

# On daily interpolation of precipitation backed with secondary information

S. Krähenmann and B. Ahrens

Institute for Atmospheric and Environmental Sciences, Goethe-University Frankfurt,  
Frankfurt am Main, Germany

Received: 29 December 2009 – Revised: 8 March 2010 – Accepted: 12 March 2010 – Published: 29 March 2010

**Abstract.** This paper investigates the potential impact of secondary information on rainfall mapping applying Ordinary Kriging. Secondary information tested is a natural area indicator, which is a combination of topographic features and weather conditions. Cross validation shows that secondary information only marginally improves the final mapping, indicating that a one-day accumulation time is possibly too short.

## 1 Introduction

Rainfall varies in both space and time. This variability increases with shorter time scales. Hence it is more difficult to interpolate with a limited number of observations on daily than on monthly or annual time scales (Haylock et al., 2008; Yatagai et al., 2009).

Many previous approaches have been made to map precipitation from gauge-observations. Some ignored the spatial covariance structure and knowledge of precipitation processes like orographic effects as Thiessen polygons (TP) and inverse distance weighting (IDW) methods do; several considered spatial covariance structure of precipitation as Kriging does (Goovaerts, 2000; Beck and Ahrens, 2006); Goovaerts pointed out that geo-statistical methods, such as Ordinary Kriging (OK) outperform traditional techniques.

In case an external variable is highly correlated to the studied variable, this correlation can be used to improve the spatial interpolation of the variable of interest. A straightforward method to introduce secondary data is by regression of rainfall versus elevation (Daly et al., 1994; Guan and Wilson, 2005). However, rainfall at a particular grid-node is derived from elevation at this point only, not taking into account surrounding point measurements. In Ordinary Co-Kriging (OCK), spatial correlations between the variable of interest and the external variable are used to modify the kriging equation system (Goovaerts, 2000). This method is highly complex as the covariances of all variables have to be estimated together.

However, while precipitation amount is known to increase with terrain height over larger accumulation times, it generally is weakly correlated to terrain height for short (e.g. daily) accumulation times (Daly, 2002; Goovaerts, 2000). Thus instead of including the external information directly into the interpolation algorithm, this study presents a way of considering external information in a modified way regarding weather conditions and greater topographical features.

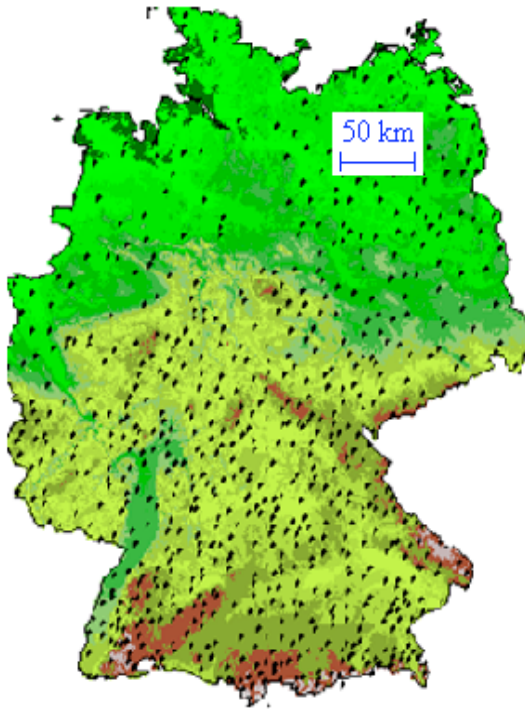
Subsequently daily precipitation data from Germany are interpolated using two techniques: (1) methods that use only daily rainfall data recorded at 759 stations in Germany (TP, IDW, OK); (2) algorithms that combine rainfall data with secondary information (Kriging of Observational Ratios (KOR), and OcK with natural area indicator (NAI) as secondary information). The prediction performance of the different algorithms is compared using cross validation (e.g., Wackernagel, 2003). Stochastic interpolation, which is related to Kriging, but not minimizing the Kriging-variance, is used for evaluating the quality of estimation uncertainty provided by techniques implementing ancillary information.

