INTERNATIONAL JOURNAL OF CLIMATOLOGY
Int. J. Climatol. **28**: 329–342 (2008)
Published online 5 June 2007 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/joc.1542

RMetS
ROYAL METEOROLOGICAL SOCIETY

Predicting the regional onset of the rainy season in West Africa

P. Laux,[a]* H. Kunstmann[a] and A. Bárdossy[b]
[a] *Institute for Meteorology and Climate Research (IMK-IFU), Forschungszentrum Karlsruhe, 82467 Garmisch-Partenkirchen, Germany*
[b] *Institute for Hydraulic Engineering, University of Stuttgart, 70569 Stuttgart, Germany*

ABSTRACT: Particularly in regions, where precipitation is limited to a few months per year only, reliable determination of the onset of the rainy season and the start of the sowing time is of crucial importance to sustainable food production. Especially since the mid-1980s, an increasing delay of onset dates in the Volta basin of West Africa has been suspected by local farmers. To investigate this speculation and develop a reliable tool to find the optimal sowing date, the onset of the rainy season in the region was analysed by means of several statistical techniques. The focus was put on the region of the Volta basin in Ghana and Burkina Faso.

In a first step, two fuzzy logic based definitions of the onset were developed using daily precipitation data and additionally accounting for important plant physiological aspects. In this context, only one definition is potentially useful to judge whether the current onset of the rainy season has already begun. In a second step, methods for predicting the onset date of the ongoing season were investigated. In this context, the detection of onset controlling variables plays a major role. Two strategies are investigated and evaluated for the prediction of the monsoon's onset dates:

1) A combination of regionalized synoptic rainfall data by means of principal component analysis (PCA) in a spatial mode and linear discriminant analysis in order to detect reliable prediction parameters and allow for a classification of the rainy season, dry season, and the onset of the rainy season using current rainfall data.
2) Linear regression models were generated to estimate the onset of the rainy season for certain regions using the onset dates of regions, where the onset has already begun.

To enhance the predictability, optimized definition parameterisation in the field of both strategies was applied.

KEY WORDS onset of the rainy season; West Africa; Volta basin; farming management strategies; decision support system; linear discriminant analysis; principal component analysis; fuzzy logic

Received 1 August 2006; Revised 15 March 2007; Accepted 18 May 2007

1. Introduction

In the Volta basin with its semi-arid to sub-humid climate from north to south, rainfall is one of the most critical factors for ecological and environmental processes. Especially in areas, where most of the agricultural production depends on rainfall, the amount of water available to plants strongly depends on the rainy season's onset, length, and end (Ati *et al*., 2002). According to Steward (1991), the onset is the most important variable to which all the other seasonal variables are related. Owing to a very high spatial and temporal variability of precipitation amounts and a non-uniform distribution of the rains during the rainy season, local farmers have problems to decide when to start with the sowing preparations. Therefore, the farmers have developed a range of strategies to cope with rainfall variability depending on site conditions. These strategies include:

(1) Exchange of information on rainfall by seasonal workers, (2) time-dependent measures, such as dry seeding, resowing, and the use of differently maturing crop varieties, (3) coping with spatial rainfall variability through the cultivation of large and widely dispersed field areas, and (4) measures for sustaining soil fertility (Graef and Haigis, 2001).

Traditional decision criteria, e.g. the observation of the behaviour of some insects or birds or flowering of certain trees, seem to fail more frequently and the demand for a scientific decision aid increases among local farmers (Roncoli *et al*., 2002). According to Sultan and Janicot (2000, 2003) and Sultan *et al*. (2003), the West African monsoon dynamics follow two distinct phases, the pre-onset and the onset phase. The pre-onset occurs in late spring when the intertropical convergence zone

* Correspondence to: P. Laux, Institute for Meteorology and Climate Research (IMK-IFU), Forschungszentrum Karlsruhe, Kreuzeckbahnstraße 19, 82467 Garmisch-Partenkirchen, Germany.
E-mail: patrick.laux@imk.fzk.de

(ITCZ) establishes itself at 5° (∼14th may), whereas the actual onset occurs when the ITCZ shifts abruptly northwards (∼24th june). Then, the ITCZ moves from 5° to 10°N, where it stays for the whole August. Following the movement of the ITCZ, the monsoon is of bimodal character in the Guinea region and of single mode in the Sahel. In these zones, the onset seldom occurs abruptly and is often preceded by short isolated showers with intermittent *dry spells* of various lengths, which are often mis-interpreted as the start of the rains *(false starts)*. Prolonged dry spells of two or more weeks after sowing are disastrous for plants, because they dry out top soil layers and prevent germination, which may lead to total crop failure or yield reduction. Thus, survival of the seedlings is the key point for agriculturists (Sultan and Janicot, 2003). For sowing, it is important to know, whether (1) the rains are continuous and sufficient to ensure enough soil moisture during planting time and (2) this level will be maintained or even increased during the growing period to avoid crop failure (Walter, 1967).

Therefore, it is essential that the most important variable, the onset of the rainy season, which coincides with the start of the growing season, is predicted online, i.e. for the ongoing or forthcoming season, on the basis of reliable scientific methods (Ati *et al*., 2002). Knowledge of the onset, cessation, and, thus, of the length of the growing/rainy season significantly supports the timely preparation of farmland, mobilisation of seed/crop, manpower, and equipment, and it will also reduce the risk of planting and sowing too late or too early (Omotosho *et al*., 2000).

Owing to the random distribution of local convection events with high precipitation rates and potential shifts of the onset dates on site scale, this article will concentrate on the determination and prediction on the regional scale. It is focussed on the first/major season's onset for the region of the Volta basin. The analysis presented will be based mainly on daily precipitation records of 29 synoptic observation stations of the Burkinabé and Ghanaian Meteorological Services. Regionalising is achieved by means of principal component analysis (PCA) in the spatial mode. The result is the zoning of five areas with high within and low between similarities of the rainfall characteristics. A comparison of the mean rainfall patterns and the climatic needs of the crop may help to judge whether the actual crop regions are optimally distributed under consideration of soil aspects.

Various definitions of the onset of the rainy season (monsoon) are currently in use. The principal research areas are West Africa, India, and Australia, i.e. areas, where water availability is scarce and limited to the rainy season. In general, two categories of definitions can be distinguished: those relying on parameters measured on the surface and those using atmospheric dynamics. The first group consists of a huge number of methods, mostly applied for agro-climatological purposes on the local scale. If the ultimate goal is the prediction of the onset and not just monitoring, it has to be defined on the regional scale (Camberlin and Diop, 2003). For the Australian monsoon, there are wind-only (e.g. Holland, 1986) and rain-only definitions (e.g. Nicholls, 1984) and a mixture of them (e.g. Troup, 1961; Hendon and Liebmann, 1990a). The second group of definitions is represented by authors like. Davidson *et al*. (1983), who analysed the appearance of certain large-scale circulation patterns in combination with the start of the rains. For West African regions, also a huge number of definitions are known. Ati *et al*. (2002) gave a good overview of various approaches. Some authors employed a rainfall-evaporation model (e.g. Benoit, 1977), others use surface pressure, temperature, and relative humidity (e.g. Omotosho, 1990, 1992). Most scientists, however, refer to precipitation itself in order to determine the onset and/or cessation of the rainy season (e.g. Walter, 1967; Ilesanmi, 1972; Ati *et al*., 2002). The advantage of this approach is that precipitation totals are readily available and it exhibits the most direct relationship rather than some other related factors. Methods using air temperature have not been used widely in West Africa, because of its uniform distribution and the small intra-seasonal variations. In fact, precipitation and not temperature is seen as the most critical factor in tropical agriculture (Stern *et al*., 1981). For rainfall-alone definitions, two further subcategories can be found in literature, a definition based on a certain threshold value (e.g. Stern *et al*., 1981) and a relative definition using a proportion relative to the total amount (Ilesanmi, 1972).

The overall goal of this investigation is twofold:

1) The development of a reasonable onset definition especially for the Volta basin on the regional scale.
2) The prediction of the major rainy season's onset for the ongoing season by means of the linear discriminant analysis and linear regression technique.

Section 2 of the present paper will cover data availability and the methodology. A fuzzy-logic approach for two different onset definitions was used to determine previous onset dates, with important plant physiological aspects being taken into account (Section 2.2). Section 2.3 will deal with the geographic assessment of the regional onset dates and Section 2.4 will describe their mean and transient features in the period 1961–2001. After this, linear discriminant analysis will be applied to verify the definition (Section 2.5). Additionally, this method may be used to classify the ongoing or forthcoming season. In Section 2.6, a predictive tool for regional onset definition using definition optimisation shall be presented. Results and limitations of this methodology will be described in this section. The summary and conclusions will be given in Section 3.

