZB (right panel) data for one single day. The analysis based on the denser HZB data set shows more variability than the analysis based on the coarser TAWES set. This is consistent with the smoothing relationship of Kriging. Figure 2 displays the quite differing forecasts with the two model set-ups. Both forecasts show larger variability than the analyses. In the following we discuss the quantification of this subjective and preliminary conclusions.

Another upscaling approach is stochastic simulation. There are several unconditional and conditional simulation methods for precipitation described in the literature (e.g., Waymire et al, 1984; Ahrens, 2003a), but there is a lack of appropriate methods for stochastic simulation conditioned on available station data. Here, conditioned sequential Gaussian simulation (e.g., Johnson, 1987; Chilès, 1999, chap. 7) is applied as implemented in the geostatistical software package *gstat* (Pebesma, 2004, and *www.gstat.org* in the World Wide Web). Sequential simulation involves the generation of a Gaussian random field, conditioned to the observed data, that honors the variogram of the random field. This conditioned simulation has been done already in Ahrens (2007) for a single event, but in this paper the simulation technique is improved by approximate normalization of the data by a logarithmic transformation respecting that precipitation is a non-Gaussian, non-negative process and applying variogram estimates for the transformed data based on rescaling of the climatological variogram with an estimated climatological range of about 100 km. Again the data is averaged within 10-km blocks.

Figure 1 (third row) shows one stochastic realization for one single day conditioned to TAWES and one realization conditioned to HZB observations. As expected the stochastic simulations are rougher than the analyses. For each day and data set an ensemble of realizations with one hundred members is generated and applied in the following comparisons. Each ensemble member is less accurate than the Kriging analysis in a squared-error sense by construction, but respects the covariance structure given by the observations. The ensemble mean field converges to a Kriging analysis with increasing number of members and is smoother than any ensemble member and thus underestimates spatial variability as does the Kriging analysis.

Optimal analysis of precipitation fields is an active field of research. We picked ordinary block Kriging as a typical and well established analysis method that is easy to implement. The

same motivation led us to apply conditioned sequential Gaussian simulation. Here, the advantages of upscaling by simulation shall be discussed. Therefore, the remaining deficiencies of the Kriging analysis and stochastic simulation upscaling are not crucial for the presented conclusions. Nevertheless, the applied methods are state-of-the-art for daily precipitation interpolation at high spatial resolution.

4. Statistics

In the following we will discuss the evaluation procedures applying a minimal set of useful statistics. Most important is the daily difference in the mean precipitation fields estimated with bias $= 1/N \sum_{x=1}^{N}(m_x - d_x)$ with the daily model forecast field m_x, the regionalized reference field d_x based on observations, and with the space index $x = 1, \ldots, N$. The relative bias rb is defined by bias$/(1/N \sum_{x=1}^{N} d_x)$. Additional statistics considered are the linear correlation coefficient $r = r(m, d)$ of the spatial fields m and d and the estimated ratio of spatial variances $vr = \mathrm{var}(m)/\mathrm{var}(d)$. Optimal values for the evaluation statistics rb, r, and vr are 0, 100, and 100%, respectively.

If a daily forecast is compared with an analysis, then the evaluation result is one value for each statistics. In case of the comparison with a daily ensemble of stochastically upscaled fields the result is an ensemble of values. Therefore, the ensemble allows easy quantification of the daily comparison uncertainty. In the following also months and a full year of daily forecasts are evaluated. Therefore, there are samples of daily results. These samples represent the variability of the quality of the daily forecasts. If the chosen reference is the daily analysis then it is easy to determine a median evaluation result, for example. But daily precipitation statistics is far from normality and consequently the daily statistics are not expected to be normally distributed. Therefore, it is difficult to generate confidence intervals or perform hypothesis tests. If daily ensembles of stochastic simulations are considered in evaluation then quantification of evaluation uncertainty of a sample of daily results gets even more difficult.

Here, the evaluation uncertainty of daily forecasts for monthly or yearly evaluation periods is quantified with a bootstrap procedure (Efron and Tibshirani, 1993). The idea of bootstrapping is to construct a large number n (here $n = 10000$) of new samples out of the original sample (e.g., the daily rbs by comparison with analysis) by random selection with replacement, to calculate the mean of the statistics (e.g., $\overline{rb}$) for each recycled sample, and yielding a distribution of evaluation means that can be summarized by boxplots, for example. In the comparison with daily reference ensembles two-step recycling is applied: first, one realization for each day is drawn and, second, this sample is resampled. These two steps are repeated n times.

Bootstrapping as described underestimates the distribution width slightly in case of persistence in the daily evaluation time series (Ahrens et al, 1998). Here, it is assumed that the persistence is small but a small underestimation of evaluation uncertainty has to be considered in the interpretation of the results.