## 2 Data

Germany consists of a complex topography, ranging from the flat maritime area to the north close to sea level, the hilly low mountain ranges in the middle part with terrain heights of 500 to 1000 m and the alpine area to the south, with an area of approximately 357 000 km<sup>2</sup>. To the south west the Black Forest and Swabian Mountains are most prominent with up to 1500 m, in the south-eastern part close to the border to Austria the elevation reaches almost 3000 m.



Correspondence to: S. Krähenmann  
(kraehenmann@iau.uni-frankfurt.de)



**Figure 1.** Orographical map of Germany with 759 daily operationally measuring rain-gauges indicated by black dots. Low areas green, and high areas brown.

### 2.1 Rain-gauge network

For this study 759 operationally measuring rain-gauges were available, operated by the German Weather Service (DWD), indicated by black dots in Fig. 1. They provide a network of continuous measurements with a fairly homogenous nationwide coverage, and an average distance between neighbouring stations of about 21 km. However, the density of the stations is reduced in higher elevations. Due to the height dependence of rainfall distribution, such biases in the station distribution can lead to systematic errors in the interpolation procedure.

### 2.2 Secondary information

Terrain elevation is known to be highly correlated with climate variables at least over longer accumulation times (Goovaerts, 2000; Daly et al., 2002). Here, a digital elevation model (DEM) is used with a spatial resolution of  $1 \text{ km}^2$  as secondary information (available from US Geological Survey, EROS Data Center, Sioux Falls, SD). The orography is shown in Fig. 1. From the DEM further information is deduced, such as steepness of terrain, aspect ratio and prominence of terrain.

### 2.3 Weather types

The propagation of rain-patterns is highly influenced by the interaction of orography and the wind-field. For example, mountain ridges are able to partially block moisture bearing air, forcing it to ascend. This leads to pronounced precipitation on windward sides, and less rain or even no rain on leeward sides of terrain. Especially in case of low wind speeds blocking is more likely, as the ability of moist air to rise above mountain ridges and propagation of rainfall towards lee-ward sides is limited.

This study distinguishes between nine weather types, characterized by wind direction and wind speed. The four main wind directions (NW, SW, SE, NE) are further split into strong wind and weak wind by a threshold of 10 m/s, the ninth weather type is characterized by no prevailing wind direction.

The horizontal wind direction for each grid point is derived from the wind component data of the 700 hPa level (Bissolli and Dittmann, 2001), distinguishing the four main wind directions. Daily average wind speed, with a spatial resolution of  $2.5^\circ \times 2.5^\circ$ , averaged over Germany, at the 850 hPa level, is noted by <http://www.esrl.noaa.gov/psd/data/gridded/tables/land.html>.

## 3 Methods

This section briefly introduces the different interpolation methods used in this study.

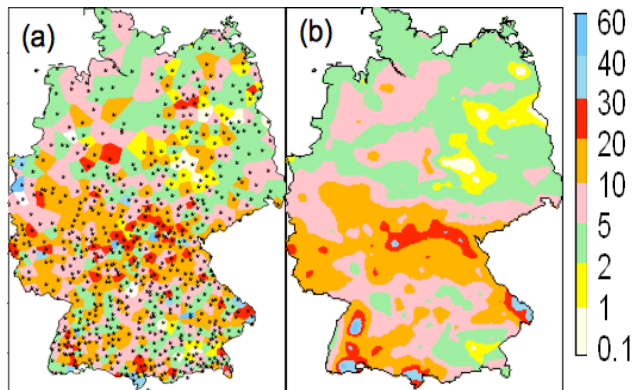
### 3.1 Univariate methods

#### 3.1.1 Thiessen Polygone (TP) Method

This is a simple interpolation method assigning to each grid cell the value of the closest observation and is also called nearest neighbour interpolation (Goovaerts, 2000). An example is given in Fig. 2a, representing daily rainfall on 1.1.2007.