2. Methodological approach

2.1. Data Set

For the statistical analyses performed, daily precipitation time series of Burkina Faso and Ghana were applied.

DOI: 10.1002/joc

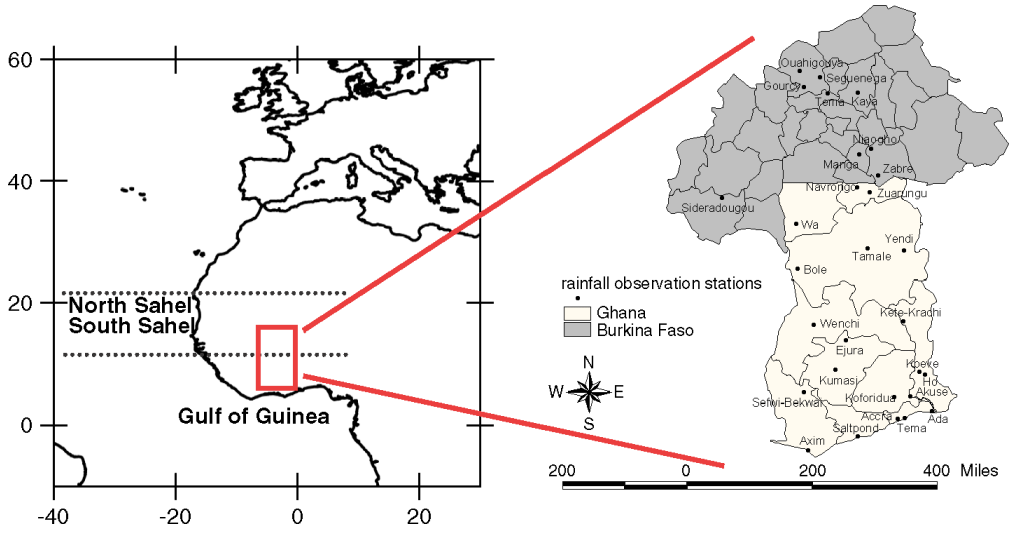

Figure 1. Spatial distribution of rainfall observation sites in the Volta basin.

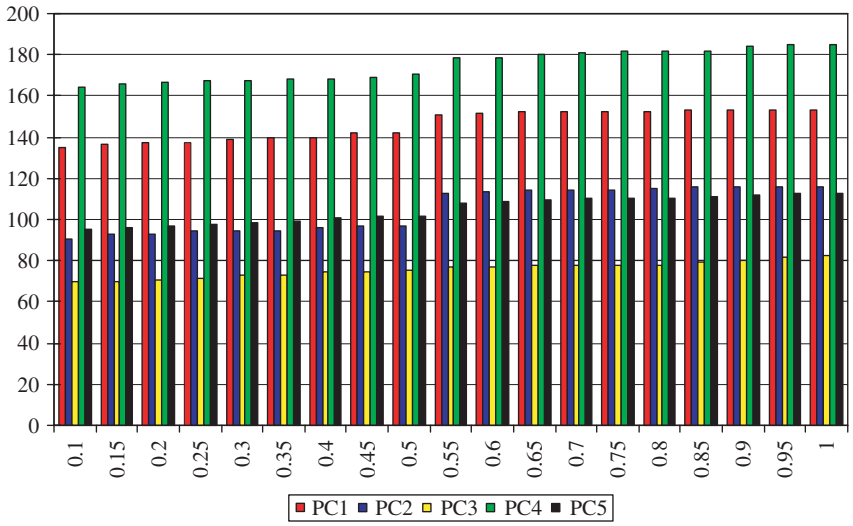

Figure 2. Mean onset dates of PC1–PC5 with varying γ value from 0.1 to 1.0 with an increment of 0.5.

The meteorological data were obtained from the *Institute Nationale de l'Environnement et des Recherches Agricoles (INERA)* at Ouagadougou (Burkina Faso), the Meteorological Service of Burkina Faso in Ouagadougou, and the *Meteorological Service Department* in Accra (Ghana). The daily data applied were checked for continuity and plausibility by calculating monthly and annual totals and cross-checking them with neighbouring stations. This work has been carried out by the Ghanaian and Burkinabé meteorological services. Owing to large data gaps in most of the observation time series, only a limited number of the meteorological observation stations available could be used. Figure 1 shows the spatial distribution of the used synoptic meteorological stations that offered continuous daily rainfall data from 1961 to 1999.

2.2. Fuzzy-logic approach to determining the rainy season onset

The definition, as used in this paper, belongs to the first group of rainfall-alone definitions with site-scaled threshold values. It was established by Stern *et al.* (1981) and modified later on by Sarria-Dodd and Jolliffe (2001) for Burkina Faso, because it tended to supply onset dates which were too late to be reasonable. Stern *et al.* (1981) considered the onset to be the first date of year, on which the following three constraints are valid simultaneously:

(1) A total of at least 25 mm of rainfall are observed within a 5-day period;
(2) The starting day and at least two other days in this 5-day period are wet (at least 0.1 mm rainfall recorded);

DOI: 10.1002/joc

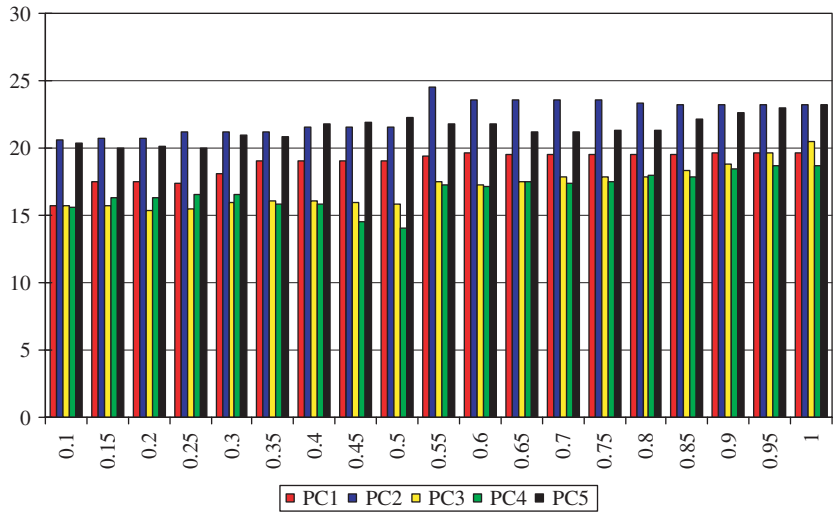

Figure 3. Standard deviation of the onset dates of PC1–PC5 with varying γ value from 0.1 to 1.0 in an increment of 0.5.

(3) No dry period of seven or more consecutive days occurring in the following 30 days.

Since this definition did not yield an onset date for every year due to the sternness of its constraints which have to be fulfilled simultaneously, a fuzzy logic approach has been established to facilitate modelling. Each definition constraint is attached to a fuzzy membership function (Figure 5) using triangular (subscript T) fuzzy numbers. Concerning e.g. the first constraint dealing with the total amount of rainfall within a 5-day period the triangular fuzzy numbers are $(18, 25, +\infty)_T$. This means that the membership grade of rainfall totals minor than 18 mm is attached to zero and totals larger than 25 mm to unity. Between 18 and 25 mm the membership grade is linearly interpolated. Then the onset date is defined as the first day of year where the product $\gamma = \gamma 1 \times \gamma 2 \times \gamma 3$ exceeds a defined threshold value (hereinafter referred to *Definition 1*). If the calculated onset date is matching the beginning of a period with high rainy day frequency within the year, the date can be regarded as reasonable. In this context sensitivity analysis to clarify the influence of γ were performed, separated for each region. Figure 2 is depicting the mean onset dates and Figure 3 the standard deviations with varying γ value from 0.1 to 1.0 in an increment of 0.5. A drastically delay of the mean onset dates as well as an increase of the standard deviations using $\gamma \geq 0.55$ can be observed, which would decrease the growing time. In turn, planting too early should also be avoided. Therefore, a threshold value of 0.4 is proposed, which is delivering reasonable onset dates for all regions within the Volta basin. *Definition 1* is only applicable for ex-post determination of the onset, because of the latter definition constraint, which represents a false start criterion. A false start occurs if the ex-post determination of the onset date using *Definition 1* overrules the onset date determined by *Definition 2*. Figure 4 is illustrating the monthly number of false starts for each region within the period 1961–2001. For predicting the onset of the ongoing season, however, this definition approach is not useful, as it would require a weather forecast of the following month due to the $\gamma 3$ constraint. For this reason, a second definition was applied, with only the fuzzy logic approach of $\gamma 1$ and $\gamma 2$ being used (hereinafter referred to *Definition 2*). In the following sections, the focus of our analysis will be on *Definition 2*.