5. Evaluation experiments

In case of localized rain events the upscaling of precipitation observations is difficult. Additionally, since the precipitation parameterization in NWP models generally involves several threshold parameters especially light precipitation forecasts are uncertain, but this is not critically interfering with the operational forecasts. Therefore, only days with more than 1 mm/day precipitation on average as observed by the HZB observations are considered wet days in evaluation. The dry days are not evaluated and thus the number of evaluated days is 206.

Table 1. Year 1999 mean results of the evaluation experiments discussed in the text. The values are for the ALADIN set-ups A1 and A2 (given in the format A1/A2). The optimum values for the given statistics mean relative bias, $\overline{rb}$, mean correlation, $\bar{r}$, and mean ratio of spatial variances, $\overline{vr}$, are 0, 100, and 100%, respectively

	$\overline{rb}$ [%]	$\bar{r}$ [%]	$\overline{vr}$ [%]
	TAWES		
ana	50/79	35/35	496/536
stoch	59/91	26/26	295/314
	HZB		
ana	35/62	34/34	317/350
stoch	48/81	31/31	285/312

Table 1 gives the mean relative biases, correlations, and variance ratios. Obviously, the HZB data set observes larger Austrian mean values and thus the overestimation of ALADIN is smaller if compared to HZB products than to TAWES data. Additionally, the analysis product offers larger mean precipitation than the stochastic upscaling method. This is because of a positive bias in the analysis due to the skewness of precipitation (Ahrens, 2006). This bias is reduced by the applied normal score transformation in stochastic upscaling. Application of a normal score transformation in the analysis is difficult since the transformation itself introduces a bias in Kriging analysis (Cressie, 1993, chap. 3.2.2.). In agreement with other studies (Creutin and Obled, 1982; Rubel and Hantel, 2001) we analyzed daily precipitation without transformation.

In either case, the forecasts by ALADIN set-up A2 (79% mean daily overestimation compared to TAWES analysis) are wetter than in set-up A1 (50% mean daily overestimation). Note that $\overline{rb}$ is an arithmetic mean of daily ratios. The 1999 relative bias against the mean TAWES analysis is only 16% and 37% in case of set-up A1 and A2, respectively. These biases have to be put into perspective by noting a potential systematic underestimation of gauge measurements ($\sim$10% in case of rainfall and more than 50% in case of snowfall) because of wind-induced or evaporation loss (Rubel and Hantel, 1999; Yang et al, 2005).

There is a strong day-to-day variation in relative bias *rb* with many days of either over- or underestimation by a factor of two as shown in Fig. 3. The figure compares the daily ALADIN forecasts with the ensembles of stochastically upscaled TAWES observations. The standard deviation of daily mean *rb*s is 120%. As the spread in the ensemble of comparison shows there is a large daily uncertainty in *rb* estimation (the mean of the daily standard deviations of *rb* are about 25%) that is comparable to the signal even for the chosen relatively large evaluation domain of 84000 km^2. Considering both the large day-to-day *rb* variability and the daily comparison spread then the difference between A1 and A2 seems to be small at daily time scales with a tendency of larger forecast amounts by A2 in summer months. If we assume that day-to-day variation and daily evaluation uncertainty are independent (which they are not because of a tendency of large daily evaluation uncertainties in case of large biases) and additionally assume that the *rb*s are normally distributed (which they are not by definition of the statistics), than about 30 evaluation days are necessary to conclude that a relative bias of 50% is significant. Since these