#### 3.1.2 Inverse Distance Weighting (IDW)

To prevent unrealistic artefacts at polygon borders, rainfall can be estimated as a linear combination of surrounding rain-gauge observations, with the weights being inversely proportional to the distance between observations to the power  $p$ . The idea of the weighting system is to put more emphasis on the observations closest to the grid cell to be estimated (Wackernagel, 2003). In this study interpolation is done from 6 surrounding observations and a distance weighting power of  $p = 1.6$  yielding the least RMSE error using cross validation with all stations in Germany for all wet days in 2007, and was held constant over the whole year.



**Figure 2.** Rain maps for Germany, 1.1.2007 (units in mm). Used interpolation methods are (a) Thiessen Polygons (TP) and (b) Ordinary Kriging (OK).

## 3.2 Geostatistical methods without ancillary variables

### 3.2.1 Ordinary Kriging (OK)

OK is a generalized least-square regression technique that allows to account for spatial dependence between observations. Like inverse distance weighting OK estimates the unknown rain amount at grid cells as a linear combination of neighbouring observations. The weights are obtained by minimization of the estimation variance, while ensuring the unbiasedness of the estimator. Instead of Euclidean distance, OK uses a semivariogram as a measure of distance in the observed rain-field. The semivariogram reflects the intuitive feeling that measurements of two rain-gauges close to each other are more alike than those further apart. Like IDW and TP the OK (without nugget effect) reproduces the observations at the station locations.

A spherical climatological variogram model with range of 160 km is chosen for interpolating daily rainfall. Input data firstly are normal score transformed, afterwards all ranges are averaged for days with at least 0.1 mm precipitation per day German wide. This has two reasons, on the one hand this yields the least RMSE in cross validation, on the other hand the stochastic simulation procedure becomes more stable. An example is given in Fig. 2b, representing daily rainfall on 1.1.2007.

## 3.3 Geostatistical methods including secondary information

The incorporation of secondary information potentially improves the estimation of the true rain-field. A straightforward approach is to predict rainfall as a function of the collocated elevation, where elevation data are available at all estimation grid-nodes. The foremost disadvantage of this approach is that the rainfall-amount at a particular grid-node is derived from the elevation only, regardless of the measurements at

surrounding rain-gauges. This approach presumes that the residual values are spatially uncorrelated. A more promising approach is to combine a geostatistical method, which is able to account for spatial correlation of rainfall, with ancillary information. In this study two types are tested: (1) Ordinary Co-Kriging (OCK), and (2) Kriging of Observational Ratios (KOR).

### 3.3.1 Ordinary Co-Kriging (OCK)

The OCK is a multivariate extension of OK. In OCK, spatial correlations between the variable of interest and the secondary variable are used to modify the Kriging system (Goovaerts, 2000). Disadvantageous is its screening of further away elevation data. Furthermore the Co-Kriging system can be unstable in case of inhomogeneous orographical structures, as it is the case in Germany. The main reason for this instability is the much higher correlation between close elevation data, than the correlation between distant rainfall data (Goovaerts, 2000; Wackernagel, 2003). To avoid instability in the subsequent modelling of direct and cross-semivariograms, elevation is computed from the 759 rain-gauge stations only, not the entire DEM. Details of elevation map do not appear in the rainfall-map as elevation is only taken for improving the estimate of spatial variability.

