Hydrology and Earth System Sciences

Dynamically vs. empirically downscaled medium-range precipitation forecasts

G. Bürger

Universität Potsdam, Institut für Geoökologie, Potsdam, Germany

currently at: Pacific Climate Impacts Consortium, Victoria, Canada

Abstract. For three small, mountainous catchments in Germany two medium-range forecast systems are compared that predict precipitation for up to 5 days in advance. One system is composed of the global German weather service (DWD) model, GME, which is dynamically downscaled using the COSMO-EU regional model. The other system is an empirical (expanded) downscaling of the ECMWF model IFS. Forecasts are verified against multi-year daily observations, by applying standard skill scores to events of specified intensity. All event classes are skillfully predicted by the empirical system for up to five days lead time. For the available prediction range of one to two days it is superior to the dynamical system.

1 Introduction

Medium-range prediction of heavy rainfall for flash-flood prone areas such as small mountainous river catchments belongs to the most important challenges of current weather forecasting. Progress in that field is obviously quite beneficial for any affected community, since early warnings in the time frame of several (3–5) days could initiate protection measures and thus avoid much of the damage that is usually brought about by flash floods. Medium-range predictability comes mainly from numerical weather prediction (NWP), where general circulation models (GCMs) simulate the global atmosphere several days into the future. But physical and numerical conditions impose a limit on the spatial resolution of GCMs, rendering their direct output fairly useless for many practical applications. Additional steps are therefore needed to derive small-scale information from GCMs.

This "downscaling"-named procedure exists in two forms, dynamical and empirical, both of which have their advantages and disadvantages. The main advantage of the dynamical approach is the foundation on first principles, which requires only a limited number of additional, empirically derived parameters to represent the unresolved scales. But the complex interplay between model dynamics and topography is difficult to represent physically so that, e.g., positional errors sometimes slip in. This problem is not encountered in empirically based methods as they are directly calibrated against the observed climate, and any potential bias should in principle be removed by the calibration. But to do so requires a considerable amount of parameters that are hard to estimate with sufficient confidence, and that introduces extra errors in the forecasts. But once these parameters are estimated, empirical model forecasts are usually much cheaper numerically.

Numerous comparisons have been conducted between different dynamical downscaling approaches (e.g. Charba et al., 2003; Ebert et al., 2003; McBride and Ebert, 2000; Richard et al., 2007; White et al., 1999). By contrast, the case of empirical downscaling is more diverse because under that term several quite different approaches are summarized. Some of them apply fairly simple statistical recipes to fit the direct model output to observations, such as model output statistics (MOS) or bias correction methods (cf. Wilks, 1995). These approaches are self-evident enough to go without further verification or comparative analysis. Other approaches are somewhat more elaborate – and may actually deserve the name of empirical *model*. Here one defines and calibrates a transfer function between large-scale atmospheric fields and small scale phenomena, such as daily temperature or precipitation. These methods have particularly come to flourish, it seems, with the advent of ensemble prediction for which their numerical simplicity can be fully exploited (Clark and Hay, 2004; Hamill et al., 2006; Liu et al., 2008).

Fig. 1. The three river basins.

Table 1. The three study areas.

Basin	Region	Area (km^2)	Stations
Alb	Black Forest (North)	150	11
Upper Danube	Black Forest (South)	520	22
Upper Iller	Alps (North)	960	53

I am not aware of any systematic comparison between dynamical and empirical methods of NWP downscaling. In this study, daily precipitation forecasts are compared that are made by two coupled systems for three small head catchments in Germany. One system is the Globalmodell (GME) downscaled by the Lokalmodell (LM, now COSMO-EU) of the German weather service (Majewski et al., 2002; Damrath et al., 2000). The other is the Integrated Forecast System (IFS) of the European Centre for Medium Range Weather Forecast (ECMWF), empirically downscaled using expanded downscaling (EDS) (White, 2002; Bürger, 1996). The systems are compared with regard to three intensity classes, by verifying binary forecasts of the corresponding events using standard scores.

2 Data and methods

2.1 The catchments

Atmospheric flow over Germany is westerly dominated, with blocking intermezzos that redirect winds northward or southward. The interplay between this flow and the orography of a catchment leads to typical precipitation characteristics. For example, while Alb and Upper Danube are in close proximity one (Alb) is located west and the other east of the Black Forest, giving them typical luv–lee characteristics with corresponding climate. The Upper Iller, on the other hand, is

located just North of the Alps and receives the greatest precipitation amounts during northerly (blocking) flow. Figure 1 displays the location of the three catchments in Southern Germany. Their main characteristics are summarized in Table 1.

For each basin, average precipitation over all reporting stations will be verified. Although varying availability of data reduces verification performance through time, each forecast system is affected equally so that a fair comparison is possible.

2.2 The GME/LM forecasting system

Unlike most other GCMs, the GME employs a gridpoint approach of a icosahedral–hexagonal type, with an almost uniform mesh size for the entire globe (i.e. without grid convergence at the poles). Until 27 September 2004, that size was ∼60 km with 31 levels, and it changed to ∼40 km and 40 levels afterwards. The model is initialized in a 3 h time interval using a data assimilation scheme that is based on optimum interpolation. Forecasts up to +174 h are issued twice daily at 0:00 and 12:00 UTC, with an additional +48 h forecast issued at 18:00 UTC. Details of the model can be found in (Majewski et al., 2002). The regional model LM is a non-hydrostatic model that operates on 35 levels and a grid spacing of 7 km covering central Europe. When nested in the GME it receives initial and boundary conditions from that model.

For this study, GME/LM forecasts were available from 2002 to 2005, issued daily at 12:00 UTC for lead times of +12 h, +24 h, +36 h, +48 h. The verified quantity was average precipitation of all grid points covering the catchment area.