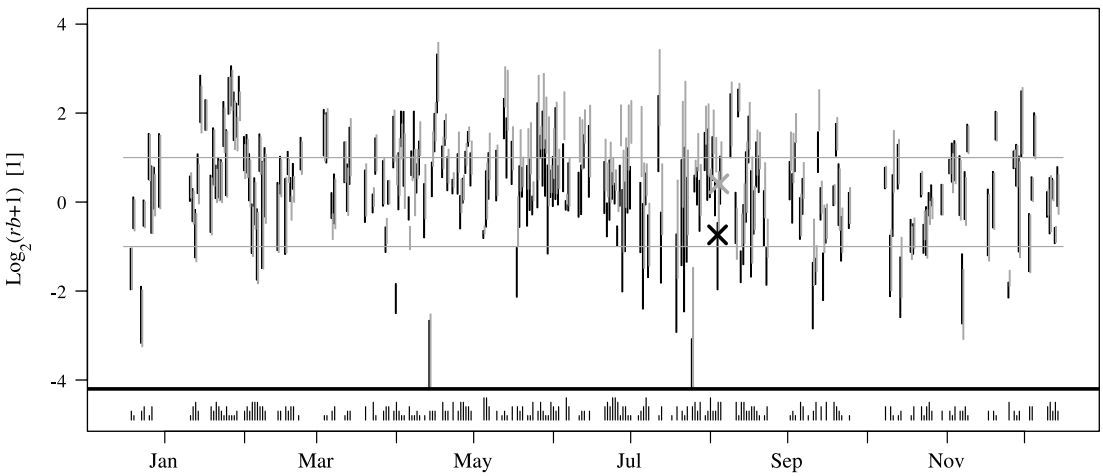

Fig. 3. Time series of daily relative bias range. The vertical lines show the relative bias range *rb* of ALADIN forecasts with set-up A1 (black lines) and set-up A2 (grey lines) compared to daily ensembles of stochastically upscaled TAWES observations. The "x" show the relative biases of the ALADIN forecasts in comparison to the TAWES analysis of the day 19 August 1999. The small bars in the lower part of the figure indicate the observed Austrian mean precipitation by HZB data in $\log_2$-scale. Days with less than 1 mm/day precipitation are not considered and thus without bars and relative bias lines. The horizontal lines indicate forecast over- or underestimation by a factor of two. The perfect value is 0

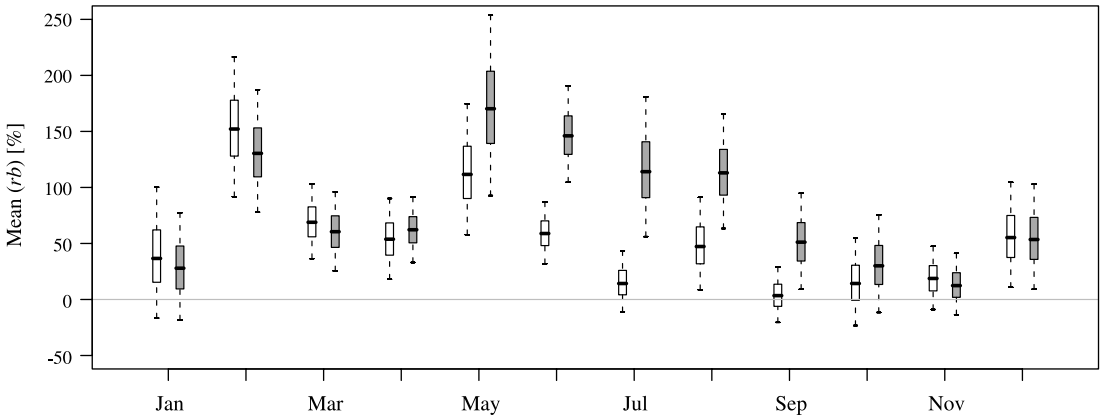

Fig. 4. Monthly boxplots summarizing the distribution of daily *rb*s of ALADIN forecasts with set-up A1 (boxplots filled white) and set-up A2 (boxplots filled grey) compared to daily ensembles of stochastically upscaled TAWES observations. The perfect value is 0%

assumptions are not justified generally more than single months of precipitation forecasts have to be evaluated.

The spread in the HZB ensembles is substantially smaller (not shown, daily standard deviations of *rb* are about 5%), i.e., the observation network density is large enough to constrain the results effectively.

Figure 4 summarizes the daily performance measured with *rb* for monthly periods. The monthly mean biases vary significantly but are significantly positive in almost all months and never significantly negative. If the systematic undercatch of precipitation gauges is taken into account, than the forecasts with set-up A1 are quite promising besides in early summer. Set-up A2 leads to larger overestimation in the summer half-year. In the winter months the difference between the set-ups is small with a small tendency of better performance with A2. Since the set-ups differ in the radiation scheme and initialization of the surface parameterization, the largest differences have been expected for the summer months (Vidale et al, 2003). But, it is disappointing albeit not unexpected that the physically enhanced set-up A2 performs worse than A1. Set-up A1 is well tested and tuned in operational day-to-day use and, therefore, the principal advantage of A2 is more than counter balanced in terms of bias. This illustrates the importance of critical application and tuning of parameters in NWP modeling.

Figure 5 shows boxplots that compare the forecasts of the year 1999 with the TAWES or HZB observations with and without upscaling. All comparisons consistently prove that ALADIN overestimates daily precipitation amounts on av-

