Paper prepared for the 123rd EAAE Seminar

PRICE VOLATILITY AND FARM INCOME STABILISATION

Modelling Outcomes and Assessing Market and Policy Based Responses

Dublin, February 23-24, 2012

Minimizing geographical basis risk of weather derivatives using a multi-site rainfall model

Ritter M.[1,2], Mußhoff O.[1] and Odening M.[2]

1 Department of Agricultural Economics and Rural Development, Georg-August University Göttingen, Göttingen, Germany
2 Department of Agricultural Economics, Humboldt University Berlin, Berlin, Germany

Matthias.Ritter@agrar.hu-berlin.de

Minimizing geographical basis risk of weather derivatives using a multi-site rainfall model

Ritter M., Mußhoff O. and Odening M.

Abstract

Weather risk is one of the main causes for income fluctuation in agriculture. Since 1997, the economic consequences of weather risk can be insured with weather derivatives, which are offered for many different weather events, such as temperature, rainfall, snow or hurricanes. It is well known that the hedging effectiveness of weather derivatives is interfered by the existence of geographical basis risk, i.e., the deviation of weather conditions at different locations. In this paper, we explore how geographical basis risk of rainfall based derivatives can be reduced by regional diversification. Minimizing geographical basis risk requires knowledge of the joint distribution of rainfall at different locations. For that purpose, we estimate a daily multi-site rainfall model from which optimal portfolio weights are derived. We find that this method allows to reduce geographical basis risk more efficiently than simpler approaches as, for example, inverse distance weighting.

Keywords: Risk management, weather risk, regional diversification, portfolio weights

JEL classification: G11, Q14, G32.

1. INTRODUCTION

Since the start of trading weather derivatives at the Chicago Mercantile Exchange in 1999, these new financial instruments have become more and more popular to insure weather risk. Opposite to usual insurances, their payoff does not depend on the actual damage, but on a specific weather event. Examples of weather indices, which are used as an underlying of weather derivatives, are cumulative temperature, heating degree days, cumulative rainfall, as well as the occurrence of snow or hurricanes. All weather derivatives have in common that the underlying weather event has to be reported by an independent weather station. Consequently, the payoff of weather derivatives cannot be influenced by buyers or sellers and thus moral hazard and adverse selection that plague conventional insurance are curbed (cf. Jin et al., 2005). As many sectors are weather sensitive (Lazlo et al., 2011; Nadolnyak and Hartarska, 2012), one would expect that weather derivatives are widely used for insuring weather risk. However, the acceptance of these risk management tools, for example, in agriculture as well as in other sectors, has fallen short of expectations so far. A major obstacle for the use of weather derivatives in practice is the existence of basis risk (Vedenov and Barnett, 2004; Chen and Roberts, 2004; Woodard and Garcia, 2008; Mußhoff et al., 2011), i.e. "the payoff of the derivative does not perfectly correspond to the shortfalls of the underlying exposure" (Berg and Schmitz 2008: 123). The deviation between actual losses and insurance payoff reduces the hedging effectiveness of weather derivatives and lowers the willingness to pay for these instruments. Basis risk can be distinguished in production basis risk and geographical basis risk. The former is caused by a low correlation between the weather variables relevant for the insurance payoff and the economic losses, while the latter arises from the difference between the weather conditions at the reference station of the derivative and the weather conditions at the buyer's location. Production basis risk is specific to a particular business and varies between sectors. Private energy consumption, for example, is closely related to outside temperature and thus the stochastic demand for electricity can be replicated fairly well by temperature indices. Hence, production basis risk is relatively low for electricity companies. In contrast, the agricultural sector is usually exposed to pronounced production basis risk because the functional relation between crop yields and weather conditions is very complex and cannot be captured by simple weather indices.

In this paper, we focus on geographical basis risk, which is unavoidable, also because weather derivatives are not provided for every existing weather station. The severity of geographical basis risk mainly depends on two factors. First, the density of reference weather stations is important. To increase liquidity of the market for weather derivatives, only a selection of standardized contracts is offered. The CME, for example, offers trading of different temperature derivatives for 24 cities in the USA, six in Canada, eleven in Europe, three in Australia and three in Japan. Moreover, it started trading standardized rainfall derivatives for ten selected US cities in 2011. Second, the geographical basis risk depends on the kind of weather variable. While temperature shows a high spatial correlation, this is not the case for precipitation, which is a local phenomenon (Xu et al., 2010). To cope with this problem, we follow Berg and Schmitz (2008) who suppose that geographical basis risk can be reduced by using a portfolio of derivatives for different locations. Identification of an optimal regional portfolio, however, requires the determination of weights for all portfolio components which reflect the correlations between their payoffs. An intuitive approach would be to weight the derivatives according to the inverse distance between the reference weather station and the buyer's location. Unfortunately, there is empirical evidence that the correlation between rainfalls at different locations frequently is not a simple decreasing function of the distance (Odening et al., 2007), which is particularly true for mountainous regions (Salsón and Garcia-Bartual, 2003). This finding suggests a statistical approach for the determination of the portfolio weights. The objective of this paper is to provide a statistical model supporting this task. For that purpose, we employ a daily multi-site rainfall model, which estimates the joint probability distribution of rainfall occurrence and rainfall amounts at different places. From that model, any rainfall based index can be derived and simulated. A daily modelling approach is sophisticated, but it has the advantage that estimation can be based on a much richer data set compared to a direct estimation of weather indices.

The contribution of our paper is twofold. First, we analyse to what extent geographical basis risk of rainfall based weather derivatives can be mitigated by optimal regional diversification. We analyse this problem for two exemplary regions in Germany. Second, to our best knowledge this is the first time that a multi-site rainfall model is utilized in the context of designing weather derivatives. We compare this approach with two simpler methods to determine optimal portfolio weights. These are historical simulation, where the optimal weights of the payoffs in the past are assumed to be the best for the future, and inverse distance weighting, where the weights just depend on the distance to the reference station.

The paper is organised as follows. In the next section, we describe the problem of optimal regional diversification in greater detail. Thereafter, we introduce the daily rainfall model and explain how it can be used to determine optimal portfolio weights. Section 3 describes the data used in this study and reports measures of the model validity. Section 4 applies the rainfall model to the diversification problem and compares the results with two benchmark methods. In Section 5, we conclude with some further discussion.