2.3 The IFS/EDS forecasting system

Unlike the GME the IFS is a spectral model. For the purpose of this study the IFS control forecast from the ensemble prediction system (EPS) was used (the operational high-resolution forecast was not available). Between 1997 and 2005 the EPS was run at gradually increasing resolution. Starting with $T_L 159$ (i.e. a spectral triangular truncation T159 with linear grid, equivalent to a grid spacing of about 120 km at mid latitudes), the system was changed in 2000 to $T_L 255$, equivalent to about 80 km; it operated on 40

levels throughout. (Later in 2006 the system was once more upgraded to $T_L 399$ or ~ 50 km, and 62 levels.) For the subsequent downscaling the following fields were selected from the 850 hPa level:

– geopotential height

– temperature

– vorticity

– specific humidity

And from the surface level

– total precipitation

was included as a fifth predictor field. All fields were interpolated on a 1×1 degree grid, using the rectangular section between the edges (4° W, 46° N) and (18° W, 56° N), which roughly covers the area of central Europe. Concatenation of all fields results in an array of dimension $825 = 5 \times (15 \times 11)$. By applying an empirical orthogonal function (EOF) analysis and retaining only the most dominant EOFs it is possible to reduce this dimension considerably. The reduction should keep as much of the field's fine structure as is necessary to represent, e.g., the major floods of concern, but not so much that one ends up fitting noise. In this case, retaining 81 EOFs was a good compromise (as further discussed in Sect. 4). It should be noted that by using the entire synoptic domain (here central Europe) the downscaling must not be confused with a simple MOS approach.

The study covers the decade from 1997 to 2005. Even for the highest available EPS resolution of about 80 km the size of the Alb basin (150 km^2) is only about 3% of the size of one grid cell (6400 km^2), so the need for downscaling is obvious. The IFS forecasts are issued at 12:00 UTC. For total precipitation (as an accumulating quantity) and a forecast lead time of $+l$ h, $l=0, 12, \ldots, 120$, the overlapping 24 h-sums of $(l+24)$ h$-l$ h were used as predictor (local precipitation is observed in 24 h-sums only).

Suppose the series of daily atmospheric predictors is given as $\boldsymbol{x}(t) = (x_1(t), \ldots, x_n(t))$, with $n=81$. On the other hand, let all station variables be concatenated to form the single vector time series $\boldsymbol{y}(t) = (y_1(t), \ldots, y_m(t))$; in our case, $m=11$. I assume that both series have been transformed to $N(0, 1)$-variates (normal with zero mean and unit variance) using the probit transformation (Ledermann et al., 1984; Bürger, 1996). This will ensure that all scales are weighted adequately by the EDS model, to be described now.

With one exception, the EDS model is just like multiple linear regression (MLR). For both one assumes a model

$$\boldsymbol{y} = \boldsymbol{x} \boldsymbol{Q} + \boldsymbol{\varepsilon} \,, \tag{1}$$

which has MLR as the least squares solution

$$\text{MLR} = \underset{Q}{\text{argmin}} \, \| \boldsymbol{x} \boldsymbol{Q} - \boldsymbol{y} \| \tag{2}$$

($\| \cdot \|$ denoting the Frobenius norm). The problem with MLR is that the simulated amplitudes are scaled by the prevailing canonical correlations between $\boldsymbol{x}$ and $\boldsymbol{y}$, and are thus *damped* relative to observations (Bürger et al., 2006). By imposing on $\boldsymbol{Q}$ the side condition that local covariance be preserved one obtains as a solution the expanded downscaling (EDS) matrix:

$$\text{EDS} = \underset{Q}{\text{argmin}} \, \| \boldsymbol{x} \boldsymbol{Q} - \boldsymbol{y} \| \,, \quad \text{subj.to} \quad \boldsymbol{Q}' \boldsymbol{x}' \boldsymbol{x} \boldsymbol{Q} = \boldsymbol{y}' \boldsymbol{y} \,. \tag{3}$$

Equation (3) describes a so called nonlinear programming problem which is numerically very complex and hard to implement. But recently the following closed-form solution of Eq. (3) was found (Bürger et al., 2009):

$$\text{EDS} = \boldsymbol{G}_x^{-1} \boldsymbol{V} \boldsymbol{U}' \boldsymbol{G}_y \tag{4}$$

Here $\boldsymbol{G}_x$ and $\boldsymbol{G}_y$ denote the Cholesky factors of $\boldsymbol{x}' \boldsymbol{x}$ and $\boldsymbol{y}' \boldsymbol{y}$, respectively, and $\boldsymbol{U}$ and $\boldsymbol{V}$ are from the singular value decomposition

$$\boldsymbol{U} \boldsymbol{\Sigma} \boldsymbol{V}' = \boldsymbol{G}_y \boldsymbol{y}' \boldsymbol{x} \boldsymbol{G}_x^{-1} \,. \tag{5}$$

Accordingly, when driven by global fields that have identical covariance to the calibrating fields of EDS, the simulated local record has covariance *identical* to the observed record. EDS is optimal among all linear maps with this property, by leaving the smallest possible error in Eq. (1). It was originally developed for the downscaling of climate scenarios, with particular emphasis on hydrologic extremes (Bürger, 2002; Menzel et al., 2006).