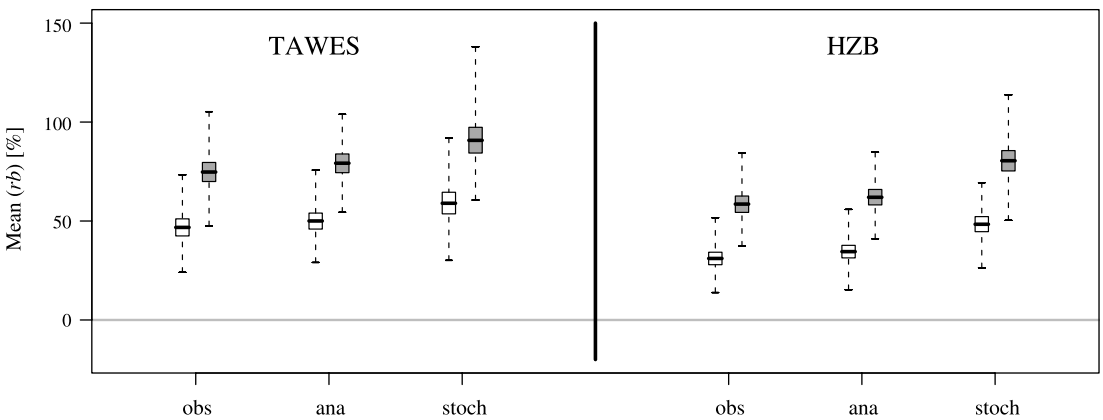

Fig. 5. As Fig. 4 but showing the 1999-mean of daily relative bias and its uncertainty in the comparison experiments with varying references. The references are based on the observational data-sets TAWES or HZB without upscaling (obs), with analysis (ana), and stochastic upscaling (stoch)

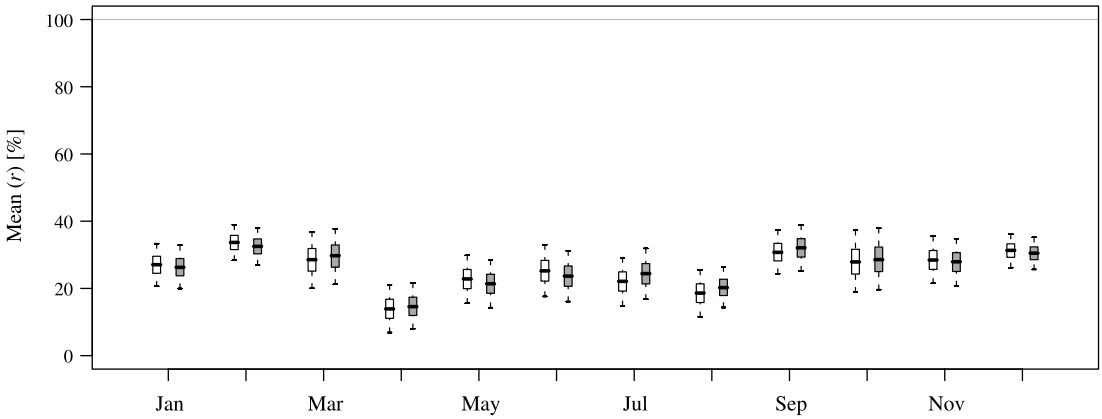

Fig. 6. Monthly boxplots summarizing the distribution of daily correlations r of ALADIN forecasts with set-up A1 (boxplots filled white) and set-up A2 (boxplots filled grey) compared to daily ensembles of stochastically upscaled TAWES observations. The perfect value is 100%

erage and that this overestimation is more pronounced in set-up A2. The larger *rb* values by comparison with stochastically upscaled data have been discussed already. The figure additionally shows that the evaluation uncertainty is slightly larger with stochastic upscaling and with comparing to TAWES instead of HZB products. These uncertainties get larger with decreasing comparison area and period length. For example, discrimination between A1 and A2 is impossible by choosing one or two winter months only. If the comparison extent (spatial and temporal) gets too small then the comparison gets insignificant due to day-to-day variation and daily uncertainty.

As with bias the day-to-day scatter in field correlation is large and evaluation uncertainty even larger (not shown) as expected for a second moment statistics. There are many wet days with insignificant correlation (the significance level of 0.31 is estimated from ensembles of unconditional stochastic simulations with prescribed mean and covariance). Inspecting monthly means of daily correlation shows that A1 and A2 do not differ significantly (Fig. 6). The correlation between stochastically upscaled observations and ALADIN direct model output (DMO) is generally small. This does not mean that the forecast precipitation patterns at larger scales do not compare well with observations since small shifts at the DMO scale drastically decrease correlation measured by r. This effect is called the ''double-penalty effect'' (small location discrepancies of sharp peaks are penalized twice, cf. Anthes, 1983). Not unexpectedly, the smaller correlations occur during the summer months where precipitation fields are generally more heterogeneous than in winter months. Forecasting of summer convection events is a well known problem of NWP modeling.