2. METHODS

2.1. Economic background

The revenue $R_{0,k}$ of a rainfall-dependant producer at location k at time 0 can be modelled as the product of the rainfall-dependant production function $Q_T(I_{T,k})$, the product's market price P and the discount factor $e^{-r\Delta t}$ (cf. Mußhoff et al., 2011):

$$R_{0,k} = Q_T(I_{T,k}) \cdot P \cdot e^{-r\Delta t} \tag{1}$$

where $I_{T,k}$ denotes a rainfall index measured at time T at location k.

The producer can hedge his/her rainfall risk by buying a weather derivative whose payoff depends on a rainfall index $I_{T,l}$ measured at location l. At time 0, the buyer pays $F_0(I_{T,l})$ for this derivative to get a payoff $F_T(I_{T,l})$ at time T, which has to be considered in the revenue function (1). With the derivative with reference weather station l, the revenue $R'_{0,k}$ becomes:

$$R'_{0,k} = R_{0,k} + F_T(I_{T,l}) \cdot e^{-r\Delta t} - F_0(I_{T,l}) \tag{2}$$

Ideally, the location of the production k and the reference weather station of the derivative l are the same. If $k \neq l$, the problem of geographical basis risk appears: The production $Q_T(I_{T,k})$ depends on the rainfall at the buyer's place k, but the derivative has a different reference station l with payoff $F_T(I_{T,l})$.

Instead of buying only one derivative for location l, the buyer can combine derivatives of the K nearest reference stations. For this combination, he/she would pay a compound premium $F_0^K(I_T)$ with

$$F_0^K(I_T) := \sum_{i=1}^{K} \omega_i F_0(I_{T,i}) \tag{3}$$

and ω_i being the weight for the ith derivative. At time T, the buyer would receive a compound payoff $F_T^K(I_T)$,

$$F_T^K(I_T) := \sum_{i=1}^{K} \omega_i F_T(I_{T,i}) \tag{4}$$

with the same weights ω_i and I_T denoting the same weather index as before, measured at K reference stations. Consequently, the revenue with K derivatives becomes:

$$R'^K_{0,k} = R_{0,k} + F_T^K(I_T) \cdot e^{-r\Delta t} - F_0^K(I_T)$$

If, hypothetically, derivatives were offered for location k, so $k = l$ in Equation (2), the revenue would be:

$$\widetilde{R}'_{0,k} = R_{0,k} + \widetilde{F}_T(I_{T,k}) \cdot e^{-r\Delta t} - \widetilde{F}_0(I_{T,k})$$

with '$\sim$' denoting the hypothetical case.

Our aim is to identify the best approximation of the hypothetical revenue $\widetilde{R}'_{0,k}$ by means of the compound revenue $R'^K_{0,k}$. This can be achieved by minimizing the (quadratic) difference between $\widetilde{R}'_{0,k}$ and $R'^K_{0,k}$:

$$\min_K \left(\widetilde{R}'_{0,k} - R'^K_{0,k}\right)^2 \text{ with}$$

$$\begin{aligned}
\widetilde{R}'_{0,k} - R'^K_{0,k} &= (R_{0,k} + \underbrace{\widetilde{F}_T(I_{T,k}) \cdot e^{-r\Delta t} - \widetilde{F}_0(I_{T,k})}_{\text{hypothetical derivative}}) - (R_{0,k} + \underbrace{F_T^K(I_T) \cdot e^{-r\Delta t} - F_0^K(I_T)}_{\text{combination of K derivatives}}) \\
&= (\widetilde{F}_T(I_{T,k}) \cdot e^{-r\Delta t} - \widetilde{F}_0(I_{T,k})) - (F_T^K(I_T) \cdot e^{-r\Delta t} - F_0^K(I_T)) \\
&= (\underbrace{\widetilde{F}_T(I_{T,k}) - F_T^K(I_T)}_{\text{payoffs}}) \cdot e^{-r\Delta t} - (\underbrace{\widetilde{F}_0(I_{T,k}) - F_0^K(I_T)}_{\text{prices}})
\end{aligned} \tag{5}$$

The first part of Equation (5) describes the difference between the payoffs from the hypothetical derivative and the compound derivative. The second term measures the difference between their prices. Note that this equation is independent of $R_{0,k}$, the revenue without a derivative. Hence, the results do not need a specification of the production function as long as the production depends on the same weather index as the derivative. The prices in Equation (5) are difficult to obtain, especially for the hypothetical derivative, which is not offered on the market. In order to avoid further assumptions on the pricing methods, we neglect the difference between the prices and concentrate on the difference between the payoffs, i.e. the approximation of $\widetilde{F}_T(I_{T,k})$ by $F_T^K(I_T)$.

As it can be seen from Equation (3) and Equation (4), the weights ω_i affect the results of the approximation, too. To get a good approximation, the optimal weights have to be determined. In the next section, a multi-site rainfall model is proposed that supports the determination of these weights. Afterwards, two benchmark approaches for calculating the weights are introduced.

2.2. Multi-site rainfall model

In the multi-site rainfall model approach (MRM), a rainfall model is adjusted to the historical daily rainfall at all locations where derivatives are offered. This model is able to simulate future rainfall paths and the optimal weights are determined on the basis of these simulations.

The model used in this study is based on the Wilks model (Wilks, 1998) as this is a widely accepted model and outperformed the hidden Markov model (Hughes et al., 1999) and the non-parametric K-nearest neighbour model (Buishand and Brandsma, 2011) in the generation of multi-site precipitation occurrence processes (Mehrota et al., 2006). In the multi-site rainfall model, the daily rainfall amount $Y_{t,k}$ at time t at location k is described as the product of a rainfall amount process $r_{t,k}$ and a rainfall occurrence process $X_{t,k}$.

$$Y_{t,k} = \underbrace{r_{t,k}}_{\text{amount}} \cdot \underbrace{X_{t,k}}_{\text{occurrence}} \tag{6}$$

These two processes are modelled separately for each station (single-site) and used later for multi-site modelling.