The model was calibrated on the data from 1997 to 2001 and validated for 2002 to 2005 (as for GME/LM). In slight deviation from a perfect prognosis approach model calibration was *not* done with observed or analyzed fields but instead with the +0 h IFS forecast. This might result in a suboptimal model fit, but it avoids any bias when the EDS is applied to real forecasts. What will further deteriorate the model calibration are the different reference times used for the global and local states. Local precipitation is reported as the 24 h-sum between 06:30 UTC of the reporting day and 06:30 UTC of the following day. The global fields, on the other hand, correspond to 00:00 UTC and 12:00 UTC valid time. Without further adjustment the IFS/EDS forecasts correspond to overlapping 24-sums in 12 h steps. To obtain non-overlapping 12 h-sums the resulting values are simply halved. This produces a "smoothed" 12 h-resolution and represents about the maximum temporal resolution that the EPS allows. Although that is sufficient for an operational early flood warning system, the very details of some of the historic floods probably need an hourly time frame to be adequately simulated hydrologically. This limitation should be kept in mind for the results shown below.
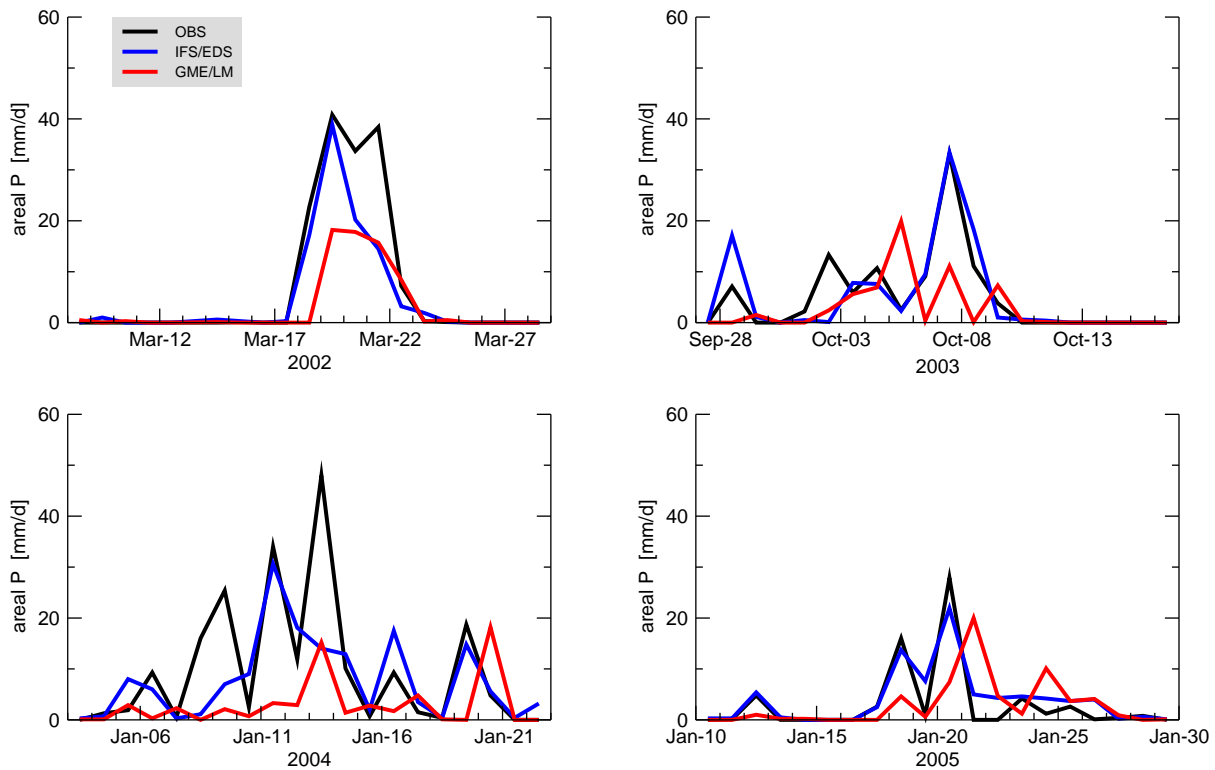
Fig. 2. For the Alb, heaviest observed (black) precipitation event of the respective year, along with the +2 d forecast of IFS/EDS (blue) and GME/LM (red).

3   Results

The two downscaling system, LM and EDS, are not directly verifiable and comparable in this setting since they are driven by different global models. If both of these driver models were verifiable and comparable, then a comparison of LM and EDS could be derived from the coupled systems considered here. Unfortunately, for the GME no evaluation nor any archived driving data exist for Europe in the time frame between 2002 and 2005.

Some indirect evaluation and comparison is nevertheless possible. According to published comparisons it is generally acknowledged that on a global scale, upper air ECMWF forecasts "exhibit smaller errors than DWD-GME forecasts" (http://www.ecmwf.int/products/greenbook). For precipitation, the comparison is more heterogeneous and seems to depend strongly on the investigated region and time, cf. http://www.emc.ncep.noaa.gov/mmb/ylin/pcpverif/scores. For example, some sources report superior performance of the GME at least over Germany (McBride and Ebert, 2000; Ebert et al., 2003) while others suggest the opposite (Bartholmes et al., 2009). In a recent comparison for Europe for the year 2008, the 500 hPa geopotential height predictions of the IFS markedly outperformed those of the GME; but this applied mainly for longer lead times (roughly > +2 d), the short-lead predictions were more similar. Predictions of local pre-

cipitation were less distinguished and more ambivalent, with the Heidke skill score being slightly better and the FBI being slightly worse for the IFS over a wide range of lead times and precipitation intensities (personal communication U. Damrath, DWD). But this comparison is only partially representative for the current study as the GME underwent significant improvements since 2004. Nevertheless, as will become apparent from comparing here the two coupled systems, some conclusions can still be drawn with regard to the downscaling models LM and EDS.

The following verification results are based on daily forecasts for the period 2002–2005, by comparing observed and predicted areal mean precipitation. I will describe in some detail the results for the smallest of the catchments, the Alb, followed by summarizing the forecasts for the other two catchments which are quite similar anyway.