Figure 6 compares forecasts to stochastically upscaled TAWES observations. The TAWES network is relatively sparse and thus the TAWES simulations are less constrained by observational information than the HZB simulations. This increases the probability for double-penalty and thus of small correlation values in TAWES comparison even in case of good forecasts. Consequently the HZB simulations compare better to the ALADIN forecasts as shown in Fig. 7. As discussed above, the Kriging analysis field is smoother. Therefore, in comparison against anal-

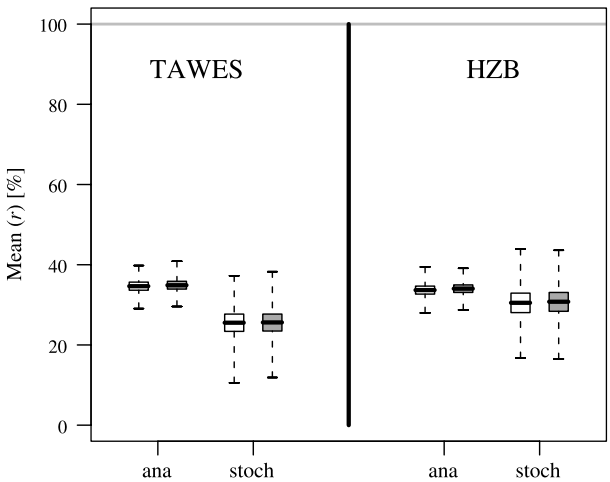

Fig. 7. As Fig. 6, but showing the 1999-mean of daily correlations r and its uncertainty in the comparison experiments with varying references. The references are based on the observational data-sets TAWES or HZB with subsequent analyses (ana) or stochastic upscaling (stoch)

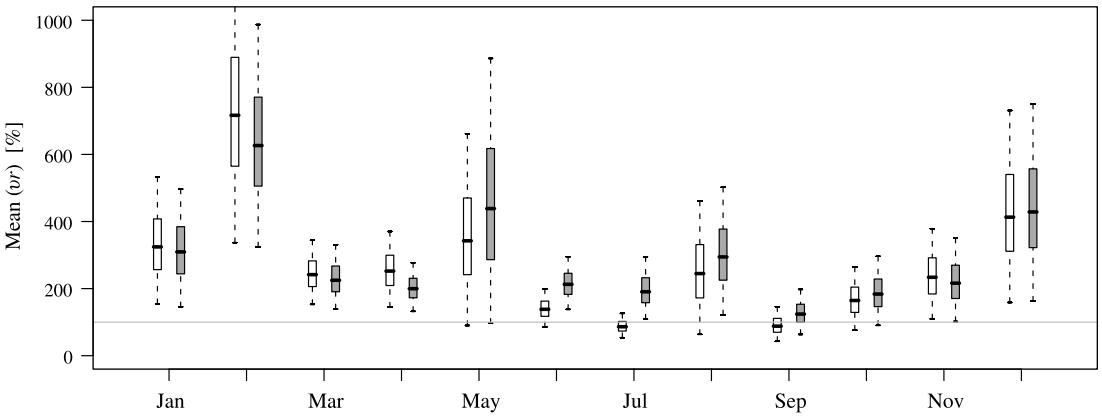

Fig. 8. Monthly boxplots summarizing the distribution of daily variance ratios *vr* of ALADIN forecasts with set-up A1 (boxplots filled white) and set-up A2 (boxplots filled grey) compared to daily ensembles of stochastically upscaled TAWES observations. The perfect value is 100%

yses the probability for double-penalty is smaller than in comparison against stochastic upscaling. This is also shown by Fig. 7. Since the constraining observational network is denser in HZB analyses than in TAWES analyses, the HZB analyses are rougher (cf. Fig. 1, second row) and the mean daily correlation is slightly smaller (Fig. 7). But, in either case the difference in pattern correlation is insignificant between A1 and A2.

The day-to-day variability of the spatial precipitation fields and in consequence of the variance ratio *vr* is large with variance over- and underestimation by ALADIN. Also, the daily evaluation uncertainty is large if compared to stochastically upscaled TAWES fields (not shown). The comparison in monthly periods by Fig. 8 shows that mean daily variability overestimation is smallest in the late summer months, in the months with highest natural heterogeneity due to intense convective rain events. But even monthly means vary substantially with large uncertainties. In view of that the set-ups A1 and A2 do not differ. Somewhat higher variabilities in summer months and smaller variabilities in winter months by A2 can easily be explained by respectively higher and smaller biases in the skewed quantity precipitation.