2.3. Geographic assessment of rainy season onset dates using the principal component analysis

Regionalisation from the site to the regional scale is achieved by means of varimax rotated PCA using daily precipitation data of the whole year based on the variance-covariance matrix to generate robust and physically interpretable components (Richmann, 1993). The variance-covariance matrix instead of the correlation matrix was used to account for the rainfall distribution

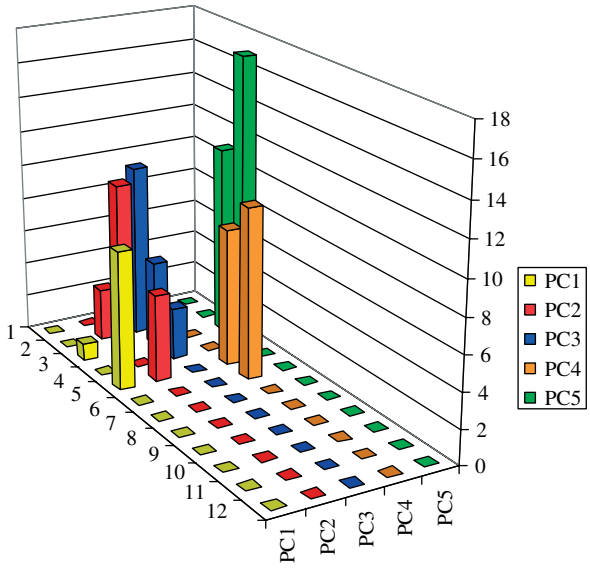

Figure 4. Number of false starts for each region and month (1961–2001).

DOI: 10.1002/joc

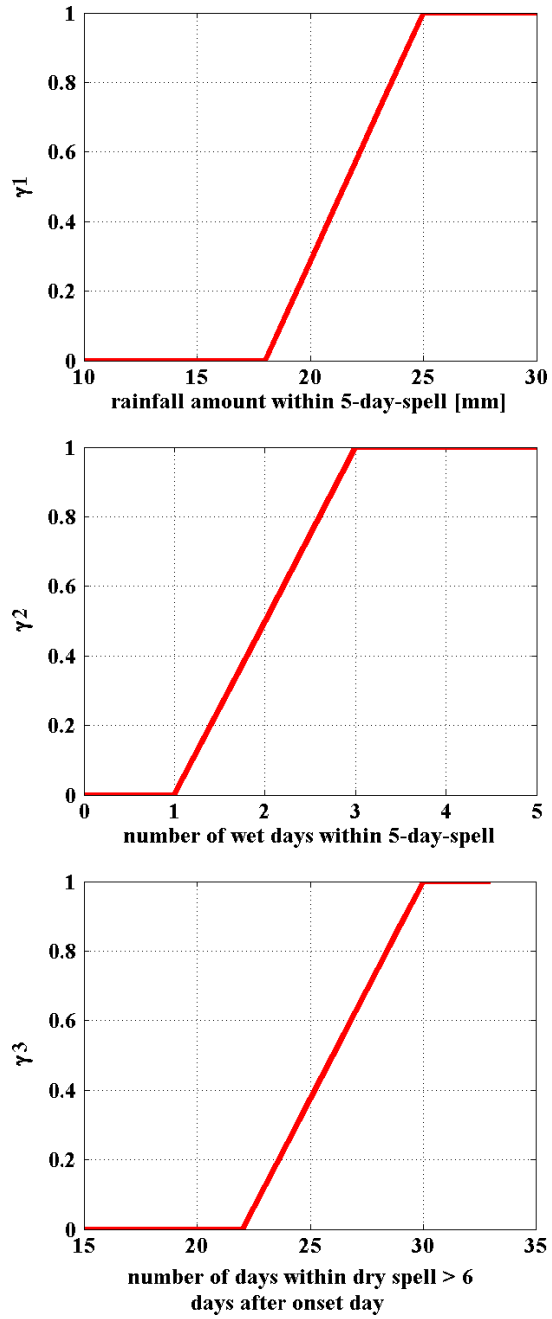

Figure 5. Membership functions representing the three definition constraints. The first day of the year, where the product of all membership grades $\gamma 1$, $\gamma 2$ and $\gamma 3$ exceeds a defined threshold value is regarded as the rainy season's onset. This figure is available in colour online at www.interscience.wiley.com/ijoc

rather than rainfall totals. Generally, PCA is used to study the covariances (or correlations) of variables in order to reduce their number by extracting only a few components which account for most of the variance of the original variables. The principal components (PCs) or factors are an independent set of linear combinations of the variables that redefine the existing variable space. The eigenvalues denote that variability that is explained by the respective factor in the principal component space. In order to reduce the components to the most important ones, only eigenvalues greater than unity were selected, thus satisfying Kaiser's criterion (Kaiser, 1958). The results were five different PCs owning high within and low between similarities, which explained about 60% of the total daily precipitation variance.

To obtain the spatial distribution of these synthetic variables, correlation analysis between the PC and all original observation data was performed. Each station was then related to the PC showing the largest correlation coefficient. By grouping all stations to the PCs, spatial information is obtained on the positions of these synthetic variables. Figure 6 shows the spatial aggregation of the observation sites with similar rainfall characteristics to five different regions. Figure 7 illustrates the smoothed long-term mean (averaging period from 1961 to 2001) rainfall per Julian day for the five regions represented by the five PCs. Low slopes of the curves (left side) reflect mean periods of drier spells, whereas high slopes characterize rain spells. For the PC 2 ($\sim$ from Julian day 180–270) and PC3 ($\sim$ from Julian day 230–260) regions, a mean dry spell can be observed, followed by a second little rainy season which is much stronger for PC3 than for PC2. The respective right figures denote the 1st derivation that highlights the annual distribution of precipitation. As far as the problem of determining the commencement of the rainy season is concerned, these graphs help to detect small rainy spells around the real mean onset date, which are often mis-interpreted as the onset (mean *false starts* are exemplarily shown for PC2). Almost all the regions show rainy spells of gradually different magnitude before the real onset. This knowledge may help to narrow the time frame of the real onset and, hence, to avoid planting too early.

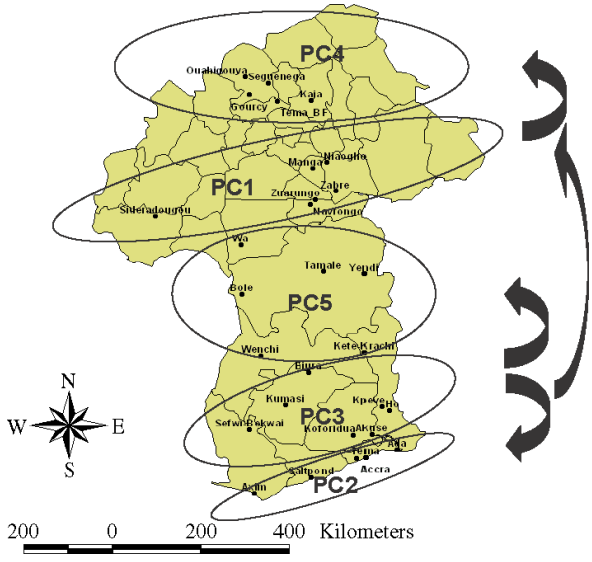

Figure 6. Spatial location of the five different regions (ellipses) corresponding to the principal components calculated by 29 synoptical rainfall stations with daily values from 1961–2001. The arrows represent the direction for predicting the rainy season's onset of one region using the current onset date of another region. This figure is available in colour online at www.interscience.wiley.com/ijoc