3.1   Alb

As a first impression, Fig. 2 displays observations and +2 d forecast of the heaviest precipitation events of each year between 2002 and 2005. It is evident that IFM/EDS performs better here than GME/LM. In March 2002, there was a sequence of 4 consecutive days with heavy precipitation. The first two of these were captured very well by the IFS/EDS, while the GME/LM merely simulated at most half of the
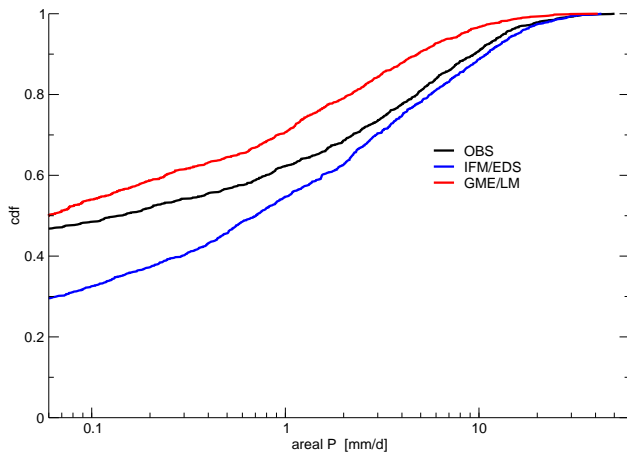
Fig. 3. Cumulative distribution functions of Alb observations (black) and simulations (+2 d) from IFM/EDS (blue) and GME/LM (red).

Table 2. Contingency table for forecasting heavy precipitation with lead +2 d, using IFS/EDS (blue) and GME/LM (red). Upper part: Q_{95}, lower part: Q_{99}.

$Q_{95}=14.6$ mm	EDS, LM$\leq Q_{95}$	EDS, LM$> Q_{95}$	
OBS$\leq Q_{95}$	1319, 1351	53, 21	1372
OBS$> Q_{95}$	33, 53	26, 6	59
	1352, 1404	79, 27	
$Q_{99}=26$ mm	EDS, LM$\leq Q_{99}$	EDS, LM$> Q_{99}$	
OBS$\leq Q_{99}$	1405, 1411	10, 4	1415
OBS$> Q_{99}$	11, 16	5, 0	16
	1416, 1427	15, 4	

scale throughout. In 2003, the difference between the forecasts is even more striking: While IFS/EDS simulates the entire series quite accurately, the GME/LM misses the events almost completely. In 2004 there was an entire series of 4 strong events, and except only in one case (IFS/EDS for the third event) they were forecast quite unsatisfactory by both systems. The 2005 forecasts are similar to those of 2003. Part of the lower predictive skill of GME/LM can be traced back to the reduced variability of that system, as demonstrated by the cumulative distribution function (cdf) of the +2 d prediction of areal precipitation, shown in Fig. 3. Compared to observations, the exceedance probability of the larger scales (on the x-axis) is considerably smaller for the GME/LM, and slightly larger for the IFS/EDS.

To focus on the predictability of strong events it is convenient to consider $Q_{95}=13.4$ mm/d and $Q_{99}=27.1$ mm/d, the upper 5% and 1% quantile, respectively, of the observed areal precipitation. Table 2 shows the contingency table of the corresponding IFM/EDS and GME/LM forecasts for a lead time of +2 d, based on the validation period (1431 days). For Q_{95}, IFM/EDS has more hits (26 vs. 6) and fewer misses (33 vs. 53) than GME/LM, but also more false alarms (53 vs. 21). This is also reflected in the general overprediction of IFM/EDS (79 events) and underprediction of GME/LM (27 events), as compared to observed 59 events. The results for Q_{99} are similar, although the number of predicted events by IFM/EDS now almost equals the number of observed events (16); no Q_{99} event is predicted by GME/LM.

The overall quality of the binary forecasts shown in Table 2 is assessed using the Gilbert skill score (GSS, also called equitable threat score). GSS measures the hit count relative to all cases where an event was observed or forecast, and scales the result in a way that random forecasts receive a zero score (Wilks, 1995). For Q_{95} (Q_{99}) this gives

GSS(IFS/EDS)=0.21 (0.19) and GSS(GME/LM)=0.06 (0.0), showing superior performance by IFS/EDS.

To gain more insight into the predictive power of our system, I have plotted in Fig. 4 the GSS for all lead times up to 5 days, using the usual thresholds of Q_{95} and Q_{99} along with 0.1 mm/d (wet/dry). For all three classes the IFS/EDS forecast shows positive skill up to a lead time of +5 d. The GME/LM forecasts are worse throughout; note that Q_{95} forecasts improve slightly from +1 d to +2 d, which indicates chance behavior in view of the small GSS values; for the Q_{99} class there is no skill beyond a lead time of +1 d. For comparison, I also show the performance of the persistence forecast, which is usually a bad predictor for precipitation due to its short memory; note again the chance behavior especially for Q_{99}.

To further analyze the dependence of forecast skill on the rarity of the event Fig. 5 shows, for a lead time of +2 d, the dependence of forecast skill on the event threshold. I show both the GSS and, as a check for under- or overprediction, the frequency bias index (FBI, the ratio of the number of forecast events to the number of observed events). For both systems the GSS decreases with rarity, but throughout it is about 0.2 larger for the IFS/EDS. With respect to the FBI, the GME/LM tends towards strong underprediction with heavier events, as compared to fairly unbiased predictions of the IFS/EDS for all thresholds.