Figure 9 compares year-long averages of daily variance ratios with analyzed and stochastically upscaled observational reference. The smoothing effect of the analyses can clearly be seen. The mean ratios against either HZB or TAWES based stochastically upscaled fields compare well with a variance overestimation of about 300% by ALADIN. In comparison with the analyses of the

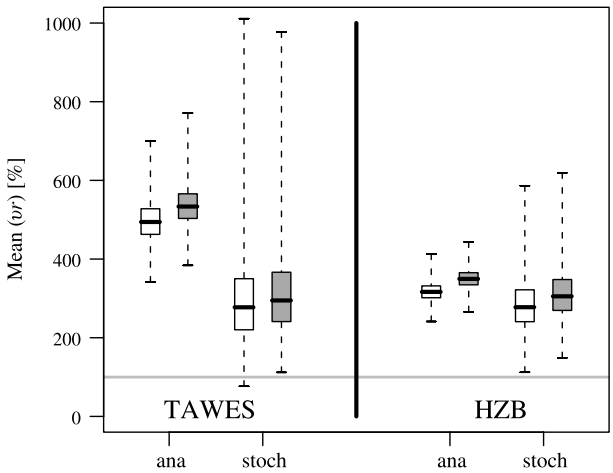

Fig. 9. As Fig. 8, but showing the 1999-mean of daily variance ratios *vr* and its uncertainty in the comparison experiments with varying references. The references are based on the observational data-sets TAWES or HZB with either analysis (ana) or stochastic upscaling (stoch)

dense HZB data the mean *vr* values are slightly larger. This indicates that spatial variance is well represented in the HZB analysis at the comparison scale of 10 km. The TAWES analyses are too smooth yielding misleading variance estimates (roughly by a factor of two in comparison to stochastic upscaling). Additionally, the evaluation uncertainty estimated by the analysis comparison is much smaller than estimated by the stochastic comparison. As a consequence the set-ups seem to differ significantly if compared to HZB analysis (no overlap of the boxes in Fig. 9), but the difference vanishes when compared against the stochastically upscaled observations.

6. Conclusions

Precipitation forecasts by numerical weather prediction or climate model precipitation scenarios have to be evaluated. This paper evaluated direct model output of the NWP model ALADIN in two set-ups with 10 km grid-spacing in Austria (total area: 84,000 km^2) against upscaled rain station data. The upscaling enables an area-to-area evaluation of precipitation fields, which is done by estimation of daily means, spatial variance and correlation patterns of precipitation forecasts and the upscaled observations. Evaluation of second moment statistics like variance or pattern correlation is important if the model output is be applied in, for example, subsequent hydrological forecasting. Hydrological modeling is a nonlinear functional on the precipitation field and therefore the hydrological forecasts are biased if precipitation variance is under- or overestimated (e.g., Ahrens, 2003b, and references therein).

Equivalently, as discussed above, the nonlinear second moment statistics are biased if the spatial variance of the upscaled precipitation variance is unrealistic. Standard precipitation analysis methods are regression based and underestimate spatial variability. The underestimation depends on the density of the available station network. It is shown that the underestimation is substantial for evaluation if the Austrian weather service's operational TAWES network (with mean next station distance of about 25 km one of the densest networks in the world) is applied.

Alternatively, stochastic simulation conditioned to the station observations is applied in upscaling. Instead of one analysis, an ensemble of realizations is considered as the observation based reference in daily evaluation. It is shown that this helps in avoidance of the bias problem and provides an easy method for quantification of daily evaluation uncertainty. The evaluation results using the coarser TAWES observation network in combination with stochastic upscaling are closer to the results using the denser HZB network than using analyzed TAWES fields. Additionally, it is conceptually advantageous to compare with a distribution of reference values as generated by stochastic simulation. This is especially beneficial if forecast ensembles are evaluated. In that case a distribution-to-distribution comparison would be possible and could replace distribution-to-single value comparisons as has to be done by, for example, usual ranked probability skill score (cf. Wilks, 1995).

Besides stochastic simulation other methods are in use for adding spatial variance to upscaled precipitation fields. For example, the orographical pattern can be considered in the analysis (Maurer, 1929; Daly et al, 1994; Smith, 2003), but this is itself based on some regression approach that hinders full variance consideration, or radar data could introduce additional spatial variability, but with the disadvantage that data sets with different measurement quality have to be mixed. In either case this paper shows how important useful variance inflation is in precipitation evaluation. Application of orographical patterns in stochastic simulation are a promising path of further research.

Additional sources of evaluation uncertainty have to be considered, especially in complex terrain like in Austria. Here, only the horizontal representativeness issue is considered. In the mountainous areas the inhomogeneous distribution of stations in the vertical (most stations at valley floors) can lead to systematic errors that are difficult to consider (e.g., Sevruk, 1997). A further systematic error is due to wind and evaporation loss of the rain gauges up to several ten percent of precipitation amount (e.g., Rubel and Hantel, 1999; Yang et al, 2005). These systematic error sources are important if absolute model performance is to be quantified.