DOI: 10.1002/joc

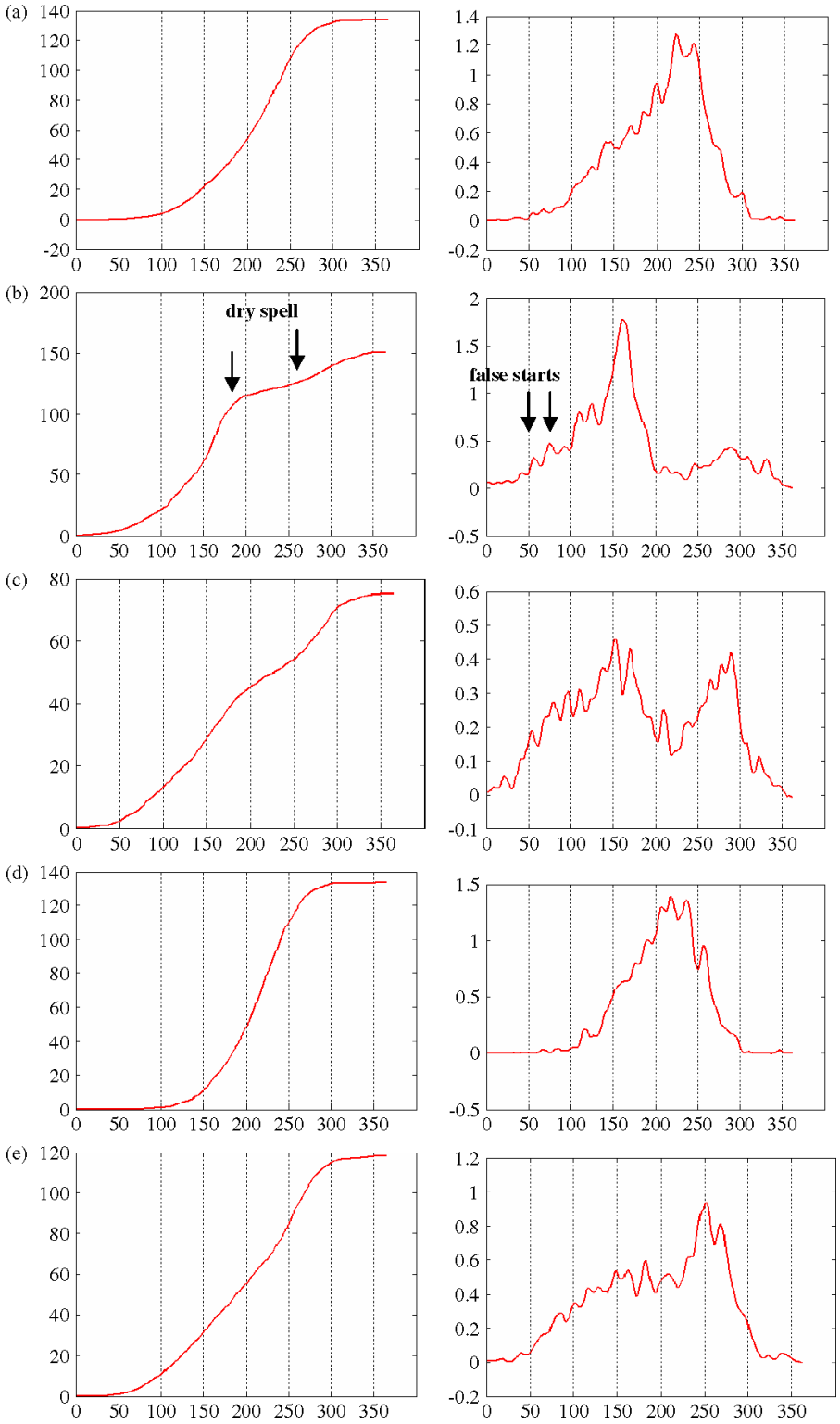

Figure 7. Mean precipitation behaviour of the five regions, represented by the five principal components: a) PC1, b) PC2, c) PC3, d) PC4, and e) PC5. Left: cumulated mean values of rainfall (mm) of all observation sites grouped into a certain region after smoothing using Savitzky–Golay algorithm. Right: 1st derivation of the cumulated mean rainfall amount (mm/d), (averaging period: 1961–2001). In the context of this work a *dry spell* refers to a period of time (more than 1 week) when an interruption of the rainy season takes place whereas a *false start* denotes the misinterpretation of the rainy seasons' onset according to the used onset definition. This figure is available in colour online at www.interscience.wiley.com/ijoc

2.4. Analysis of past onset dates and mean precipitation of the different rainfall regions

The mean values of all observation sites associated to one PC were used as input for *Definition 1* and *Definition 2* in order to derive past onset dates. Linear regression analysis and the F-test were applied to compute trends

DOI: 10.1002/joc

of the onset dates and their statistical significance. In the linear regression model, the dependent variable y is assumed to be a linear function of an independent variable x plus an error ε introduced to account for all other factors:

$$y_i = b_0 + b_1 x_i + \varepsilon_i \quad (1)$$

Using the '*least squares methodology*' (Middleton, 2000) - taking the line which minimizes the sum of squares of the error term in Equation (1) -, the fitted line is:

$$\hat{y}_i = b_0 + b_1 x_i \quad (2)$$

where $\hat{y}_i$ are the values of y estimated from the fitted line and ε are the differences between the observed values of$y$and these estimated values. b_0 is the intercept of the y axis and b_1 is the slope of the regression line and, simultaneously, the trend value of the dependent variable. In order to test the null hypothesis when the slope $b_1 = 0$, it is useful to divide the total variance $s_y^2 = \sum(y_i - \overline{y})^2/(N-1)$ of the dependent variable into two independent parts: the variance due to the regression and the error variance $\sum \varepsilon^2/(N-2)$. These calculations are indicated best by the *Analysis of Variance* (ANOVA) table for linear regression. Under the null hypothesis of there being no effect due to the regression, the mean squares due to regression and the mean squares due to the variance result in the variance due to the error. Hence, the MSR/MSE ratio is expected to follow the F distribution, and the calculated value can be compared with the tabulated value for a certain level of significance and 1 and ($N-$ 2) degrees of freedom. If the calculated value is larger than the tabulated value, the null hypothesis is rejected and the regression is seen to be significant. Figure 8 shows the onset dates from 1961 to 2001 using both definitions and their respective linear regression functions. Generally, the rains commence in the southernmost areas and propagate northwards under the influence of the migration of the ITCZ, moving back and forth across the equator in a semi-annual cycle following the sun's zenith point.

Table I lists the mean onset date, standard deviation, linear trend, and its significance for each region. A classification of the significance is offered in Table II. In comparison to *Definition 2, Definition 1* yields later mean onset dates and lower standard deviations. Differences in the standard deviations are plainest for the northernmost regions of PC1 und PC2. Generally, *Definition 1* yields more significant trends than *Definition 2*. However, *Definition 2* produces the highest trend for PC1 (0.88 days/year), namely, an onset date strongly delayed by 35 days per 40 years. When looking at the time series of the onset dates, the smallest agreement of onset dates using both definitions is reached with PC3 and PC4. For PC4, for instance, the onset in 1963 takes place on Julian day No. 181 according to *Definition 1*. When applying *Definition 2,* the onset takes place 60 days earlier.

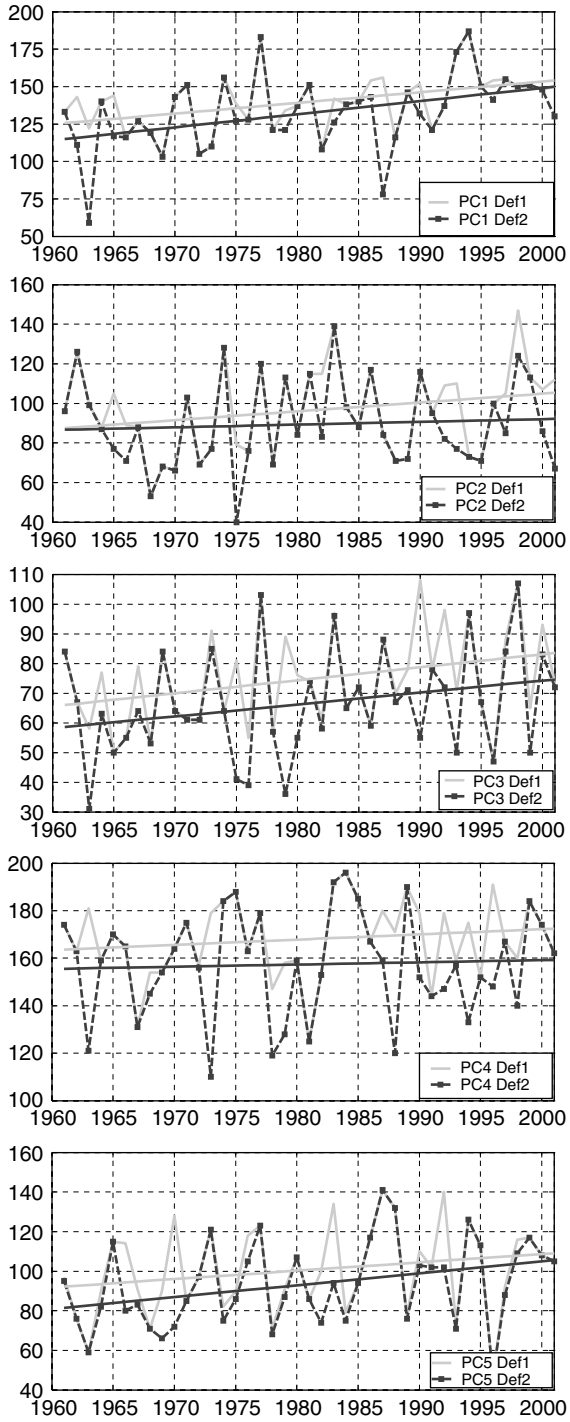

Figure 8. onset dates (Julian day) of the major rainy season using *Definition 1* and *Definition 2* (dashed lines) and linear trend lines (solid lines). The features of both definitions can be found in Section 2.3.