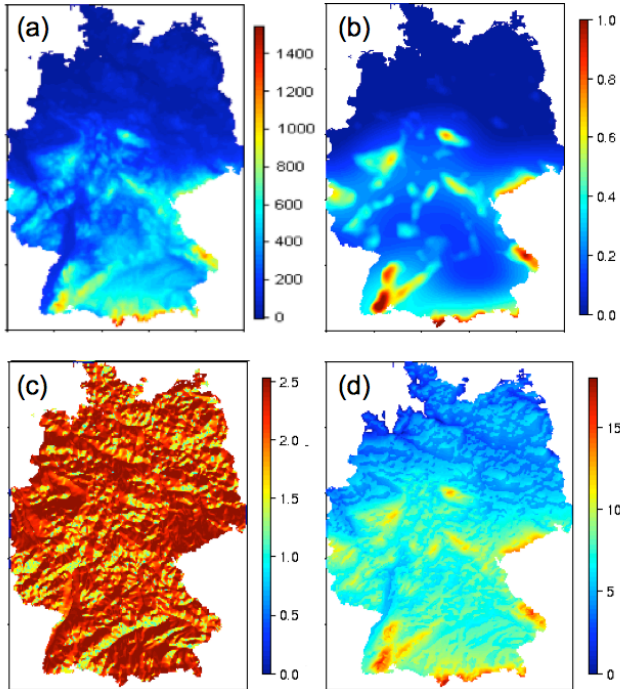
### 3.3.2 Kriging of Observational Ratios (KOR): natural area indicators as secondary information

This is an alternative method to OCK, which decouples the regression and the interpolation part. Here, OK is performed on ratios obtained dividing daily observations by a natural area indicator. Ratios have the advantage of avoiding negative values as well as allowing for non-linear relations between primary and secondary variable.

Primarily an indicator value specific to the prevailing weather condition is calculated for each grid-node, combining orographical parameters deduced from a DEM (elevation, slope, prominence of terrain, and wind facing direction), which is explained in the following. The weather specific indicator value links the orography to the observed rainfall (averaging rainfall for similar conditions). Thereafter the natural area indicator is calculated, such as to maximize the correlation between a single day observation and his respective indicator value specific to the prevailing weather conditions.

### Orographical parameters

To only account for main orographical features, elevation data ( $h$ ) are smoothed using a Gaussian filter with radius 9 km (Fig. 3a), inspired by a study of Smith et al. (2003) which found that the dominant spatial-scale of lifting and rainfall is in the Alps about 10 km.



**Figure 3.** Illustration of terrain features deduced from the DEM: (a) smoothed DEM using Gaussian-filter with radius 5 km [m], (b) prominence of terrain, (c) relative facing-value with wind from NW, wind speed >10 m/s, (d) indicator values [-].

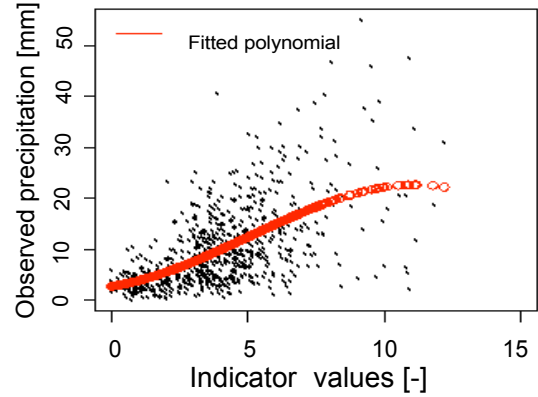
Slope is calculated as the maximum change in elevation over the distance between the cell and its eight neighbours.

The aspect depicts the down-slope direction of the maximum rate of change in elevation from each node to its neighbours, and is useful for distinguishing windward from leeward sides as a function of wind direction, and wind speed. For each single day a facing-value is deduced from the aspect ratio given the mean wind direction, and the wind speed. The relative down-weighting of the area on the leeward side is smaller in case wind-speed exceeds 10 m/s (Fig. 3c). The weights are calculated, such as to maximize the correlation between observed rainfall and facing-value.

The so called prominence (prom) of terrain specifies the effectiveness of terrain to alter the precipitation-field. According the method suggested by Daly et al. (2002) for each grid node the maximum elevation difference to grid nodes within radius 100 km is determined. Grid nodes with elevation differences greater than 500 m are assigned 1, smaller than 100 m 0, or otherwise a value between 0 and 1 (Fig. 3b).