This paper quantified the relative performance of an ALADIN set-up that is physically enhanced by a more expensive and conceptually improved radiation parameterization and continuous surface simulation against a set-up applied in operational NWP. In terms of spatial patterns and variability forecasting there are only minor differences between these set-ups if these differences are put into perspective with the evaluation uncertainties. Both set-ups overestimate daily precipitation by more than 35% on average in the conducted evaluation experiments. But the physically enhanced set-up forecasts significantly more daily precipitation in summer months than the operationally applied set-up (twice as much and more). Therefore, we conclude that the operationally well tested and tuned set-up performs better. In our opinion this shows again that a tuned model set-up with harmonically interplay-

ing physical parameterizations is not easily outplayed by improving single model components that risks the harmony of the components (i.e., that risks the careful error balance) of a numerical weather prediction model.

On the specific sources of intensified summer precipitation we presently can only speculate: We assume a large-scale increase of soil moisture and enhanced soil moisture-precipitation feedback (cf. Schär et al, 1999). But this is a problem of further research beyond the focus of the present paper on the challenge of upscaling of precipitation data in model evaluation.

Acknowledgements

Data are provided by the Austrian National Weather Service ZAMG, Vienna, Hydrographische Zentralbüro, BMLFUW, Vienna. Both authors acknowledge financial support through the Austrian Science Foundation FWF under grant P16815.

References

Adler RF, Kidd C, Petty G, Morissey M, Goodman HM (2001) Intercomparison of global precipitation products: The Third Precipitation Intercomparison Project (PIP-3). Bull Amer Meteor Soc 82: 1377–1396

Ahrens B (2003a) Rainfall downscaling in an Alpine watershed applying a multiresolution approach. J Geophys Res 108 (DOI: 10.1029/2001JD001485)

Ahrens B (2003b) Evaluation of precipitation forecasting with the limited area model ALADIN in an Alpine watershed. Meteorol Z 12: 245–255

Ahrens B (2006) Distance in spatial interpolation of daily rain gauge data. Hydrol Earth Sys Sci 10: 197–208

Ahrens B (2008) On evaluation of precipitation fields with rain station data. In: Interfacing Geostatstics and GIS (J. Pilz, ed). Wien: Springer (in print)

Ahrens B, Karstens U, Rockel B, Stuhlmann R (1998) On the validation of the atmospheric model REMO with ISCCP data and precipitation measurements using simple statistics. Meteorol Atmos Phys 68: 127–142

Ahrens B, Jasper K, Gurtz J (2003) On ALADIN precipitation modeling and validation in an Alpine watershed. Ann Geophys 21: 627–637

Aldworth J, Cressie N (2003) Prediction of nonlinear spatial functionals. J Stat Plan Infer 112: 3–41

Anthes RA (1983) Regional models of the atmosphere in middle latitudes. Mon Wea Rev 111: 1306–1335

Atkinson P, Lloyd C (1998) Mapping precipitation in Switzerland with ordinary and indicator Kriging. J Geogr Inform Dec Anal 2: 65–76

Beck A, Ahrens B (2004) Multiresolution evaluation of precipitation forecasts over the European Alps. Meteorol Z 13: 55–62

Beck A, Ahrens B, Stadlbacher K (2004) Impact of nesting strategies on precipitation forecasting in dynamical downscaling of reanalysis data. Geophys Res Lett 31: 5 (DOI: 10.1029/2004GL020115)

Benoit R, Schär C, Binder P, Chamberland S, Davies HC, Desgagné M, Girard C, Keil C, Kouwen N, Lüthi D, Maric D, Müller E, Pellerin P, Schmidli J, Schubiger F, Schwierz C, Sprenger M, Walser A, Willemse S, Yu W, Zala E (2002) The real-time ultrafinescale forecast support during the special observing period of the MAP. Bull Amer Meteor Soc 83: 85–109

Bubnova R, Hello G, Bernard P, Geleyn J-F (1995) Integration of the fully elastic equations cast in the hydrostatic pressure terrain-following coordinate in the framework of the ARPEGE/Aladin NWP system. Mon Wea Rev 123: 515–535

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

Chilès J-P (1999) Geostatistics: modeling spatial uncertainty. New York: Wiley

Ciach G, Morrissey M, Krajewski WF (2000) Conditional bias in radar rainfall estimation. J Appl Meteorol 39: 1941–1946

Cressie N (1993) Statistics for spatial data. NY: Wiley (revised edition)

Creutin J, Obled C (1982) Objective analyses and mapping techniques for rainfall fields: an objective comparison. Water Resour Res 18: 413–431

Daly C, Neilson R, Phillips D (1994) A statistical-topographic model for mapping climatological precipitation over mountainous terrain. J Appl Meteorol 33: 140–158

Efron B, Tibshirani RJ (1993) An introduction to the bootstrap. New York: Chapman & Hall

Ferretti R, Paolucci T, Zheng W, Visconti G, Bonelli P (2000) Analyses of the precipitation pattern in the Alpine region using different cumulus convection parameterizations. J Appl Meteorol 39: 182–200