To find out which γ criterion mainly is responsible for the positive trends (delayed onset dates) and since our focus is lying on *Definition 2*, the relationship between the two γ criteria was investigated and linear trends of the amount of regional precipitation and the number of wet days were analysed. For this purpose, monthly mean values and monthly resolved values of the rainfall amount ($\gamma 1$) and the number of wet days

DOI: 10.1002/joc

Table I. Mean onset dates, standard deviations, trends, and their levels of significances for which the null hypothesis is being rejected of the regions (see Figure 6) using *Definition 1* and *Definition 2* (see Section 2.3).

Definition 1	Mean onset (Julian day)	Standard dev. (days)	Trend (days/year)	Signif. (%)
PC1	140	19	0.71	99
PC2	96	22	0.45	88
PC3	75	16	0.44	96
PC4	168	16	0.22	70
PC5	101	22	0.42	99
Definition 2	**Mean onset (Julian day)**	**Standard dev. (days)**	**Trend (days/year)**	**Signif. (%)**
PC1	132	24	0.88	99
PC2	89	22	0.14	36
PC3	67	18	0.40	91
PC4	157	22	0.09	25
PC5	93	21	0.60	99

Table II. Criteria for judging the level of significance P for which the null hypothesis is being rejected.

$P < 80\%$	Not significant
$80\% \leq P < 90\%$	Slightly significant
$90\% \leq P < 99\%$	Moderately significant
$P \geq 99\%$	Highly significant

($\gamma 2$) were computed. Table III presents the correlation coefficient between the amount of mean monthly rainfall and the mean monthly number of wet days. Except for the region corresponding to PC2, all correlations are very high. Heavy convective showers, associated with high amounts of rainfall and simultaneously a slightly increasing number of wet days, lead to lower correlation coefficients during the rainy season, especially in the mean onset month. Figure 9 exemplarily shows the monthly resolved correlation coefficient of the amount of rainfall and the number of wet days for the region PC1. In this case, the mean onset month is May with a correlation coefficient of ~0.2. Again, trend analysis of monthly precipitation amounts and wet days was applied using linear regression analysis and verified by the F–test. Table IV summarizses the results of the trend analysis for all regions. Significant values ($\alpha < 0.05$) and very significant values ($\alpha < 0.01$) are highlighted with light and dark grey colour. Both the number of rainy days and the precipitation amount reveal a number of significant negative trends. Overall, PC5 exhibits the most significant trends ($P < 0.01$) in the number of wet days. As regards the mean onset dates, it is presumed that for PC1, both variables might be responsible for the delayed onset. They exhibit significant and strong negative trends in April, a formerly frequent rainy season's onset month. PC2 with its normal onset dates around March displays a significant negative trend of the number of rainy days in this month and significant negative trends of precipitation one month earlier and later. PC3 having the earliest onset month with frequent starting dates in February also exhibits significant trends with high amounts of decreasing rainy days and rainfall in this month. Regarding PC4, the number of rainy days in April (trend = −0.05 days/year) seems to play a major role for the onset shift. In PC5, both parameters are important again (high negative trends in March). These results are in good agreement with the trend analysis of the rainy season's shift using *Definition 2*.

Table III. Correlation coefficients of monthly mean rainfall amount and monthly mean number of wet days (averaging time period: 1961–2001). PC1 to PC5 are representing different regions in the Volta basin (Figure 7). All coefficients are highly significant.

PC1	0.95
PC2	0.77
PC3	0.91
PC4	0.96
PC5	0.97

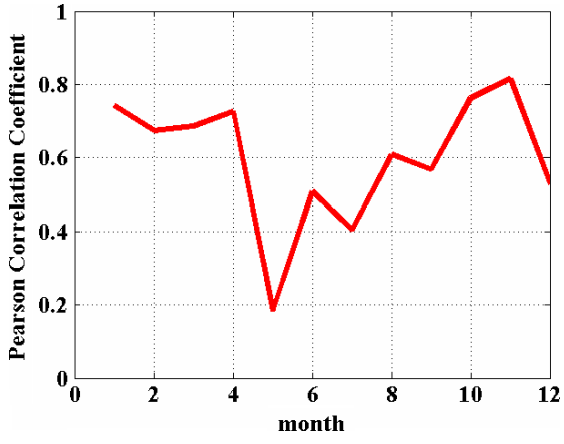

Figure 9. Monthly resolved correlation coefficient of rainfall amount and number of wet days for the region corresponding to PC1 (1961–2001). This figure is available in colour online at www.interscience.wiley.com/ijoc

DOI: 10.1002/joc

Table IV. Monthly trends of number of rainy days (R.D.) [days/year] and precipitation amounts (P.A.) [mm/year] for all regions. Significant results with an error probability $\alpha = 5\%$ and $\alpha = 1\%$ are highlighted with light and dark grey.

	PC1		PC2		PC3		PC4		PC5	
Month	R.D.	P.A.	R.D.	P.A.	R.D.	P.A.	R.D.	P.A.	R.D.	P.A.
1	−0.01	−0.06	−0.05	−0.44	−0.03	−0.27	–	–	−0.02	0.13
2	−0.03	−0.23	−0.04	−0.61	−0.14	−0.65	−0.01	−0.02	−0.12	−0.18
3	−0.06	−0.14	−0.08	0.03	−0.07	−0.42	−0.01	–	−0.2	−0.56
4	−0.08	−0.71	−0.06	−0.66	−0.04	0.05	−0.05	−0.21	−0.12	0.14
5	0.01	−0.15	−0.06	0.11	−0.06	−0.73	−0.06	0.14	−0.12	−0.28
6	0.02	0.28	−0.18	−6.85	−0.1	−1.04	0.01	−0.5	−0.13	0.41
7	0.02	−0.52	−0.13	−2.07	−0.06	−0.65	−0.01	−0.36	−0.13	−0.68
8	−0.02	0.75	−0.06	−0.6	0.04	−0.26	−0.04	−0.53	−0.11	−0.6
9	−0.04	−0.3	0.02	−0.38	−0.01	−0.32	−0.07	−0.39	−0.11	−0.16
10	−0.01	−0.02	–	0.36	–	−0.68	0.01	−0.14	−0.1	0.34
11	−0.05	−0.09	−0.14	−0.65	−0.1	−0.77	–	0.01	−0.13	−0.16
12	−0.02	0.02	−0.12	−0.22	−0.05	−0.02	−0.02	−0.04	−0.06	−0.2

$\alpha = 0.05$
$\alpha = 0.01$

2.5. Predicting the regional rainy season's onset using linear discriminant analysis and simple rainfall indices

Linear discriminant analysis (*LDA*) was carried out to allocate each day to one of the four classes: (1) dry season, (2) transition, (3) onset of the rainy season, and (4) wet season. The transition class was introduced in order to check if there is rather an abrupt or a gradual onset according to the discriminant model and the onset definition. A poor hit ratio for this transition class and reclassification of these days to the dry seasons' class was expected to guarantee a good discrimination between onset and dry season. A similar approach was applied by Sarria-Dodd and Jolliffe (2001) to derive a linear discriminant function that distinguishes between 'Real' and 'False' starts of the onset of the rainy season for Burkina Faso and Mali. LDA generally is a very useful tool for detecting the variables that enable the researcher to discriminate between two or more groups and for assigning cases to different groups with a better-than-chance accuracy.