#### Calculation of the indicator

Nine climatological indicator values are determined, one for each weather type. Observations are averaged for days of the



**Figure 4.** Precipitation versus natural area indicator for 1.1.2007. A 3rd order fit, which maximizes the correlation between trend-value and observation is shown (correlation coefficient = 0.577).

same weather type, yielding nine climatological observations (obs.clim) for the  $k$  rain-gauges:

$$\text{obs.clim}_{i,j,k} = \frac{1}{N} \sum_{n=1}^N \text{obs}_{i,j,k} \quad (1)$$

where  $N$  is the number of days of the same weather type,  $i$  the wind direction (NW, SW; SE; NE; no wind), and  $j$  the wind speed (above or below 10 m/s).

The indicator values are calculated combining previously defined orographical parameters:

$$\text{Ind}_{i,j} = \log(\text{slope } h) \text{ prom} + \log(\text{facing}_{i,j} h) \quad (2)$$

such as to maximize the correlation to the respective climatological observations at the  $k$  locations (Fig. 3d):

$$\text{cor}(\text{Ind}_{i,j,k}, \text{mean}(\text{obs.clim}_{i,j,k})) = \max \quad (3)$$

Finally a natural area indicator (NAI) is calculated, which is adjusted to daily rain-gauge observations. A non-linear regression line (3rd order polynomial) defines daily rain-gauge observations as a function of indicator values of the respective weather condition (Fig. 4):

$$\text{obs}_k \sim a + b \text{Ind}_k + c \text{Ind}_k^2 + d \text{Ind}_k^3 \quad (4)$$

yielding parameters  $a$ ,  $b$ ,  $c$ , and  $d$ . These parameters are applied on indicators to calculate the NAIs:

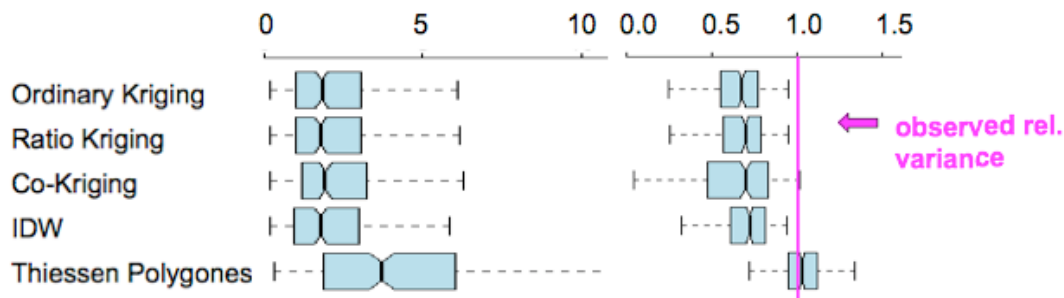
$$\text{NAI} = a + b \text{Ind} + c \text{Ind}^2 + d \text{Ind}^3 \quad (5)$$

This allows for non-linear relations between daily rainfall and indicator value. The rain-gauge observations are divided by the NAI values at the  $k$  locations:

$$\text{ratio}_k = \frac{\text{obs}_k}{\text{NAI}_k} \quad (6)$$

yielding  $k$  ratios, which are normalized before interpolation is done.





**Figure 5.** Comparison of different interpolation methods using cross validation. The box plots indicate RMSE (left), and relative variances (right) found for all days in 2007.

### 3.4 Modelling uncertainty

The probabilistic way to model the uncertainty of a variable at any grid-node consists of viewing the unknown value as the realization of a random variable, and deriving its conditional cumulative distribution function (ccdf). The ccdf fully models the uncertainty at a grid-node, since it gives the probability that the unknown variable is now greater than any given threshold. Under the multiGaussian model, the ccdf at any location is Gaussian and completely characterized by its mean and variance, which corresponds to the Kriging estimate and variance. The approach requires a prior normal score transform of the input data to ensure that at least the univariate distribution is normal (Deutsch and Journel, 1998). The normal score ccdf then is back-transformed to yield the ccdf of the original variable.