Occurrence process. The single-site daily occurrence process $X_{t,k}$ for location k is modelled as a zero-one process for rain (1) or no rain (0).

$$X_{t,k} = \begin{cases} 0 & \text{if day } t \text{ is dry at location } k \\ 1 & \text{if day } t \text{ is wet at location } k \end{cases}$$

$X_{t,k}$ is assumed to follow a first-order, two-state Markov process implying that the probability of rainfall occurrence just depends on the weather condition of the previous day (cf. Roldán and Woolhiser, 1982; Wilks, 1998). The process is described by the transition probabilities $p_{t,k}^{01}$ and $p_{t,k}^{11}$, which capture the probability of rain if it rained on the previous day or not.

$$\begin{aligned} p_{t,k}^{01} &= \Pr\{X_{t,k} = 1 \mid X_{t-1,k} = 0\} \\ p_{t,k}^{11} &= \Pr\{X_{t,k} = 1 \mid X_{t-1,k} = 1\} \end{aligned} \tag{7}$$

In the following, we write $p_{t,k}^{01/11}$ as an abbreviation for $p_{t,k}^{01}$ and $p_{t,k}^{11}$.

Wilks (1998) models these probabilities as monthly constants. In this paper, however, we model them as daily changing values within a year and approximate them by truncated Fourier series. Between years, transition probabilities stay constant, i.e. $p_{t,k}^{01/11} = p_{t+365,k}^{01/11}$. The coefficients of the Fourier series are estimated by maximizing log-likelihood functions (Woolhiser and Pegram, 1979). The order of the Fourier series is determined by means of the Akaike information criterion (AIC).

Following Wilks (1998), the single-site occurrence process can be simulated step by step by using standard normally distributed random numbers[1] $\varepsilon_{t,k} \sim N(0,1)$ and a starting value $X_{0,k}$:

$$X_{t,k}^{\text{sim}} = \begin{cases} 1 & \text{if } \Phi[\varepsilon_{t,k}] \le p_{t,k}^{01/11} \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

Amount process. The single-site daily rainfall amount process $r_{t,k}$ for location k is assumed to follow a mixed exponential distribution (cf. Woolhiser and Roldán, 1982; Wilks, 1998):

$$f[r_{t,k}] = \frac{\alpha_{t,k}}{\beta_{t,k}} \exp\left[\frac{-r_{t,k}}{\beta_{t,k}}\right] + \frac{1-\alpha_{t,k}}{\gamma_{t,k}} \exp\left[\frac{-r_{t,k}}{\gamma_{t,k}}\right] \tag{9}$$

with $\beta_{t,k} \ge \gamma_{t,k} > 0$ and $0 < \alpha_{t,k} < 1$ for all t. It is the sum of two exponential distributions, one with a higher mean $\beta_{t,k}$ and one with a lower mean $\gamma_{t,k}$, mixed by the parameter $\alpha_{t,k}$.

Wilks (1998) also models these parameters as monthly constant, whereas in this paper, they are approximated by truncated Fourier series. Again, the coefficients of the Fourier series were estimated by maximizing log-likelihood functions (Woolhiser and Pegram, 1979) and their orders are determined by the AIC. For simplicity, we here assumed that the orders of the Fourier series for $\alpha_{t,k}$, $\beta_{t,k}$ and $\gamma_{t,k}$ are equal. Otherwise, the computational effort would increase cubically. When the parameters $\alpha_{t,k}$, $\beta_{t,k}$ and $\gamma_{t,k}$ are estimated for location k, the amount process for location k can be simulated via:

$$r_{t,k}^{\text{sim}} = r_{\min} - \delta_{t,k} \ln\left[\Phi\left(z_{t,k}\right)\right] \tag{10}$$

Here, $r_{\min}$ describes the minimal amount that is detected as rain (0.1 mm), and $\delta_{t,k}$ is given by

$$\delta_{t,k} = \begin{cases} \beta_{t,k} & \text{if } u_t \le \alpha_{t,k} \\ \gamma_{t,k} & \text{if } u_t > \alpha_{t,k} \end{cases}$$

with a uniform $[0,1]$ random variable u_t deciding which exponential distribution is chosen. The variate $z_{t,k}$ in Equation (10) follows a standard normal distribution and is independent of $\varepsilon_{t,k}$.

Multi-site modelling. After the single-site occurrence processes and the single-site amount processes are estimated for every location, these processes can be used for the multi-site modelling. Similar to the single-site occurrence process (8), the multi-site occurrence process is simulated by

[1] Instead of using standard normally distributed random numbers, a uniform random variable $u_{t,k}$ could be compared with $p_t^{01/11}(k)$ (rain if $u_{t,k} \le p_{t,k}^{01/11}$) in Equation (8). However, the notation with the standard normal cumulative distribution function Φ was chosen because it is more practical for the multi-site modelling.

$$X_{t,k}^{\text{sim}} = \begin{cases} 1 & \text{if } \Phi[\varepsilon_{t,k}] \le p_{t,k}^{01/11} \\ 0 & \text{otherwise} \end{cases} \tag{11}$$

for all locations k simultaneously. The $\varepsilon_{t,k}$ are again temporally independent, but now spatially correlated random variables: $\boldsymbol{\varepsilon}_t \sim N_{K+1}(\mathbf{0},\boldsymbol{\Sigma})$. The entries $\sigma(k,l)$ of the covariance matrix $\boldsymbol{\Sigma}$ describe the spatial correlations between the random variables for location k and l, $\sigma(k,l) = \text{Corr}\left[\varepsilon_{t,k}, \varepsilon_{t,l}\right]$.

Computing these correlations separately for each pair of stations leads to a large number of parameters that have to be estimated[2] as well as to the problem of a possibly not positive semidefinite matrix $\boldsymbol{\Sigma}$. We overcome these problems by defining $\sigma(k,l)$ by a function of the distances between the stations:

$$\sigma(k,l) = \exp\left(-d_1 D(k,l)^{d_2}\right) \tag{12}$$

with $D(k,l)$ being the distance between locations k and l (Wilks, 1998). The correlation $\sigma(k,l)$ decreases with increasing distance $D(k,l)$. The parameters d_1 and d_2 are estimated so that they minimize the summarized quadratic difference between the resulting correlation $\xi(k,l)$ and the historical correlation $\xi^0(k,l)$ for all pairs (k,l), i.e. $\sum_{k,l}\left[\xi(k,l) - \xi^0(k,l)\right]^2$.