GSS and FBI are by far not the only scores to assess the quality of a binary forecast, or equivalently, summarizing the entries of a contingency table. If in an experiment the total number of cases is of no interest, there remain 3 independent quantities (degrees of freedom) by which that table is characterized, leaving infinitely many possibilities to combine them into a score. Looking at the table in more economic terms, a very interesting and simplifying approach is provided by the cost-loss model (Thompson, 1952). Namely, once an event is forecast and precautionary measures have been taken, it does not matter – from an economic point of view – whether the event actually occurs or not.
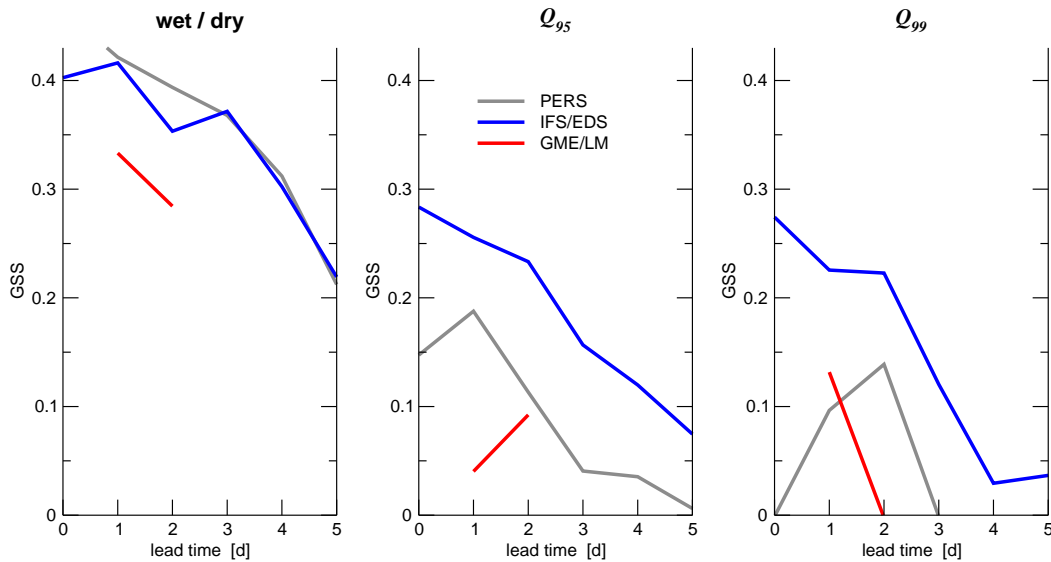
Fig. 4. For the Alb, skill (GSS) of the two forecast systems vs. lead time, using three different event classes. IFS/EDS: blue, GME/LM: red. For comparison, persistence is used as well (gray).
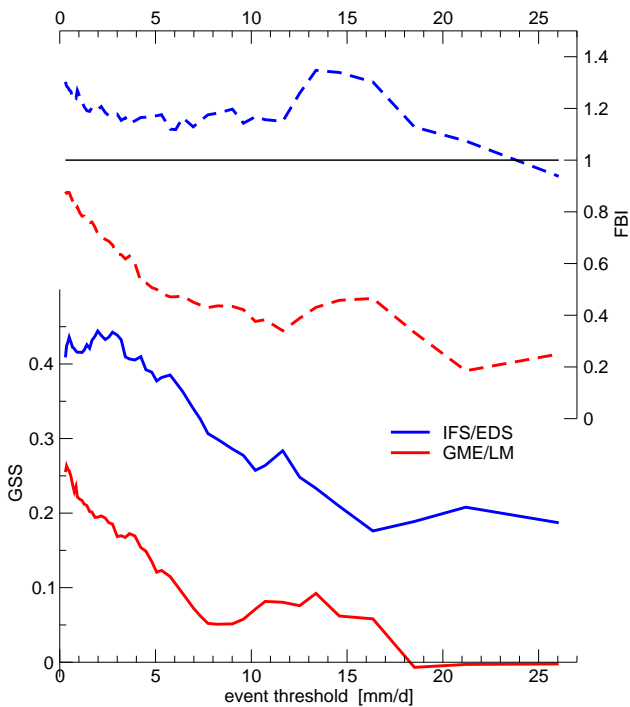

Fig. 5. GSS and FBI dependence on event threshold (rarity), using +2 d predictions for the Alb.

Therefore, only two degrees of freedom remain: the probability of a miss, $P_M = P(O > Q \wedge F \leq Q)$, and the probability of a forecast being issued, $P_F = P(F > Q)$. With the cost of precautionary measures being C and that of a loss incurred

from a miss being L, the expected daily expenses amount to:

$$e = L \cdot P_M + C \cdot P_F . \tag{6}$$

If no forecast is issued, no investment costs are generated but each occurring event is a miss. If the event probability is P_E, the costs to be expected are

$$e_0 = L \cdot P_E . \tag{7}$$

In general, if

$$\frac{P_E - P_M}{P_F} > \alpha , \tag{8}$$

where $\alpha = C/L$ denotes the cost/loss ratio, the expected reduction of losses outweighs the investment from the protection and the forecast has positive economic effects.