Stochastic simulation (SGI, Deutsch and Journel, 1998; Ahrens and Beck, 2008) has been used to assess local uncertainty from the local distribution of simulated values; that is the ccdf at  $u$  is approximated by:

$$F(u; z|(n)) \approx \frac{1}{L} \sum_{l=1}^L i^{(l)}(u; z) \quad (7)$$

where  $i^{(l)}(u; z) = 1$  if  $z^{(l)}(u) \leq z$  and 0 otherwise, with  $z$  the simulated value,  $n$  the neighbouring data, and  $L$  the number of simulations. In theory, as the number of realizations tends to infinity, the local distribution of simulated values should match that provided by Kriging within a similar framework (Goovaerts, 2001).

## 4 Results

The performance of the different interpolation methods is assessed and evaluated using cross validation. The idea behind cross validation is to remove each rain-gauge observation once in turn from the input dataset and to re-estimate the rainfall-amount from the remaining dataset using an interpolation method. The evaluation-criteria are the root mean square of the prediction error (RMSE, perfect value 0), and

the variance of the interpolated rain-field relative to the variance of the input data (perfect value equals 1).

The relative variance reflects the ability of the different interpolation algorithms to maintain the spatial variability within the rain-field. The relative variance is calculated as the ratio of the estimated variability and the true variability given the left-out rain-gauges.

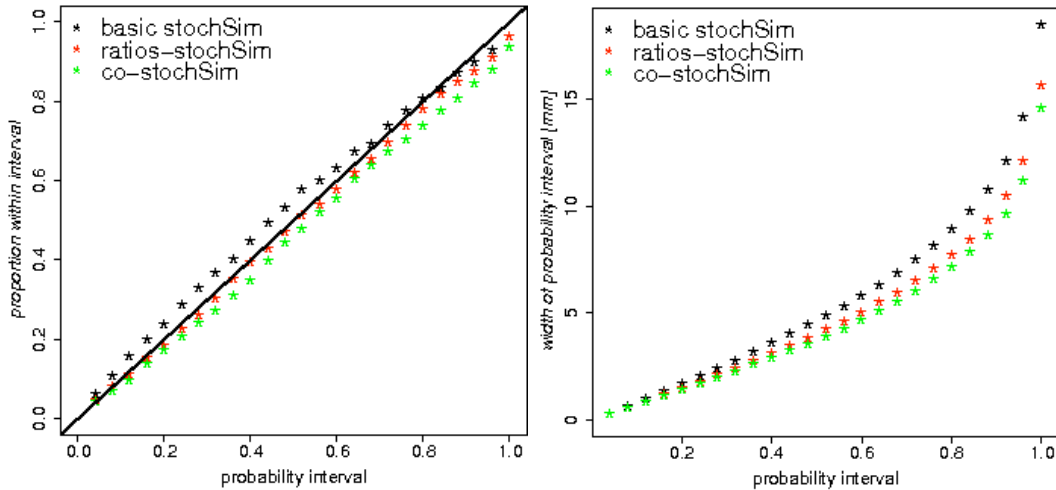
The evaluation is summarized in Fig. 5. The box plots for RMSE indicate the range of interpolation errors for all wet days (at least 0.1 mm station mean) in 2007. The IDW and the three Kriging methods (OK, OCK, KOR) perform equally well, while Thiessen Polygons yields a clearly larger RMSE value.

In terms of relative variance Thiessen Polygons yield the best score. Hence, as the method can not distinguish between close by rain gauges (more alike) and further away rain-gauges (less similar), cross-validation yields a large RMSE, but highlights its strength in maintaining the spatial variability. IDW performs on average slightly better than the three Kriging methods (Fig. 5). However, IDW does not provide a direct uncertainty measure as geostatistical algorithms do (Kriging, SGI).

Of particular interest is the result for OCK in terms of relative variance, the large spread of the box plot suggests that OCK retains the spatial variability in some cases better but many of the cases worse than OK. Overall KOR and OCK do not outperform OK, neither in terms of RMSE nor in terms of relative variance. This indicates that the gained value, if there is any, does not appear in this evaluation exercise and might be a consequence of the generally low correlation coefficient (e.g. 1.1.2007, Fig. 4).