If the single-site parameters $\alpha_{t,k}$, $\beta_{t,k}$ and $\gamma_{t,k}$ and the covariance matrix $\boldsymbol{\Sigma}$ are estimated for every location, the multi-site amount process is - similar to the single-site amount process in Equation (10) - determined for all locations k simultaneously by

$$r_{t,k}^{\text{sim}} = r_{\min} - \delta_{t,k} \ln[\Phi(z_{t,k})] \tag{13}$$

Here, $r_{\min}$ describes again the minimal amount that is detected as rain, and δ_t is given by

$$\delta_{t,k} = \begin{cases} \beta_{t,k} & \text{if } \dfrac{\Phi\left[\varepsilon_{t,k}\right]}{p_{t,k}^{01/11}} \le \alpha_{t,k} \\ \gamma_{t,k} & \text{if } \dfrac{\Phi\left[\varepsilon_{t,k}\right]}{p_{t,k}^{01/11}} > \alpha_{t,k} \end{cases} \tag{14}$$

In the case of rainfall occurrence, $\Phi\left[\varepsilon_{t,k}\right]$ is always smaller than or equal to $p_{t,k}^{01/11}$ (see Equation (8)). Assuming that there is no rain at the nearest neighbours of location k, so that $\Phi[\varepsilon_{t,l}] > p_{t,l}^{01/11}$ for location l being close to location k. Because the correlations of the $\varepsilon_{t,k}$ are

[2]The correlations $\sigma(k,l)$ cannot be derived directly from the data as only the correlations between the historical occurrence processes are observable, $\xi^0(k,l) = \text{Corr}\left[X_{t,k}^0, X_{t,l}^0\right]$. The correlation of the random variables $\sigma(k,l) = \text{Corr}\left[\varepsilon_{t,k}, \varepsilon_{t,l}\right]$ can be obtained via trial and error by guessing $\sigma(k,l)$ for every pair of locations and comparing the resulting $\xi(k,l)$ with the historical $\xi^0(k,l)$. By using an adequate algorithm, this procedure can be repeated until $\xi(k,l)$ and $\xi^0(k,l)$ match.

high for near neighbours, $\Phi\lfloor\varepsilon_{t,k}\rfloor$ is then close to $p_{t,k}^{01/11}$, thus $\frac{\Phi\lfloor\varepsilon_{t,k}\rfloor}{p_{t,k}^{01/11}} > \alpha_{t,k}$. This leads to the smaller parameter $\gamma_{t,k}$ in Equation (14), so only little rainfall is expected if location k is in the center of a dry area. If there is rain at the nearest neighbours, however, $\Phi[\varepsilon_{t,k}]$ is quite small, so that $\frac{\Phi\lfloor\varepsilon_{t,k}\rfloor}{p_{t,k}^{01/11}} \le \alpha_{t,k}$, which leads to the higher mean $\beta_{t,k}$ in Equation (14).

In fact, a modified version of Equation (14) is used, where $\beta_{t,k}$ is replaced by the following term as it better reflects the rainfall at the margins of wet regions (see Wilks (1998) for details):

$$\delta_{t,k} = \begin{cases} \gamma_{t,k} + 2\left[\beta_{t,k} - \gamma_{t,k}\right]\left(1 - \dfrac{\Phi(\varepsilon_{t,k})}{\alpha_{t,k} \cdot p_{t,k}^{01/11}}\right) & \text{if } \dfrac{\Phi\left[\varepsilon_{t,k}\right]}{p_{t,k}^{01/11}} \le \alpha_{t,k} \\ \gamma_{t,k} & \text{if } \dfrac{\Phi\left[\varepsilon_{t,k}\right]}{p_{t,k}^{01/11}} > \alpha_{t,k} \end{cases}$$

As for the occurrence process, the $\mathbf{z}_{t,k}$ in Equation (13) are temporally independent, but spatially correlated random variables: $\mathbf{z}_{t,k} \sim N_{K+1}(\mathbf{0}, \mathbf{Z})$. The entries of $\mathbf{Z}$ describe the correlations between the random variables, $\zeta(k,l) = \mathrm{Corr}\lfloor z_{t,k}, z_{t,l}\rfloor$, but only the historical correlations between the amounts are observable, $\eta^0(k,l) = \mathrm{Corr}\left[Y_{t,k}^0, Y_{t,k}^0\right]$. Similar to Equation (12), the $\zeta(k,l)$ are obtained by a function of the distance between locations:

$$\zeta(k,l) = \exp\left(-d_3 D(k,l)^{d_4}\right) \tag{15}$$

with d_3 and d_4 minimizing $\sum_{k,l}\left[\eta(k,l) - \eta^0(k,l)\right]^2$.

With all parameters at hand, the multi-site precipitation process can be simulated for all locations k - following Equation (6) - as the product of the simulated multi-site occurrence process (Equation (11)) and the simulated multi-site amount process (Equation (13)). This multi-site model can be used to simulate future rainfall $Y_{t,k}^{\mathrm{MRM}}$ for all locations simultaneously resulting in future rainfall indices $I_{T,k}^{\mathrm{MRM}}$ and future payoffs of the derivative $F_T(I_{T,k}^{\mathrm{MRM}})$. Then, the optimal weights ω_i^{MRM} are obtained by linear regression depending on the number of derivatives K used for the combination. The equation system used for the regression consists of n^{MRM} equations of the form

$$\begin{aligned} \widetilde{F}_T(I_{T,k}^{\mathrm{MRM}}) &= \omega_1^{\mathrm{MRM}} F_T(I_{T,1}^{\mathrm{MRM}}) + \omega_2^{\mathrm{MRM}} F_T(I_{T,2}^{\mathrm{MRM}}) + \ldots + \omega_K^{\mathrm{MRM}} F_T(I_{T,K}^{\mathrm{MRM}}) + \varepsilon \\ &= \sum_{i=1}^{K} \omega_i^{\mathrm{MRM}} F_T(I_{T,i}^{\mathrm{MRM}}) + \varepsilon \end{aligned} \tag{16}$$

for n^{MRM} years of simulated rainfall data. ε denotes the error term.

2.3. Benchmark approaches

Historical simulation. The first benchmark approach is historical simulation (HS), which is actually based on the empirical rainfall distribution. It analyzes how the derivative would have performed in the past and assumes that the best weights for the historical payoffs will perform best in the future. The optimal weights ω_i^{HS} are obtained by linear regression with n^{HS} equations of the form

$$\begin{aligned}\widetilde{F}_T(I_{T,k}) &= \omega_1^{\mathrm{HS}} F_T(I_{T,1}) + \omega_2^{\mathrm{HS}} F_T(I_{T,2}) + \ldots + \omega_K^{\mathrm{HS}} F_T(I_{T,K}) + \varepsilon \\ &= \sum_{i=1}^{K} \omega_i^{\mathrm{HS}} F_T(I_{T,i}) + \varepsilon\end{aligned}$$

for n^{HS} years of historical rainfall data. ε denotes again the error term.