For the case described in Table 2, suppose the cost for protection against a rather moderate Q_{95}-event is $C = 10$ k€, and the loss is $L = 100$ k€, then $e_0 = 0.05 * L = 5$ k€. Using the IFS/EDS forecasts one gets a value of about $e = 3.3$ k€, which amounts to 1.7 k€ savings *per day*; GME/LM forecasts yield savings of about 500 €. This is the situation for forecasts of lead +2 d. Figure 6 displays the expected daily expenses for all lead times and both event classes. Considerable savings are to be expected for Q_{95} events when using IFS/EDS forecasts for up to lead +5 d. For Q_{99} events, using $C = 100$ k€ and $L = 1$ M€, the same is true for forecasts of up to lead +3 d. Using GME/LM forecasts gives moderate savings, except for the +2 d forecast of Q_{99} which entails no savings. The two examples above used a cost/loss ratio of $\alpha = 0.1$; according to Eq. (8), positive economic effects can be expected also for smaller values of α.
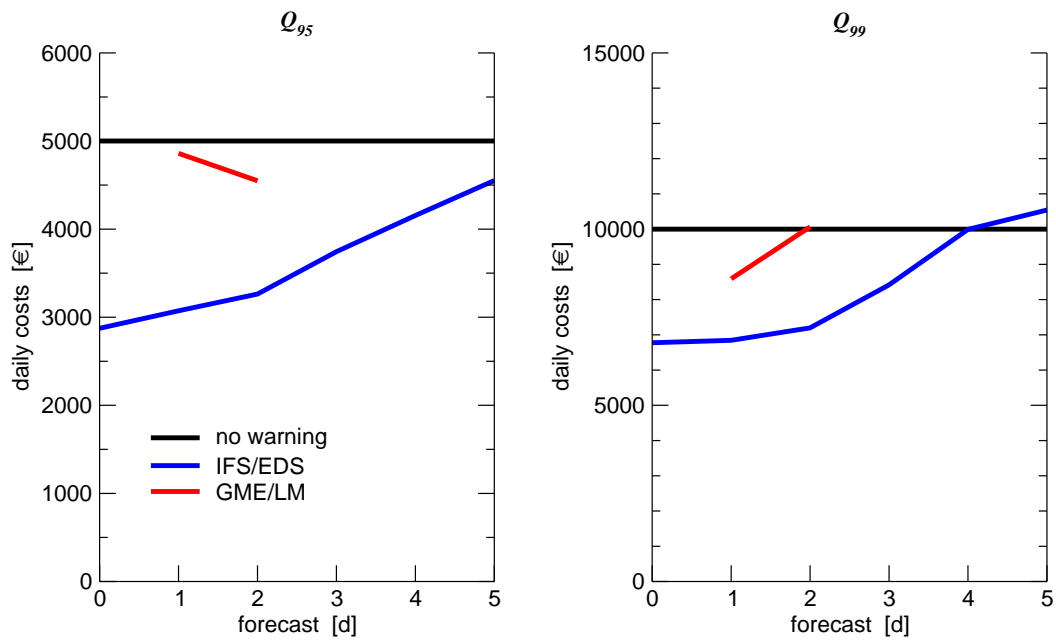
Fig. 6. Expected daily expenses for the Alb, for a Q_{95} and Q_{99} event with no warnings (black) or warnings from IFS/EDS (blue) or GME/LM (red).
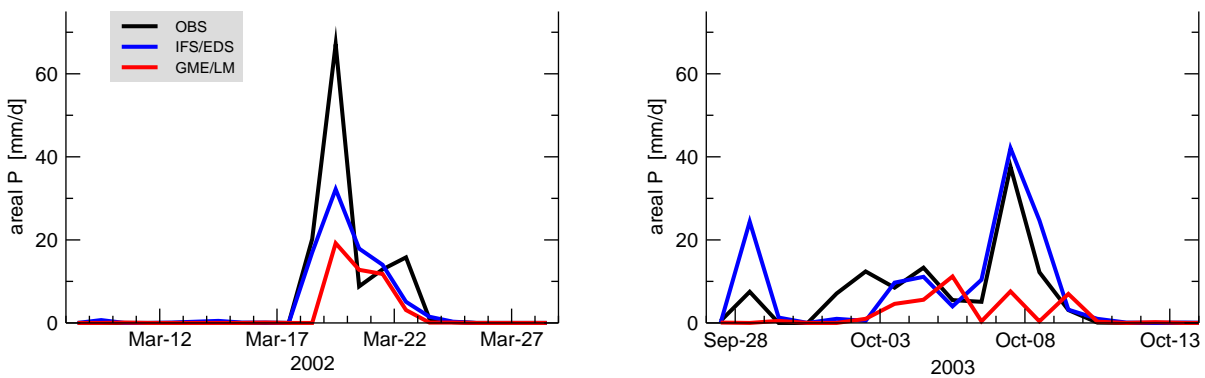

Fig. 7. Similar to Fig. 2, for the Upper Danube. (Note that observations ended in 2003).

3.2 Upper Danube and Upper Iller

Despite different geographical and climatic conditions, the verification results for the Upper Danube and Upper Iller are similar to the Alb. For the Upper Danube, where data ended in 2003, the +2 d forecast of the most extreme yearly events is depicted in Fig. 7. In 2002, the most extreme event was observed on 19 March with 67 mm/d. Here the IFS/EDS forecast (32 mm/d) is only moderately better than that of GME/LM (19 mm/d). In 2003, IFS/EDS forecasts are again superior to GME/LM. Forecasts for the Upper Iller are generally worse than those of the other catchments. This is exemplified by the yearly maxima shown in Fig. 8. Especially the 2004 and 2005 forecasts are bad for both systems. The general superiority of IFS/EDS to GME/LM is apparent from

Fig. 9. It shows that for all event classes and lead times the GSS is comparable to the skill of the Alb shown in Fig. 4. Only the Q_{99} skill for lead time +2 d is exceptionally high for the Upper Danube (GSS=0.54). It is unknown whether this is a random effect (data ended in 2003) or indicative of a real feature.