Different methods yield models of uncertainty that can greatly differ, and a legitimate question is whether the choice of a technique (e.g. OK, OCK, KOR) can be supported by the data. Cross-validation can be used to build uncertainty models which are then compared with observations that have been temporarily removed one at a time.

At any test location  $u$ , knowledge of the ccdf  $F(u; z|(n))$  allows the computation of a series of symmetric p-probability intervals (PI) bounded by the  $(1-p)/2$  and  $(1+p)/2$  quantiles



**Figure 6.** Accuracy plot indicating the ability of the different interpolators (OK, OCK, KOR) to estimate the true width of uncertainty measure (left), and width of uncertainty measure for PIs (right) for operational rain-gauges in Germany on 1.1.2007.

of that cdf. For example, the 0.5-PI is bounded by the lower and upper quantiles  $[F^{-1}(u;0.25|(n)), F^{-1}(u;0.75|(n))]$ . A correct modelling of local uncertainty would implicate that there is a 0.5 probability that the actual  $z$ -value at  $u$  falls into that interval or, equivalently, that over the study area, 50% of the 0.5-PI include the true value. If a set of  $z$ -measurements and independently derived cdfs are available at  $N$  locations,  $u_j, \{[z(u_j), F(u_j, z|(n))], j = 1, \dots, N$ , the fraction of true values falling into the symmetric  $p$ -PI can be computed as:

$$\bar{\xi}(p) = \frac{1}{N} \sum_{j=1}^N \xi(u_j; p) \quad \forall p \in [0, 1] \tag{8}$$

with  $\xi(u_j; p) =$

$$\begin{cases} 1 & \text{if } F^{-1}(u_j; (1-p)/2) < z(u_j) \leq F^{-1}(u_j; (1+p)/2) \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

Here, the cdfs are inferred through 100 SGI (OK-, OCK-, KOR- equation).

The scattergram of the estimated versus expected fractions is called ‘‘accuracy plot’’. Figure 6 (left) shows the daily rainfall accuracy plot computed for both OK, OCK, and KOR cdfs using cross-validation of 759 observations at all wet days in 2007. The accuracy plot shows that OK overestimates the uncertainty for PI smaller than 0.6 (black stars above black line), e.g. the 0.5 PI derived from the OK contains 60% of the true values, while KOR, and OCK perform better. For PIs greater than 0.6 OK performs better. Here, COK contains too few true values.

### 5 Conclusions

Secondary information has to be considered in terms of accumulation time and complexity of terrain. Combining OK with a NAI only slightly improves the evaluation results independent of the applied combination approach. Introduction of secondary information with observational ratios yields a small improvement in terms of uncertainty measure, as could be shown with the accuracy plot (Fig. 6, left).

As is shown, the number of true values within interval is closer to the expected number (given by the PI) when incorporating NAI values. Otherwise the width of the probability interval seems too wide indicating overestimation of the true uncertainty. This is not true for probability intervals greater than 0.6. Overall the gain is minimal. The foremost reason for this is the too short accumulation time, as the spatial distribution of precipitation tends to be better defined by topographical/external parameters for longer accumulation times (Goovaerts, 2000). New techniques introducing secondary information should be investigated, e.g. by means of a stratified variogram which is separately inferred from rain-gauges within similar topographical features. Another way might be to determine parameters such as NAI on a physically meaningful basis, or adding further parameters such as humidity or stability (Haiden et al., 2008), instead of determining them statistically. The main challenge is the fact that any simple, static, either topographical or weather-type based index can capture only a small part of the complex processes involved.

**Acknowledgements.** Precipitation data were obtained from the German Weather Service (DWD). The authors also acknowledge funding from the Hessian initiative for development of scientific and economic excellence (LOEWE) at the Biodiversity and Climate Research Centre (BiK-F), Frankfurt/Main.