The advantage of this method is that it is easy to implement. The results, however, strongly depend on the length n^{HS} of the used dataset. If the index is calculated only once a year, many years are needed. In particular, if the strike level of a put (call) option is set quite high (low), the derivative rarely leads to a payoff and the weights depend on very few observations.

Inverse distance weighting. The inverse distance weighting (IDW) approach does not depend on any historical weather data, but only on the location of the weather stations and their distance to the reference station. It is based on the observation that the correlation between weather events decreases by increasing distance, so that the weights $\omega_i^{\mathrm{IDW}p}$ are determined by the inverse of the distance $D(0,i)$ between locations 0 and i to the power of p, $p = 1,2,3$, normalized by the sum of the weights (Shepard, 1968):

$$w_i = \frac{1}{\mathrm{D}(0,i)^p}$$

$$\omega_i^{\mathrm{IDW}p} = \frac{w_i}{\sum_i w_i} \tag{17}$$

It is important to note that this approach in contrast to the two other methods does not depend on any weather data. For the two other approaches, a weather station close to the production place is needed to calculate the dependency on the neighbours' weather stations where derivatives are offered. This restriction is not necessary for the inverse distance weighting, so this approach is much more flexible and can be used for any location. Another advantage of this method is its simplicity. No historical data is needed, the weights just depend on the geographical distances. This in turn is also the disadvantage: Any spatial dependency beyond the distance is neglected.

3. DATA

3.1. Rainfall data

The dataset used in this study is provided by the German Meteorological Service (DWD). It contains daily rainfall data for 49 German weather stations (see Fig. 1) from 01/01/1973 to 31/12/2010 (38 years). 47 stations are situated all over Germany and represent most of the stations where weather derivatives are available from the private supplier of weather derivatives 'Celsius Pro' [3]. The other two of the 49 stations represent farms as they are stations in small towns: one is in Koßdorf, a town in the south of Brandenburg (eastern Germany), the other one is situated in Nordhausen in the south of the mountain range Harz in the center of Germany. Those two locations were chosen because of different geographical conditions: Brandenburg is a plain region, whereas the Harz is a mountainous area. Agriculture plays an important role in both regions.

Figure 1. Map of the stations used in this study; Koßdorf and Nordhausen (gray) represent two farms, the other stations (white) represent those places where derivatives are available

Source: own picture

So far, the notation was very general, neither the rainfall index nor the payoff function were specified in Section 2. For the rainfall index $I_{T,k}$ at location k, we will use exemplarily the sum of rainfall in May (in mm),

---

[3] Celsius Pro: http://www.celsiuspro.com/

$$I_{T,k} = \sum_{t=1}^{T} Y_{t,k}^{\text{May}}$$

and for the derivative a put option on this index with payoff

$$F_T(I_{T,k}) = \lambda_k \cdot \max\left(0, \text{Strike}_k - I_{T,k}\right) \tag{18}$$

λ_k describes the tick size which converts the index outcome into a monetary amount. It can be set differently for every location k, but for simplification we assume $\lambda_k = \lambda = 1$. Strike_k is set to the historical average of rainfall in May at location k. So this option has positive payoff if the rainfall in May is lower than the historical average. The month May was chosen because rain is very important for farmers in Germany in this period.

3.2. Validation

To examine the performance of the different approaches in predicting the portfolio weights, we use an out-of-sample prediction, i.e., we split the whole dataset into a training and a testing dataset. The training dataset includes the rainfall data from 1973 to 2000, which is used to calculate the weights. Then, it is analyzed how these weights perform in the testing phase from 2001 to 2010. For these ten years, the hypothetical payoffs at the buyer's place, $\widetilde{F}_T(I_{T,k})$, and the compound payoffs, $F_T^K(I_T)$, are calculated and compared using the root mean square error (RMSE):

$$\text{RMSE}\left(\widetilde{F}_T(I_{T,k}), F_T^K(I_T)\right) = \sqrt{\frac{1}{N} \cdot \sum_{n=1}^{N} \left(\widetilde{F}_T(I_{T_n,k}) - F_T^K(I_{T_n})\right)^2} \tag{19}$$

with $N = 10$ years of testing data.

4. RESULTS

4.1. Ex post analysis

At first, the whole dataset (1973-2010) is analyzed for examining the assumption that a linear combination of several neighbours represents the rainfall $Y_{t,k}$ at a place k better than just the nearest neighbour. For this purpose, we calculate the RMSE (Equation (19)) between the actual rainfall at the buyer's place (Koßdorf or Nordhausen) and the rainfall at the nearest station by distance (Dresden and Brocken, respectively), i.e. $K = 1$. Then, using linear regression, we calculate the optimal combination of the two nearest stations to approximate the actual rainfall ($K = 2$). The aggregated rainfall is again compared with the actual rainfall at the buyer's place[4]. This procedure is repeated for more and more stations, sorted by their distance to the buyer's

[4]Another way of sorting the neighbours is the correlation of the historical payoffs of the put option. The RMSE for the daily rainfall, however, almost does not change if the neighbours are sorted by correlation instead of distance.

place. The results for Koßdorf and Nordhausen are depicted in the first part of Table 1 for $1 \le K \le 10$. Apparently, the error decreases about 20% if more stations are used than only one, but for more than about five stations, the error remains almost constant.

The buyer, however, is not primarily interested in mimicking the daily rainfall $Y_{t,k}$ but rather in approximating the hypothetical payoff $\widetilde{F}_T(I_{T,k})$ that he/she would get from a derivative without geographical basis risk. Therefore, the calculations of the RMSE are repeated with the payoffs from a put option on cumulative rainfall in May defined in Equation (18). Table 1 shows that for the payoff from the put option, the approximation error decreases by 29% (Koßdorf) and 17% (Nordhausen) with an increasing number of included derivatives[5]. However, for the payoff from the put option, there are only 38 (yearly) values, while for the daily rainfall, there are 365 [days] x 38 [years] = 13870 values available. Because of the low number of observations, the calculation of the RMSE is stopped at $K = 7$ to avoid overfitting caused by too many parameters relative to the number of observations[6].

The findings of an decreasing RMSE motivate that using several stations instead of the nearest one better replicates the payoff the buyer needs and reduces the geographical basis risk.