Two pre-requisites have to be noted: the number of groups must not exceed the number of variables describing the data set and the groups must have the same covariance structure. Given a sample $X = (x_1, x_2, \ldots, x_p)$ of n observations on a vector of p variables from g populations $\zeta 1$, ..., ζg, the linear discriminant function(s) Y are of the form:

$$Y_{1,\ldots,nu} = \sum_{k=1}^{p} \upsilon_k X_k \tag{3}$$

with nu is being the number of discrimination functions Y. The number of functions is the minimum of $(g-1)$, where g is the number of categories in the grouping variable, or p, the number of discriminating (independent) variables. X_k is denoting the variables describing the data set. The discrimination coefficients υ_k have to be determined such that the discrimination between the groups is best. Determination of the coefficients of the discriminating function is quite simple. In principle, the discriminating functions are formed such that the separation (= distance) between the groups B is maximized, the distance within the groups W is minimized, and, hence, the ratio (= discriminant criteria) of B and W is maximized. B, W, and the overall weighted mean $\overline{\mu}$ and the population mean μ_j are computed as follows:

$$B = \sum_{j=1}^{g} \pi_j (\mu_j - \overline{\mu})(\mu_j - \overline{\mu}) \tag{4}$$

$$W = \sum_{j=1}^{g} \sum_{i \in j} \pi_j (x_i - \mu_j)(x_i - \mu_j) \tag{5}$$

$$\overline{\mu} = \frac{1}{n} \sum_{j=1}^{g} \pi_j \mu_j \tag{6}$$

$$\mu_j = \pi_j \sum_{i \in j} x_i \tag{7}$$

πj where:

x_i Prior probability of an observation belonging to the group j

Number of observations, $i = 1, \ldots, n$

The overall goal is to find predictor variables for all regions, which are reliable in representing the onset dates. These potential predictor variables are simple rainfall indices that describe the *rainfall amount* (VRI_x) and *number of wet days* (VRA_x) 30, 25, 15, 10, and 5 days before the potential onsets and γ values, respectively. In this case, three discriminant functions (not shown) are required. Similar to multiple regression analysis, not all the candidate predictors contribute to a significant improvement of the discriminant model and there is no established method for assessing which predictors

DOI: 10.1002/joc

Table V. Confusion matrix for PC1 to assess the performance of the discrimination using the most suited variables and γ threshold with the main focus on the onset class (PC1 is representing the area illustrated in Figure 6).

$\gamma = 0.5$		Class membership after application of linear discriminant analysis [%]			
		Dry season	Transition	onset	Wet season
Predetermined class membership (%)	Dry season	**75.07**	16.28	2.23	6.42
	Transition	54.58	**22.74**	5.19	17.49
	onset	12.08	8.07	**63.24**	16.61
	Wet season	16.02	12.17	11.96	**59.85**

Table VI. Confusion matrix for PC2 to assess the performance of the discrimination using the most suited variables and γ threshold with the main focus on the onset class (PC2 is representing the area illustrated in Figure 6).

$\gamma = 0.45$		Class membership after application of linear discriminant analysis [%]			
		Dry season	Transition	onset	Wet season
Predetermined class membership (%)	Dry season	**76.02**	15.44	2.50	6.04
	Transition	53.39	**22.03**	7.42	17.16
	onset	15.31	10.52	**55.98**	18.19
	Wet season	18.61	12.10	10.61	**58.68**

Table VII. Confusion matrix for PC3 to assess the performance of the discrimination using the most suited variables and γ threshold with the main focus on the onset class (PC3 is representing the area illustrated in Figure 6).

$\gamma = 0.5$		Class membership after application of linear discriminant analysis (%)			
		Dry season	Transition	onset	Wet season
Predetermined class membership (%)	Dry season	**83.37**	14.68	1.24	0.71
	Transition	50.72	**38.10**	3.47	7.71
	onset	8.09	11.30	**70.91**	9.70
	Wet season	5.38	6.95	8.43	**79.24**

Table VIII. Confusion matrix for PC4 to assess the performance of the discrimination using the most suited variables and γ threshold with the main focus on the onset class (PC4 is representing the area illustrated in Figure 6).

$\gamma = 0.4$		Class membership after application of linear discriminant analysis (%)			
		Dry season	Transition	onset	Wet season
Predetermined class membership (%)	Dry season	**80.32**	15.86	1.38	2.44
	Transition	52.99	**27.97**	4.01	15.03
	onset	7.20	5.80	**79.44**	7.56
	Wet season	13.94	11.50	10.57	**63.99**

are useful when the assumption of equal variance is dropped (Mason, 1998). Typically, stepwise multiple regression analysis is a screening method to choose the variables for the discriminant model (Ward and Folland, 1991). According to these results (not shown here) four parameters were selected: VRI25, VRA5, and both the $\gamma 1$ and $\gamma 2$ values. Starting from the computed onset date, 40–10 days before this date were taken as *dry season* class, 9–1 day before as *transition*, the date itself and another 2 days as *onset,* and 15–30 days afterwards as *wet season*. Owing to the lack of an adequate measure of model performance, like the R^2 statistic for the regression analysis, validation is a very important component of model construction. Jack–knife validation was used in order not to reduce the number of observations for the training of the model. Tables V–IX present the confusion matrices of the classification using the four above-mentioned variables and optimal γ value

DOI: 10.1002/joc

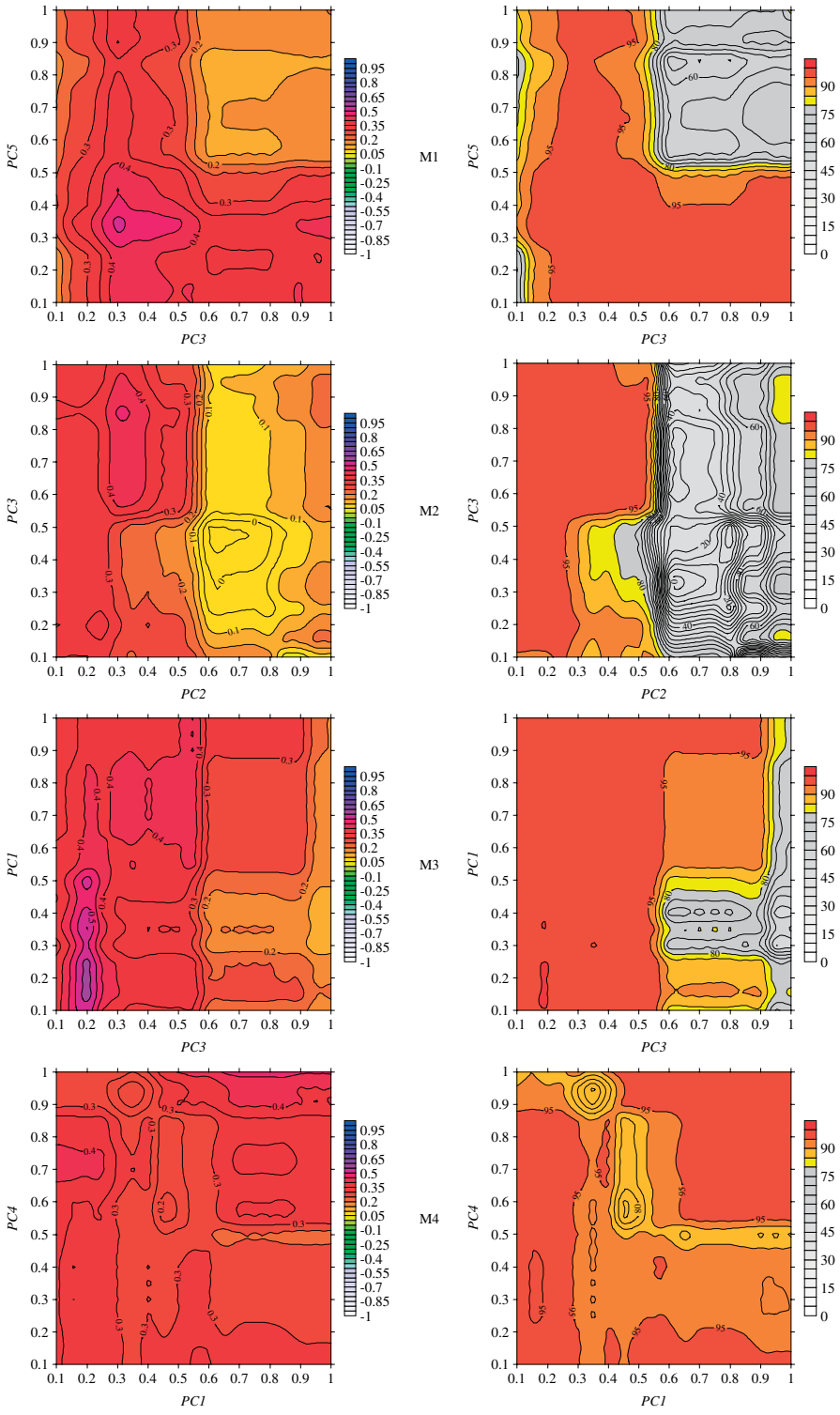

Figure 10. Correlation coefficients between the onset dates of two regions using varying γ values from 0.1 to 1.0 (left column) and level of significance (right column) for the selected models (M1, M2, M3, M4).

for *definition 2*. The percentage of cases on the diagonal depicts the percentage of correct classification (hit ratio). The number of hits should be compared with the number expected by chance: The prior probabilities for the four classes are: 52, 15, 7, and 27%.