Frei C, Häller E (2001) Mesoscale precipitation analysis from MAP SOP rain-gauge data. MAP Newsl 15: 257–260

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

Frei C, Christensen J, Déqué M, Jacob D, Vidale P (2003) Daily precipitation statistics in regional climate models: evaluation and intercomparison for the European Alps. J Geophys Res 108: 4124–4142

Geleyn J, Hollingsworth A (1979) An economical analytical method for the computation of the interaction between scattering and line absorption of radiation. Contrib Atmos Phys 52: 1–16

Goovaerts P (1999) Performance comparison of geostatistical algorithms for incorporating elevation into the mapping of precipitation, http://www.geocomputation.org/1999/023/gc_023.htm

Grasso LD (2000) The differentiation between grid spacing and resolution and their application to numerical modeling. Bull Amer Meteor Soc 81: 579–580

Johnson M (1987) Multivariate statistical simulation. New York: Wiley

Journel A (1974) Geostatistics for conditional simulation of ore bodies. Econ Geol 69: 673–687

Maurer J (1929) Die Niederschlagsverteilung im schweizerischen Hochgebirge. Beiträge zur Physik der freien Atmosphäre 15: 102–106

Morcrette J-J (1991) Radiation and cloud radiative properties in the European Center for Medium Range Weather Forecasts Forecasting System. J Geophys Res 96: 9121–9132

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

Pebesma E (2004) Multivariable geostatistics in S: the gstat package. Comp Geosci 30: 683–691

Rubel F, Hantel M (1999) Correction of daily rain gauge measurements in the Baltic Sea drainage basin. Nord Hydrol 30: 191–208

Rubel F, Hantel M (2001) BALTEX 1/6-degree daily precipitation climatology 1996–1998. Meteorol Atmos Phys 77: 155–166

Schär C (2001) Alpine numerical weather prediction 2000–2020: a look back to the future. MAP Newsl 14: 6–12

Schär C, Lüthi D, Beyerle U (1999) The soil-precipitation feedback: a process study with a regional climate model. J Climate 12: 722–741

Sevruk B (1997) Regional dependency of precipitation-altitude relationship in the Swiss Alps. Climatic Change 36: 355–369

Smith R (2003) A linear upslope-time-delay model for orographic precipitation. J Hydrol 282: 2–9

Tustison B, Harris D, Foufoula-Georgiou E (2001) Scale issues in verification of precipitation forecasts. J Geophys Res 106: 11775–11784

Uppala S, Kallberg P, Simmons A, Andrae U, da Costa Bechtold V, Fiorino M, Gibson J, Haseler J, Hernandez A, Kelly G, Li X, Onogi K, Saarinen S, Sokka N, Allan R, Andersson E, Arpe K, Balmaseda M, Beljaars A, van de Berg L, Bidlot J, Bormann N, Caires S, Chevallier F, Dethof A, Dragosavac M, Fisher M, Fuentes M, Hagemann S, Hólm E, Hoskins B, Isaksen L, Janssen P, Jenne R, McNally A, Mahfouf J-F, Morcrette J-J, Rayner N, Saunders R, Simon P, Sterl A, Trenberth K, Untch A, Vasiljevic D, Viterbo P, Woollen J (2005) The ERA-40 reanalysis. Quart J Roy Meteor Soc (submitted)

Vidale P, Lüthi D, Frei C, Seneviratne S, Schär C (2003) Predictability and uncertainty in a regional climate model. J Geophys Res 108 (DOI: 10.1029/2002JD002810)

Waymire E, Gupta V, Rodriguez-Iturbe I (1984) A spectral theory of rainfall intensity at the meso-β scale. Water Resour Res 20: 1453–1465

Wilks D (1995) Statistical methods in the atmospheric sciences. International Geophysics Series, vol. 58. San Diego: Academic Press

Wilson C (2001) Review of current methods and tools for verification of numerical forecasts of precipitation, COST717–Working Group Report WDF_02_200109_1, available at http://www.smhi.se/cost717/

Yang D, Kane D, Zhang Z, Legates D, Goodison B (2005) Bias corrections of long-term (1973–2004) daily precipitation data over the northern regions. Geophys Res Lett 32(19): L19501

Young C, Nelson B, Bradley A, Smith J, Peters-Lidard C, Kruger A, Baeck M (1999) An evaluation of NEXRAD precipitation estimates in complex terrain. J Geophys Res 104: 19691–19703

Corresponding author's address: B. Ahrens, Institut für Atmosphäre und Umwelt, Johann Wolfgang Goethe-Universität Frankfurt, Altenhöferallee 1, 60438 Frankfurt a.M., Germany (E-mail: bodo.ahrens@iau.uni-frankfurt.de)