Table 1. Root mean square error (RMSE) for the daily rainfall and for the payoff from the put option in May for Koßdorf and Nordhausen in dependence of the number of K nearest neighbours used for linear combination (sorted by distance), data 1973-2010

	$K=1$	$K=2$	$K=3$	$K=4$	$K=5$	$K=6$	$K=7$	$K=8$	$K=9$	$K=10$
Daily rainfall										
RMSE (Koßdorf)	2.73	2.44	2.34	2.30	2.26	2.25	2.25	2.25	2.24	2.24
RMSE (Nordhausen)	2.75	2.47	2.34	2.22	2.20	2.20	2.19	2.19	2.19	2.19
Put option										
RMSE (Koßdorf)	7.16	6.69	6.04	5.78	5.39	5.38	5.05	-	-	-
RMSE (Nordhausen)	8.02	7.05	6.88	6.86	6.85	6.75	6.68	-	-	-

Source: own calculation

4.2. Out-of-sample prediction

In the previous section, the whole dataset was used to determine the best weights for the combination via regression from an *ex post* perspective. In this part, the multi-site rainfall model and the two benchmark approaches for predicting the best weights are used to compare their performance. For this purpose, we split the dataset into training data (1973-2000, 28 years) and

[5]The neighbours are again sorted by distance. For the payoff from the put option, the difference between sorting the neighbours by distance or by correlation is larger than for the daily rainfall. However, the way of sorting does not change the overall message so that we will continue with sorting by distance only.
[6]Bentler and Chou (1987) set a minimal ratio of the sample size to the number of free parameters of 5:1.

testing data (2001-2010, 10 years). The training data is used to determine the weights using the multi-site rainfall model and using historical simulation (The inverse distance weighting is independent of the weather data.). The performance of these weights is then checked for the testing data.

For the multi-site rainfall model, the parameters are adjusted to the training data for each location. The empirical transition probabilities $p_{t,k}^{01}$ and $p_{t,k}^{11}$ (see Equation (7)) for Koßdorf and Nordhausen as well as the fitted Fourier series are shown in Fig. 4 in the appendix. The parameters for the mixed exponential distribution for the rainfall amount $\alpha_{t,k}$, $\beta_{t,k}$ and $\gamma_{t,k}$ (see Equation (9)) for the two places are depicted in Fig. 5 in the appendix. The results for the decorrelation function (Equation (12)) describing $\mathbf{\Sigma}$ are $d_1 = 0.0012$ and $d_2 = 1.0022$, so that the correlations of the rainfall occurrence between two different locations range from 0.3610 to 0.9855. For Equation (15) describing $\mathbf{Z}$, the estimated parameters are $d_3 = 0.0075$ and $d_4 = 0.9373$ and the resulting correlations vary between 0.0143 and 0.9230. With the parameters estimated for all places and the covariance matrices $\mathbf{\Sigma}$ and $\mathbf{Z}$, the multi-site rainfall model is specified and used to forecast the rainfall in May in Koßdorf, Nordhausen and all other places 10000 times. A linear regression provides the optimal weights for the simulated data. After all the weights were determined with different methods, they are used to combine the test data of the K nearest neighbours. These compound payoffs are then compared with the hypothetical payoff in the testing phase by the RMSE.

As an example, Table 2 shows the optimal weights for replicating the payoff of the derivative for Koßdorf with $K = 5$ nearest neighbours, i. e. Dresden, Leipzig/Halle, Cottbus, Lindenberg and Potsdam, according to the different calculation methods. It can be seen that the historical simulation puts more weight on Lindenberg (0.555) and Potsdam (0.254) than on Cottbus (0.019) even though Cottbus is nearer. The negative weight for Leipzig/Halle (-0.420) indicates that the farmer should sell this contract. The results for the multi-site rainfall model differ significantly: The highest weight is put on Leipzig/Halle (0.350), followed by Dresden (0.268). By definition, the weights from inverse distance weighting decrease with increasing distance and the decline is stronger for higher p.

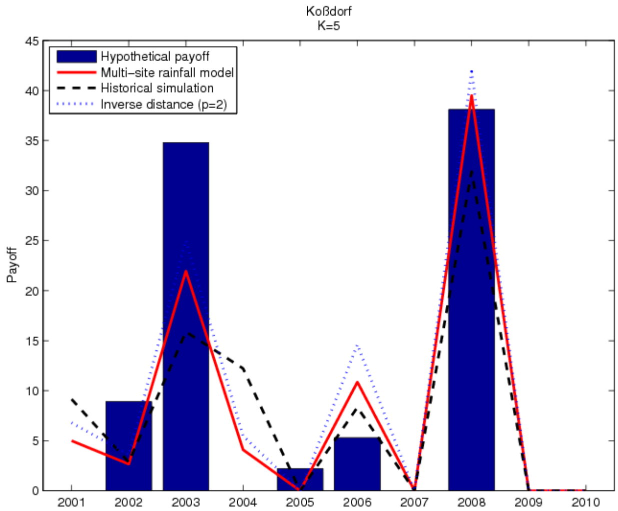
These weights describe how the nearest neighbours have to be combined to approximate the hypothetical payoff in Koßdorf. The resulting compound payoff for these weights for Koßdorf and $K = 5$ from 2001 to 2010 are shown in Fig. 2 in comparison with the hypothetical payoff the farmer would receive if this option was offered for Koßdorf. The compound and the hypothetical payoff for different weights are compared using the RMSE, which is shown in the last column of Table 2. In this application, the historical simulation performs worst and the multi-site rainfall method outperforms the other approaches.