4 Discussion

In the above analysis I have assessed the capability of the two coupled systems IFS/EDS and GME/LM to forecast precipitation for a small river basin several days in advance. The IFS/EDS was able to skillfully predict medium-sized events (Q_{95}) up to a lead time of 5 days and strong events (Q_{99}) up
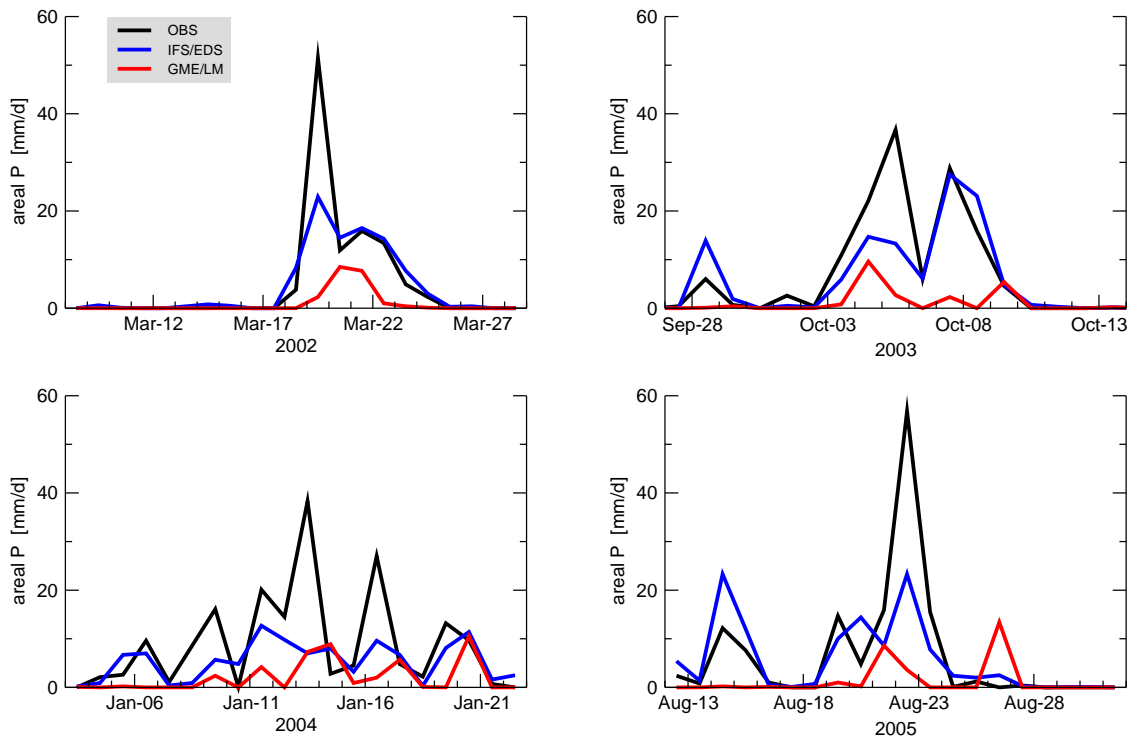
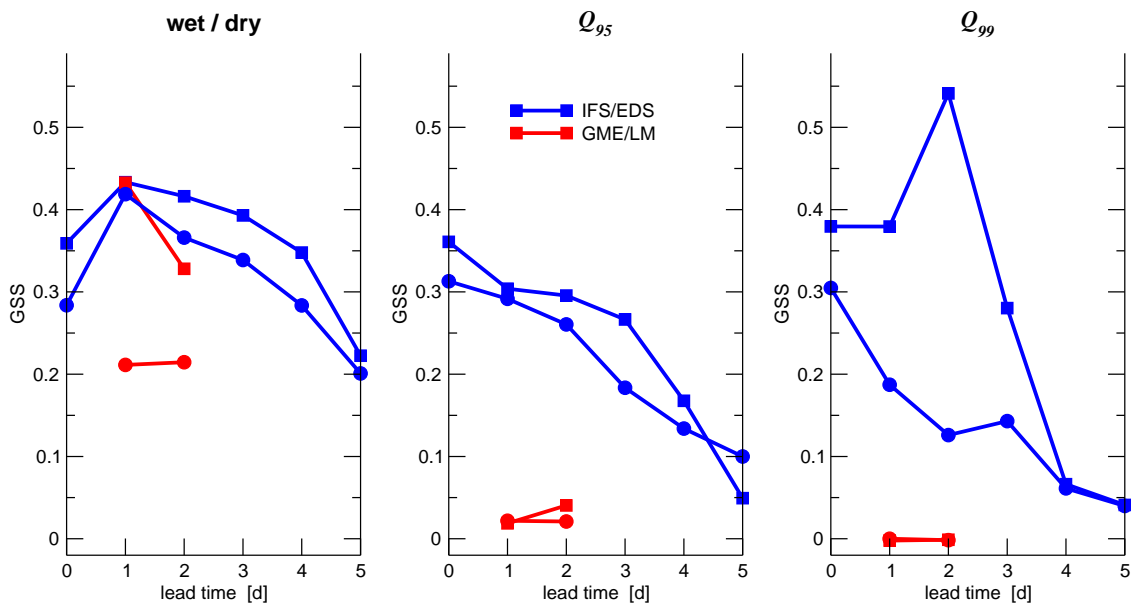**Fig. 8.** Similar to Fig. 2, for the Upper Iller.

Fig. 9. Similar to Fig. 4, the GSS for the other two catchments. Upper Danube (square) and Upper Iller (circle).

to at least 3, maybe 5 days in advance. For all event classes and lead times that system outperformed the GME/LM system. The crucial question is now which of the components makes the difference. But since the systems are so deeply intermingled one feels that deciding that question is hard if

not impossible. For example, while IFS and EDS are coupled through 5 predictor fields LM obtains *all* its boundary conditions from GME. And among those there might always be "bad" fields which are not in the IFS predictor set and which make the difference.

Whatever the reported differences in skill are, they appear marginal in relation to those of, e.g., Fig. 2 or Fig. 5. These +2 d forecasts reveal a gap in skill that can hardly be ascribed to the driving models. This is further supported by the GME/LM failure to reproduce local variability, as evidenced by Fig. 3 and compared to the fairly unbiased local variability of the GME at least in its 2008-version. Therefore, I would ascribe the observed skill differences mainly to the downscaling. It is not unlikely that what we see here is related to the well known luv-lee problem of many high-resolution dynamic models (Baldauf and Schulz, 2004; Elementi et al., 2005), where in mountainous terrain too much rainfall is produced on the luv side and too little on the lee side.

The luv-lee problem illuminates the differences between dynamical and empirical downscaling models mentioned in the introduction: Being genuinely three-dimensional the dynamical models simulate high-resolution precipitation for an entire domain. But the parameterizations of the unresolved scales – here: the advection of falling rain – introduce imperfections that over complex terrain can have a large impact on the water balance. Empirical models, on the other hand, have "seen" the luv–lee characteristic during calibration and "remember" it when confronted with a particular weather type. But a large-scale/small-scale relation like this may as well be more complicated, nonlinear for example, which would then require a revision of the transfer function class and a re-fitting with extra parameters. One should note, however, that if this revision comes *after the fact* independent validation with the same data is no longer possible. Some a priori physical insight is therefore desirable even for empirical models.