Edited by: M. Dolinar

Reviewed by: T. Haiden and another anonymous referee

## References

- Ahmed, S. and de Marsily, G.: Comparison of geostatistical methods for estimating transmissivity using data on transmissivity and specific capacity, *Water Resour. Res.*, 23, 1717–1737, 1987.
- Ahrens, B.: Distance in spatial interpolation of daily rain gauge data, *Hydrol. Earth Syst. Sci.*, 10, 197–208, 2006, <http://www.hydrol-earth-syst-sci.net/10/197/2006/>.
- Ahrens, B. and Beck, A.: On upscaling of rain-gauge data for evaluating numerical weather forecasts, *Meteorol. Atmos. Phys.*, 99(3–4), 155–167, 2008.
- Bissolli, P. and Dittmann, E.: The objective weather type classification of the German Weather Service and its possibilities of application to environmental and meteorological investigations, *Meteorol. Z.*, 10(4), 253–260, 2001.
- Daly, C., Neilson, R. P., and Phillips, L. D.: A statistical-topographic model for mapping climatological precipitation over mountainous terrain, *J. Appl. Meteorol.*, 33, 140–158, 1994.
- Daly, C., Gibson, W. P., Taylor, G. H., Johnson, G. L., and Pasteris, P.: A knowledge-based approach to the statistical mapping of climate, *Clim. Res.*, 22, 99–113, 2002.
- Daly, C.: Variable influence of terrain on precipitation patterns: Delineation and Use of Effective Terrain Height in PRISM, Oregon State University, Corvallis, available at: <http://www.prism.oregonstate.edu/pub/prism/docs/effectiveterrain-daly.pdf>, 2002.
- Deutsch, C. V. and Journel, A. G. (Eds.): *GSLIB: Geostatistical Software Library and User's Guide*, Oxford Univ. Press, Oxford, UK, 369 pp., 1998.
- Goovaerts, P.: Geostatistical approaches for incorporation elevation into the spatial interpolation of rainfall, *J. Hydrol.*, 228, 113–129, 2000.
- Goovaerts, P.: Geostatistical modelling of uncertainty in soil science, *Geoderma*, 103, 3–26, 2001.
- Guan, H. and Wilson, J. L.: Geostatistical Mapping of Mountain Precipitation Incorporating Autosearched Effects of Terrain and Climatic Characteristics, *J. Hydrometeorol.*, 6(6), 1018–1031, 2005.
- Haiden, T. and Pistotnik, G.: Intensity-dependent parameterization of elevation effects in precipitation analysis, *Adv. Geosci.*, 20, 33–38, 2009, <http://www.adv-geosci.net/20/33/2009/>.
- Haylock, M. R., Hofstra, N., Klein Tank, A. M. G., Klok, E. J., Jones, P. D., and New, M.: A European daily high-resolution gridded dataset of surface temperature and precipitation, *J. Geophys. Res.*, 113, D20119, doi:10.1029/2008JD010201, 2008.
- Pebesma, E. J.: Multivariable geostatistics in S: the gstat package, *Comput. Geosci.*, 30, 683–691, 2004.
- Prudhomme, C. and Reed, D. W.: Mapping of extreme rainfall in a mountainous region using geostatistical techniques: A case study in Scotland, *Int. J. Climatol.*, 19, 1337–1356, 1998.
- Smith, R. B., Quingfang, J., Fearun, M. G., Tabary, P., Dorninger, M., Doyle, J. D., and Benoit, R.: Orographic precipitation and air mass transformation: An Alpine example, *Q. J. R. Meteorol. Soc.*, 129, 433–454, 2003.
- Wackernagel, H. (Eds.): *Multivariate Geostatistics*, 3rd Edn., Springer, Berlin, 2003.
- Yatagai, A., Arakawa, O., Kamaguchi, K., Kawamoto, H., Nodzu, M. I., and Hamada, A.: A 44-year daily gridded precipitation dataset for Asia based on a dense network of rain gauges, *SOLA: Scientific Online Letters of the Atmosphere*, 5, 137–140, doi:10.2151/sola.2009-035, 2009.