Table 2: Optimal weights for replicating the payoff in Koßdorf, $K = 5$, calculated by different methods and the resulting RMSE when comparing the compound with the hypothetical payoff (neighbours sorted by their distance to Koßdorf)

	Dresden	Leipzig/Halle	Cottbus	Lindenberg	Potsdam	RMSE
Multi-site rainfall model	0.268	0.350	0.130	0.154	0.057	5.33
Historical simulation	0.288	-0.420	0.019	0.555	0.254	8.25
Inverse distance ($p = 1$)	0.277	0.221	0.190	0.156	0.156	5.63
Inverse distance ($p = 2$)	0.364	0.233	0.172	0.116	0.115	5.53
Inverse distance ($p = 3$)	0.456	0.233	0.148	0.082	0.081	5.59

Source: own calculation

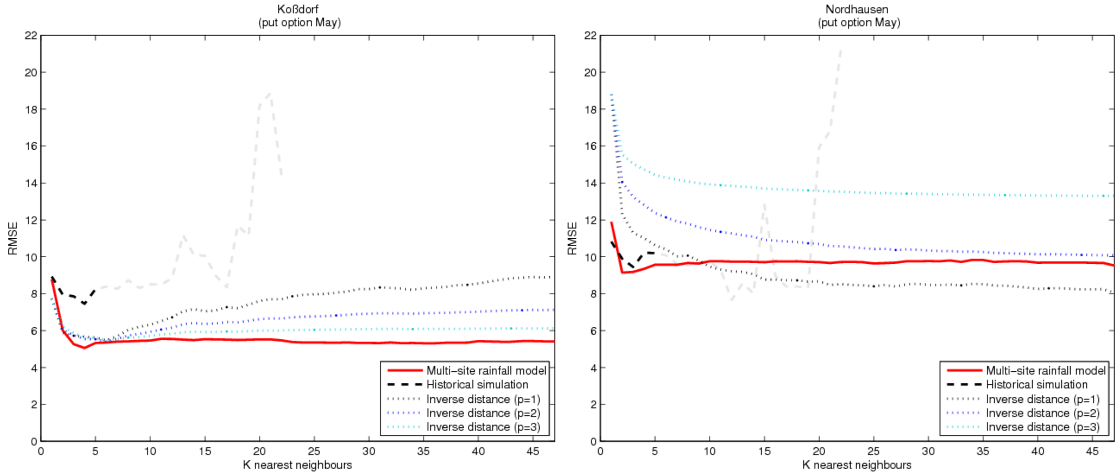
Figure 2. Hypothetical and compound payoff for Koßdorf 2001-2010, $K = 5$



Source: own calculation

When changing the value of K, different weights and different values of the RMSE are obtained. The results of the RMSE for Koßdorf and Nordhausen for different K are shown in Fig. 3 and Table 3. For Koßdorf, we can see that the error decreases for the multi-site rainfall model and the IDW models by using more than one place. For the multi-site rainfall model, a decline of around 40% is achieved. As before, the error stays constant for $K > 5$ (MRM) or increases again (IDW). The historical simulation obviously performs poorly in predicting the weights. The values for the historical simulation for $K > 5$ are unstable as they are the result of overfitting caused by not enough observations for the regression (28). The graphs for Nordhausen show that the historical simulation can arbitrarily outperform the other approaches for a large K. Also the IDW shows a varying performance: The IDW1 model offers the best results for Nordhausen, while it was the IDW3 model for Koßdorf. The error for the IDW method for Nordhausen is notably high for $K = 1$. This is due to the fact that this method puts a weight of 1 if there is only one neighbour, as it can be seen in Equation (17), whereas the other approaches also allow for weights different from 1. Nordhausen's nearest neighbour Brocken is situated much higher (1142 m AMSL) than Nordhausen (185 m AMSL) leading to different rainfall.

Figure 3. Root mean square error for Koßdorf (left) and Nordhausen (right) for the out-of-sample prediction, test data 2001-2010; the values for the historical simulation for $K > 5$ (gray) are the result of overfitting



Source: own calculation

Table 3. Reduction of weather risk and Geographical Basis Risk (GBR) for Koßdorf and Nordhausen, weights from multi-site rainfall model

	Koßdorf			Nordhausen		
	RMSE	Change of weather risk	GBR	RMSE	Change of weather risk	GBR
$K = 0$	16.66	-	-	17.59	-	-
$K = 1$	8.77	-47%	-	11.90	-32%	-
$K = 2$	6.01	-64%	-31%	9.14	-48%	-23%
$K = 3$	5.27	-68%	-40%	9.18	-48%	-23%
$K = 4$	5.06	-70%	-42%	9.34	-47%	-21%
$K = 5$	5.33	-68%	-39%	9.57	-46%	-20%

Source: own calculation

All in all, the weights obtained by the multi-site rainfall model appear to be the most reliable ones and lead to a reduction of the geographical basis risk of between 20% and 40%, even though this method can be outperformed by the two models in some cases. Furthermore, the results indicate that a small number of derivatives is sufficient: For the multi-site rainfall model, the increase is the highest for $K = 4$ (Koßdorf) or $K = 2$ (Nordhausen).

Table 3 also exhibits the approximation error for $K = 0$. This describes the deviation if no derivative is bought, i. e. if all weights are set to zero. We can assume that the weather-dependant producer from Section 2 with production site l could perfectly hedge his/her weather risk by buying a weather derivative with a rainfall-dependant payoff $F_T(I_{T,l})$ with reference station l, i. e. the production basis risk and the geographical basis risk are zero in this case. Then, the RMSE for $K = 0$ represents a measure for the producer's weather risk. However, the derivative is not provided for place l, so that the producer has to buy derivatives for other reference stations. As it can be seen from Table 3, buying one derivative reduces his/her weather risk by 47% (Koßdorf) or 32% (Nordhausen). With the weights from the multi-site rainfall model and the optimal value K, the weather risk decreases up to 70% (Koßdorf) or almost 50% (Nordhausen). The remaining 30% to 50% describe the geographical basis risk that can not be hedged.

5. Discussion and Conclusion

In this article, we showed that geographical basis risk inherent to rainfall based weather derivatives can be reduced by combining derivatives of adjacent reference stations. By approximating the payoff of a hypothetical derivative by the weighted combination of the K nearest neighbours, the error decreases about 20% when using more than one neighbour.

For calculating the optimal weights for the portfolio, a multi-site rainfall model was proposed, which is calibrated to the historical data and then simulates future rainfall. The performance of this model was compared to two benchmark approaches: An intuitive one, where those weights are chosen which performed best in the past, and a very easy one just depending on the difference to the reference station. It turned out that the multi-site rainfall model leads to a reduction of the geographical basis risk between 20% and 40% and in general outperforms the other methods, which, however, sometimes also perform well.

One crucial point of our analysis is that we neglected transaction costs. If the buyer had to pay transaction costs for every single contract, the advantage of the better approximation would vanish. The intention of this research was rather that the seller offers this portfolio of weather derivatives to the buyer, so the transaction costs incurred only once. It is also much easier for the supplier of the derivatives to carry out this analysis.