It may be summarized that for almost all regions (except for the region corresponding to PC2), the hit ratio for the *onset* class ranges between 60 and 80% with an a priori probability of 7%. Also the classes of *dry season and wet season* show satisfactory results. The fact that the *transition* class cannot be discriminated well indicates that the definition for the onset operates successfully. Most of the cases of this class were reassigned to the dry seasons' class.

DOI: 10.1002/joc

Table IX. Quality parameters of the four regression models and test of stochastic independence and normal distribution of the residuals (for the model numbers see Table XI).

Model number	R^2	Signif. (%)	Residuals D–W	K–S test (asymp. sig.)
M1	0.27	99	1.58	0.64
M2	0.21	99	2.14	0.47
M3	0.33	99	1.57	0.97
M4	0.22	99	2.18	0.27

Table X. Test for normality of the data at the 95% significance level ($\alpha = 0.05$) in order to declare the significance of the correlation coefficients. The result is '1', if the mutual distributions follow a normal distribution, '−0.5' if the data of the target PC is normally distributed and the data of the independent PC is not normally distributed. For the case that the data of the target PC is not normally distributed, but the data of the independent PC is normally distributed, the result is '0.5', and otherwise the result is '0' (no normal distribution of the data of both PCs).

Model	γ target	γ independent	K–S test	Lilliefors test	Jarque–Bera test
M1	0.35	0.3	1	1	1
M2	0.3	0.85	1	1	1
M3	0.2	0.2	1	1	1
M4	1	0.8	1	1	−0.5

2.6. Predicting the regional rainy season's onset using linear regression analysis and optimized γ values

Assuming PC3 to be the first rainy season's onset region of the year and knowing the date, linear regression models can be generated to predict successively the onsets of PC2, PC5, PC1, and PC4. The arrows in Figure 6 represent the direction of regression modelling. In order to enhance the performance of the regression models, the best γ threshold combination of two regions has to be found. For this purpose, the linear correlation coefficient between the onset dates of two regions was computed by displacing the respective γ threshold against the other. Figure 10 displays the correlation coefficients between the onset dates of two regions using varying γ values ranging from 0.1 to 1 with an increment of 0.05. In order to declare the significance off the correlation coefficients the data has to be checked towards normal distribution. Various test for normality, inter alia Kolmogorov–Smirnov (K–S) test, Jarque–Bera test and Lilliefors test, are commonly used. The shortcoming of the K–S test is that it compares the data with a standard normal distribution with zero mean and unity standard deviation. Since the onset dates are Julian Days and therefore impossibly negative, K–S test is (without applying a z-transformation before) rather improper for that purpose. Jarque–Bera test and Lilliefors test are evaluating the hypothesis that the data have a normal distribution with unspecified mean and variance. These tests are based on the skewness and kurtosis of the data sample. For a true normal distribution, the skewness should be near zero and the kurtosis should be near three. Jarque–Bera test is an asymptotic test which should therefore not be applied with small sample sizes (<50) (Judge *et al.*, 1988). Both distributions of Julian Days for each γ threshold combination and each model (Table IX) have been tested for normality at the 95% significance level ($\alpha = 0.05$). All the three tests have been applied and compared for each possible γ combination of the models M1–M4 (not shown here). Regarding to the K–S test, all mutual distributions of the models M1–M4 follow a normal distribution performing z-transformation before testing. Jarque–Bera test and Lilliefors test deliver slightly different test results. Table X is presenting the respective test result for the special γ combination used for each model. The prerequisite of normality is mostly fulfilled at the 95% significant level (except for the Jarque–Bera test for model 4). γ combinations with the highest absolute value of their correlation coefficients were used to generate the regression models. The onset of the rainy season of PC3 is best to estimate the onset of PC2, PC5, and PC1. The onset of PC4 can be estimated with PC1's onset. The respective regression equations, correlation coefficients, and best γ combinations are listed in Table XI. Quality parameters of the four different models (M1-M4) are summarized in Table IX. The time series of the dependent and independent variables (Julian onset dates) can be treated as stochastic independent, since they contain annual values. The quality of the regression model can be expressed via the R^2 which represents the explained variance fraction and the significance is verified using F–test. The Durbin–Watson test was applied in order to check the autocorrelation viz. independence of the residuals. Values between 1.5 and 2.5 can be regarded as independent whereas values next to zero and four are strong autocorrelated (Durbin and Watson, 1950). The non-parametric and distribution free K–S test was applied to check if the residuals are normal distributed, which is a necessary prerequisite for the correctness of the F–test. The K–S test essentially looks at the most extreme absolute deviation and determines the probability that this deviation could be explained by a normally distributed data set. The value listed as *asymptotical significance (2-tailed)* gives this probability (P) as a number between 0 and 1. A value of 0.05, for instance, means that only 5% of normally distributed data sets are expected to yields deviations as large as that reported for the extreme absolute deviation. In general, a *2-tailed asymptotical significance* value of 0.05 is considered good evidence that the data set is not normally distributed. A value greater than 0.05 implies that there is insufficient evidence to suggest that the data set is not normally distributed, but in turn, it does not provide evidence that the data set is normally distributed. Looking at the frequency distribution (histogram) of the standardized residuals, their mean value and standard deviation can circumstantiate the assumption of normality or non-normality. According to Table IX the models are highly significant but explain generally low fractions of the total variances (M1 = 27%,

DOI: 10.1002/joc

Table XI. Optimal γ threshold combination, linear regression equations, and correlation coefficients to estimate the onset dates of the target PC using the onset dates of the independent PC (the PCs are representing the areas illustrated in Figure 6).

Number	Target PC	Indep. PC	γ threshold target PC	γ threshold Indep. PC	Regression equation	r
M1	PC5	PC3	0.35	0.3	PC5 = 52.36 + 0.61 PC3	0.52
M2	PC2	PC3	0.3	0.85	PC2 = 44.33 + 0.57 PC3	0.46
M3	PC1	PC3	0.2	0.2	PC1 = 78.64 + 0.80 PC3	0.57
M4	PC4	PC1	1	0.8	PC4 = 119.7 + 0.40 PC1	0.47

Table XII. Confusion matrix for PC5 to assess the performance of the discrimination using the most suited variables and γ threshold with the main focus on the onset class (PC5 is representing the area illustrated in Figure 6).

$\gamma = 0.25$		Class membership after application of linear discriminant analysis (%)			
		Dry season	Transition	onset	Wet season
Predetermined class membership (%)	Dry season	**80.68**	13.56	1.44	4.32
	Transition	49.95	**30.98**	2.15	16.92
	onset	9.99	8.76	**68.34**	12.91
	Wet season	7.61	10.77	12.70	**68.92**

M2 = 21%, M3 = 33% and M4 = 22%). The standardized residuals of all four models are independent (D–W test) and normal distributed (K–S test). Figure 11 is additionally illustrating the empirical frequency distribution of the standardized residuals of model M4 with zero mean and standard deviation of 0.99. It can be seen that the empirical distribution follows approximately the normal distribution. Since these constraints are fulfilled, the regression models can be applied.

3. Summary and conclusions

Since the onset of the rainy season is of major interest for farming management strategies, it is important to find a reasonable way for its prediction. In this paper, basically two different methodologies operating on different predictive time frames were presented to predict the onset of the current rainy season on regional scale in the Volta basin: (1) LDA using simple rainfall indices and (2) Linear Regression Analysis (LRA) using onset dates of regions with earlier onset. Both methods are easily applicable, since they depend solely on daily precipitation amounts of 29 weather stations. In comparison to the paper of Sarria-Dodd and Jolliffe (2001), which is dealing with the discrimination into 'Real' and 'False' onsets of the rainy season in Burkina Faso and Mali using LDA, this approach allows one to distinguish between the three classes *dry season, onset of the rainy season* and *wet season* in Ghana and Burkina Faso. Their work is based solely on local scale, whereas this methodology aims at regional scale using a sophisticated fuzzy logic onset definition due to the high local rainfall variability on the observation site scale, but it is also applicable on local scale (not shown within this paper). However, the core of the definitions, used in Sarria-Dodd and Jolliffe (2001) as well as in this work is remaining the same. Both definitions are derived from the work of Stern *et al.* (1981) which is accounting for agricultural needs tropical West Africa.