A major drawback in the current setup of the IFS/EDS is the determination of the number of EOFs to be retained. Here it was done by simply cross-checking some validation statistics for various lead times, and selecting a number that appeared optimal on average. For Alb, Upper Danube, and Upper Iller this was 81, 79, and 114 EOFs, respectively. Due to data limitations this was done using the entire dataset, so the verification statistics shown above are not fully independent. However, dependence on the number of EOFs was in general fairly weak over a broad range of values, so that the main results are not affected by this choice. This step should nevertheless be improved in future work, for example, by using more elaborate cross validation techniques.

It should be noted that probabilistic versions of the IFS/EDS system exist and have also been applied to the three catchments (cf. OPAQUE, http://brandenburg.geoecology. uni-potsdam.de/projekte/opaque). This was done simply by replacing the deterministic IFS forecast by the ensemble prediction system of the ECMWF (Bürger et al., 2009). In these applications, the use of probabilistic information indeed improves the forecasts, especially for the longer lead times beyond +3 d.

Edited by: A. Gelfan

References

Bürger, G.: Expanded downscaling for generating local weather scenarios, Clim. Res., 7, 111–128, 1996.

Bürger, G.: Selected precipitation scenarios across Europe, J. Hydrol., 262(1–4), 99–110, 2002.

Bürger, G., Fast, I., and Cubasch, U.: Climate reconstruction by regression-32 variations on a theme, Tellus A, 58(1), 227–235, 2006.

Bürger, G., Reusser, D., and Kneis, D.: Early flood warnings from empirical (expanded) downscaling of the full Ensemble Prediction System, Water Resour. Res., doi:10.1029/2009WR007779, in press, 2009.

Baldauf, M. and Schulz, J. P.: Prognostic precipitation in the Lokal-Modell (LM) of DWD, COSMO Newsletter, 4, 177–180, 2004.

Bartholmes, J. C., Thielen, J., Ramos, M. H., and Gentilini, S.: The european flood alert system EFAS – Part 2: Statistical skill assessment of probabilistic and deterministic operational forecasts, Hydrol. Earth Syst. Sci., 13, 141–153, 2009,
http://www.hydrol-earth-syst-sci.net/13/141/2009/.

Charba, J. P., Reynolds, D. W., McDonald, B. E., and Carter, G. M.: Comparative verification of recent quantitative precipitation forecasts in the National Weather Service: A simple approach for scoring forecast accuracy, Weather Forecast., 18(2), 161–183, 2003.

Clark, M. P. and Hay, L. E.: Use of medium-range numerical weather prediction model output to produce forecasts of streamflow, J. Hydrometeorol., 5(1), 15–32, 2004.

Damrath, U., Doms, G., Frühwald, D., Heise, E., Richter, B., and Steppeler, J.: Operational quantitative precipitation forecasting at the German Weather Service, J. Hydrol., 239(1–4), 260–285, 2000.

Ebert, E. E., Damrath, U., Wergen, W., and Baldwin, M. E.: The WGNE assessment of short-term quantitative precipitation forecasts, Bull. Am. Meteorol. Soc., 84(4), 481–492, 2003.

Elementi, M., Marsigli, C., and Paccagnella, T.: High resolution forecast of heavy precipitation with Lokal Modell: analysis of two case studies in the Alpine area, Nat. Hazards Earth Syst. Sci., 5, 593–602, 2005,
http://www.nat-hazards-earth-syst-sci.net/5/593/2005/.

Hamill, T. M., Whitaker, J. S., and Mullen, S. L.: Reforecasts: An Important Dataset for Improving Weather Predictions, Bull. Am. Meteorol. Soc., 87(1), 33–46, 2006.

Ledermann, W., Churchhouse, R. F., and Vajda, S.: Handbook of Applicable Mathematics: Statistics, John Wiley & Sons, 1984.

Liu, X., Coulibaly, P., and Evora, N.: Comparison of data-driven methods for downscaling ensemble weather forecasts, Hydrol. Earth Syst. Sci., 12, 615–624, 2008,
http://www.hydrol-earth-syst-sci.net/12/615/2008/.

Majewski, D., Liermann, D., Prohl, P., Ritter, B., Buchhold, M., Hanisch, T., Paul, G., Wergen, W., and Baumgardner, J.: The op-

erational global icosahedral–hexagonal gridpoint model GME: Description and high-resolution tests, Month. Weath. Rev., 130(2), 319–338, 2002.

McBride, J. L. and Ebert, E. E.: Verification of quantitative precipitation forecasts from operational numerical weather prediction models over Australia, Weather Forecast., 15(1), 103–121, 2000.

Menzel, L., Thieken, A. H., Schwandt, D., and Bürger, G.: Impact of climate change on the regional hydrology – scenario-based modelling studies in the German Rhine catchment, Nat. Haz., 38(1), 45–61, 2006.

Richard, E., Buzzi, A., and Zangl, G.: Quantitative precipitation forecasting in the Alps: The advances achieved by the Mesoscale Alpine Programme, Q. J. R. Meteorol. Soc., 133(625), 831–846, 2007.

Thompson, J. C.: On the operational deficiencies in categorical weather forecasts, Bull. Amer. Meteor. Soc, 33, 223–226, 1952.

White, B. G., Paegle, J., Steenburgh, W. J., Horel, J. D., Swanson, R. T., Cook, L. K., Onton, D. J., and Miles, J. G.: Short-term forecast validation of six models, Weather Forecast., 14(1), 84–108, 1999.

White, P. W.: IFS Documentation, ECMWF, Reading, 2002.

Wilks, D. S.: Statistical methods in the atmospheric sciences, Academic Press, San Diego, 1995.