Without increasing transaction costs, there is no reason why the number of stations used for the combination should be constrained. The results show, however, that the error is minimal if about 5 stations are used. This makes sense as the correlation decreases with increasing distance, so that stations which are located further away do not make any difference in the approximation.

This study provides first evidence for the reduction of geographical basis risk by combining weather derivatives with different reference stations. It has to be examined, however, if the results can be generalized and transferred to other geographical regions or indices. As the network of rainfall measuring weather stations in Germany is quite tight, it would be interesting to elaborate if these results can be transferred to other regions, where the distance to the nearest neighbours would amount to several hundred kilometres.

Acknowledgements

The financial support from the German Research Foundation via the CRC 649 "Economic Risk", Humboldt-University Berlin, is gratefully acknowledged. Moreover, the authors would like to thank the participants of the CRC 649 Conference 2011 for their helpful comments and discussions on this topic.

REFERENCES

Bentler, P. M. and Chou, C.-P. (1987). Practical issues in structural modelling. *Sociological Methods Research* 16:78-117.

Berg, E. and Schmitz, B. (2008). Weather-based instruments in the context of whole-farm risk management. *Agricultural Finance Review* 68:119-133.

Buishand, T. A. and Brandsma, T. (2001). Multisite simulation of daily precipitation and temperature in the Rhine Basin by nearest-neighbor resampling. *Water Resources Research* 37:2761-2776.

Chen, G. and Roberts, M. C. (2004). Weather derivatives in the presence of index and geographical basis risk: Hedging dairy profit risk. Proceedings of the NCR-134 Conference on Applied Commodity Price Analysis, Forecasting, and Market Risk Management. St. Louis, MO.

Hughes, J. P., Guttorp, P., and Charles, S. P. (1999). A non-homogeneous hidden Markov model for precipitation occurrence. *Applied Statistics* 48:15-30.

Jin, Y., Rejesus, R. M., and Little, B. B. (2005). Binary choice models for rare events data: a crop insurance fraud application. *Applied Economics* 37:841-848.

Lazo, J. K., Lawson, M., Larsen, P. H., and Waldman, D. M. (2011). US economic sensitivity to weather variability. *Bulletin of the American Meteorological Society* 92:709-720.

Mehrotra, R., Srikanthan, R., and Sharma, A. (2006). A comparison of three stochastic multi-site precipitation occurrence generators. *Journal of Hydrology* 331:280-292.

Mußhoff, O., Odening, M., and Xu, W. (2011). Management of climate risks in agriculture - will weather derivatives permeate? *Applied Economics* 43:1067-1077.

Nadolnyak, D. and Hartarska, V. (2012). Agricultural disaster payments in the southeastern US: do weather and climate variability matter? *Applied Economics* 44:4331-4342.

Odening, M., Mußhoff, O., and Xu, W. (2007). Analysis of rainfall derivatives using daily precipitation models: Opportunities and pitfalls. *Agricultural Finance Review* 67:135-156.

Roldán, J. and Woolhiser, D. A. (1982). Stochastic daily precipitation models: 1. a comparison of occurrence processes. *Water Resources Research* 18:1451-1459.

Salsón, S. and Garcia-Bartual, R. (2003). A space-time rainfall generator for highly convective Mediterranean rainstorms. *Natural Hazards and Earth System Sciences* 3:103-114.

Shepard, D. (1968). A two-dimensional interpolation function for irregularly-spaced data. Proceedings of the 1968 ACM National Conference, pages 517-524.

Vedenov, D. V. and Barnett, B. J. (2004). Efficiency of weather derivatives as primary crop insurance instruments, *Journal of Agricultural and Resource Economics* 29:387-403.

Wilks, D. S. (1998). Multisite generalization of a daily stochastic precipitation generation model. *Journal of Hydrology* 210:178-191.

Woodard, J. D. and Garcia, P. (2008). Basis risk and weather hedging effectiveness. *Agricultural Finance Review* 68:99-117.

Woolhiser, D. A. and Pegram, G. G. S. (1979). Maximum likelihood estimation of Fourier coefficients to describe seasonal variations of parameters in stochastic daily precipitation models. *Journal of Applied Meteorology* 18:34-42.

Woolhiser, D. A. and Roldán, J. (1982). Stochastic daily precipitation models: 2. a comparison of distributions of amounts. *Water Resources Research* 18:1461-1468.

Xu, W., Filler, G., Odening, M., and Okhrin, O. (2010). On the systemic nature of weather risk. *Agricultural Finance Review* 70:267-284.

APPENDIX

Figure 4. Empirical and estimated rainfall probabilities for Koßdorf (left) and Nordhausen (right), data 1973-2000

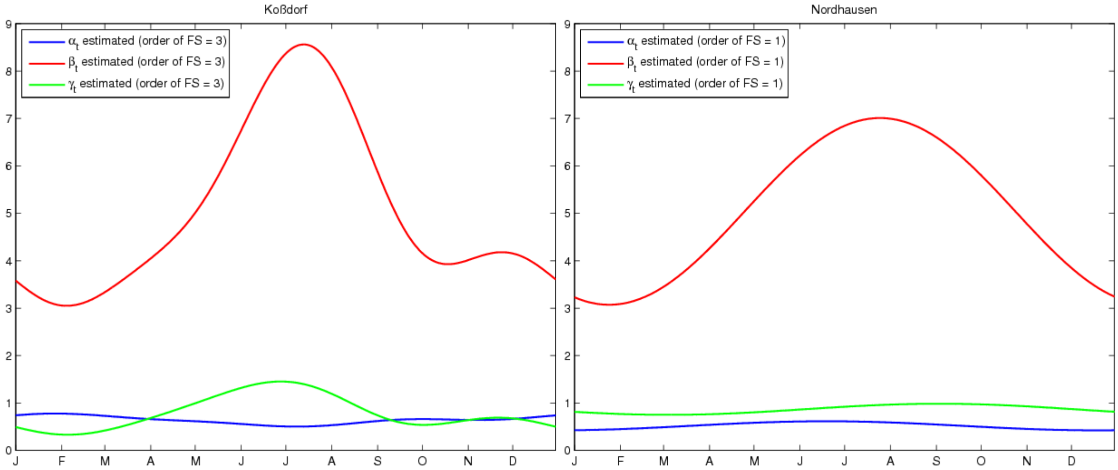


Figure 5. Estimated parameters for the mixed exponential distributions for Koßdorf (left) and Nordhausen (right), data 1973-2000