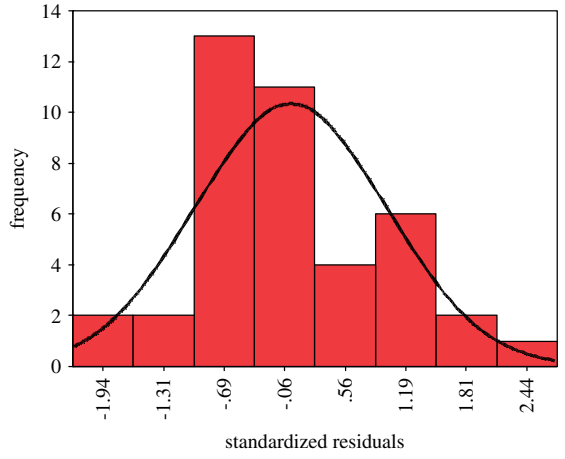

Figure 11. Histogram of the standardized residuals for model M4 (Figure 10). The black line represents the normal distribution. This figure is available in colour online at www.interscience.wiley.com/ijoc

To perform the regionalisation, PCA was applied to group the stations into five different regions showing zonal arrangement. Additionally, significant linear trends in the regional onset dates were found and analysed with respect to their potential reasons. Since both definition criteria (precipitation amount and number of wet days) are highly correlated for the *definition 2* (Section 2.2), it can be stated that a general decrease in both parameters took place. Regionally and monthly resolved analysis was presented in this paper. In a former study of

DOI: 10.1002/joc

Neumann *et al*. (2007), also significant precipitation trends on site scale were detected and many of them could be associated with climate indices, i.e. North Atlantic oscillation, Southern oscillation index, tropical North Atlantic oscillation and tropical South Atlantic oscillation. Hence, it may be concluded that these trends are likely to have large-scale reasons.

LDA is potentially useful to judge day by day whether the rainy season for the current year has already begun. It will not deliver the future onset dates, whereas LRA can potentially predict the onset dates a few weeks ahead under the assumption of the onset date of one region being already known. This also draws also a conceptual difference between the work of Sarria-Dodd and Jolliffe (2001) and this work. In our approach, an attempt has been made to integrate a priori information of regions with early rainy season's onset. For the prediction of the first onset region (PC3) in particular, the LDA has to be apply exclusively. Afterwards, LRA and LDA can be used co-instantaneously, depending on the predictive time frame desired. Furthermore, both methods can be cross-checked in the future. However, there are also limitations in applying these methods. For a daily decision of whether the time to plant is right or not, daily precipitation data of preferably all measurement sites are needed. In fact, they are measured and quoted day by day, but data collection and quality control strongly delay their availability for this application. Potential solutions of this problem consist in the usage of rain gauges with satellite transmission technology or meso-scale meteorological modelling within the Volta basin. In this case, validation studies of the model output using the measured data would be essential. None of the methods presented is suitable for extrapolation (years ahead). While the method of least squares regression often gives optimal estimates of the unknown parameters, it is very sensitive to the presence of outliers in the data used to fit the model. Few outliers can sometimes seriously skew the results of a least squares analysis. If too few outliers are included in the model, more robust regression methods coping with these problems can be used, e.g. based on a least median of squares estimator. To improve the performance of the models, it is important to include new data as soon as they are available and update the linear regression equations. Before, the prerequisites must be checked out carefully.

Acknowledgements

This work was funded by the German Ministry of Education and Research (BMBF) within the GLOWA-Volta project (http://www.glowa-volta.de). Financial support is gratefully acknowledged.

References

Ati OF, Stigter CJ, Oladipo EO. 2002. A comparison of methods to determine the onset of the growing season in Northern Nigeria. *International Journal of Climatology* **22**: 731–742.

Benoit P. 1977. The start of the growing season in northern Nigeria. *Agricultural Meteorology* **18**: 91–99.

Chamberlin P, Diop M. 2003. Application of daily rainfall principal component analysis to the assessment of the rainy season characteristics in Senegal. *Climate Research* **23**: 159–169.

Davidson NE, McBride JL, McAvaney BJ. 1983. The onset of the Australian monsoon during winter MONEX: Synoptic aspects. *Monthly Weather Review* **111**: 496–516.

Durbin J, Watson GS. 1950. Testing for serial correlation in least squares regression I. *Biometrika* **37**: 409–428.

Graef F, Haigis J. 2001. Spatial and temporal rainfall variability in the Sahel and effects on farmers' management strategies. *Journal of Arid Environments* **48/2**: 221–231.

Hendon HH, Liebmann B. 1990. A composite study of onset of the Australian summer monsoon. *Journal of the Atmospheric Sciences* **47**: 2909–2923.

Holland GJ. 1986. Interannual variability of the Australian summer monsoon at Darwin: 1952–82. *Monthly Weather Review* **114**: 594–604.

Ilesanmi OO. 1972. An empirical formulation of the onset, advance and retreat of rainfall in Nigeria. *Journal of Tropical Geography*, **34**: 17–24.

Judge GG, Hill RC, Griffiths WE, Lutkepohl H, Lee T-C. 1988. *Introduction to the Theory and Practice of Econometrics*. Wiley: New York.

Kaiser HF. 1958. The varimax criterion for analytic rotation in factor analysis. *Pyrometrical* **23**: 187–200.

Mason SJ. 1998. Seasonal forecasting of south African rainfall using a non-linear discriminant analysis model. *International Journal of Climatology* **18**: 147–164.

Middleton GV. 2000. *Data Analysis in the Earth Sciences Using MATLAB*. Prentice Hall: Saddle River, New Jersey.

Neumann R, Jung G, Laux P, Kunstmann H. 2007. Climate trends of temperature, precipitation and river discharge in the Volta Basin of West Africa. *International Journal of River Basin Management* **5**: 17–30.

Nicholls N. 1984. A system for predicting the onset of the Australian wet season. *Journal of Climatology* **4**: 425–435.

Omotosho JB. 1990. Onset of thunderstorms and precipitation over Northern Nigeria. *International Journal of Climatology* **10**: 840–860.

Omotosho JB. 1992. Long-range prediction of the onset and end of the rainy season in the West African Sahel. *International Journal of Climatology* **12**: 369–382.

Omotosho JB, Balogun AA, Ogunjobi K. 2000. Predicting monthly and seasonal rainfall, onset and cessation of the rainy season in West Africa using only surface data. *International Journal of Climatology* **20**: 865–880.

Richmann MB. 1993. Comments on: The effect of domain shape dependence on principal component analysis. *Journal of Climatology* **13**: 203–218.

Roncoli C, Ingram K, Kirshen P. 2002. Reading the Rains: Local knowledge and rainfall forecasting in Burkina Faso. *Society and Natural Resources* **15**: 409–427.

Sarria-Dodd DE, Jolliffe IT. 2001. Early detection of the start of the wet season in semiarid tropical climates of western Africa. *International Journal of Climatology* **21**: 1251–1262.

Stern RD, Dennett MD, Garbutt DJ. 1981. The start of the rains in West Africa. *Journal of Climatology* **1**: 59–68.

Stewart JI. 1991. Principles and performance of response farming. In *Climatic Risk in Crop Production. Models and Management for the Semi-Arid Tropics and Sub-Tropics*, Ford W, Muchow RC, Bellamy ZA (eds). CAB International: Wallingford.

Sultan B, Janicot S. 2000. Abrupt shift of the ITCZ over West Africa and intraseasonal variability. *Geophysical Research Letters* **27**: 3353–3356.

Sultan B, Janicot S. 2003. The west African Monsoon Dynamics, Part II: The "Preonset" and "Onset" of the Summer Monsoon. *Journal of Climate* **16**: 3407–3427.

Sultan B, Janicot S, Diedhiou A. 2003.. The west African monsoon dynamics, Part I: Documentation of Intraseasonal Variability. *Journal of Climate* **16**: 3389–3406.

Troup AJ. 1961. Variations in upper tropospheric flow associated with the onset of the Australian summer monsoon. *Indian Journal of Meteorology and Geophysics* **12**: 217–230.

Walter MW. 1967. Length of the rainy season in Nigeria. *Nigerian Geographical Journal* **10**: 123–128.

Ward MN, Folland CK. 1991. Prediction of seasonal rainfall in the north Nordeste of Brazil using eigenvectors of sea-surface temperatures. *International Journal of Climatology* **11**: 711–743.

DOI: 10.1002/joc