Homogeneity Analysis of Rainfall event by Using ROI Approach

Badreldin G. H. Hassan , Isameldin A. Atiem Dep.
of Civil Eng., Faculty of Engineering Science
University of Nyala
Nyala, Sudan
badr36hassan@yahoo.com, isam_eldin@hotmail.com

Feng Ping
Civil Engineering Department
Tianjin University, Tianjin, China
fengping@tju.edu.cn

Abstract—**The main purpose of this paper is to present and apply one of the statistical methodologies - the region of influence (ROI) approach - and form homogenous region/regions for rainfall frequency analysis in order to extract rainfall information guidelines for the Sudan basin. 15 gauging stations are selected, with recorded annual data ranging from 25 to 102 years in length. The aim here is to provide a regional curve with the capacity of taking into account the spatial pattern of variation of hydrologic phenomena across many gauging sites and which can be used for estimating rainfall quantiles at both gauged and ungauged sites within a specified region. This regional curve can also provide the possibility of the transfer of hydrologic behavior of a region to a site of interest in order to improve at-site estimates.**

The study results are analyses and compared. At first rainfall quantile for selected sites were estimated. In addition, the regression analysis between estimated quantiles (for 100 years) and the geographical features of the basin (drainage area (A), longitude (X_{dist}.) and latidute (Y_{dist}.)) was derived. In an aim to extend the methodologies to the case of ungauged site, a nonlinear regression model is derived and the results are investigated.

Keywords-At-site Frequency Analysis; Homogeneity; Regional frequency analysis; Region of influence (ROI); Regression Analysis

I. INTRODUCTION

Water is one of our most important natural resources, and there are many conflicting demands upon it. Skilful management of our water bodies is required if they are to be used for such diverse purpose as domestic and industrial supply, crop irrigation, transport, recreation , sport and commercial fisheries, power generation and waste disposal. So, it needs no probability to state that water constitutes one of the most essentials for life existence, and enough has been said about the role of water in the various biological processes sustaining life.

An essential step in the design of any containment structure and in water resource management is estimating the probabilities of occurrence for the events that the structure is designed to alleviate or that will play an important role in our decisions. The statistical method for making such probability predictions has been revolutionized by computers, and it seems that the current method based on regional analysis is much better than the conventional at-site procedure. At-site frequency analysis has a weak point on accuracy in case of short record length. Therefore, sufficient amount of data are needed for at-sitefrequency analysis. In Sudan, more reliable quantiles can be obtained by regional frequency analysis since there are not enough data at site.

The research tend to analysis the rainfall in Sudan, to develop design procedures for estimating rainfall magnitudes for given probabilities of nonexceedence at gauged and ungauged sites in the region with aim at improving adequate measures in the design of water resources structures, such as culverts, bridges, reservoirs, spillways, etc, and aiming towards a greening economy and sustainable development.

In this study, the region of influence (ROI) method was applied to annual maximum rainfall data. In addition, application of this method was examined. L-moments such as L-CV, L-skewness, and L-kurtosis were estimated for each site and divided regions. The aim of this study has been to adapt, analyse and evaluate an Index Flood regionalisation approach for Sudan. Preliminary investigations and results revealed the need for a flexible approach to region identification in order to correspond to the complex natural environment in Sudan. The potential approach has to allow for the highest degree of homogeneity possible for the respective sites of interest under the constraints of the hydrologic conditions and data availability.

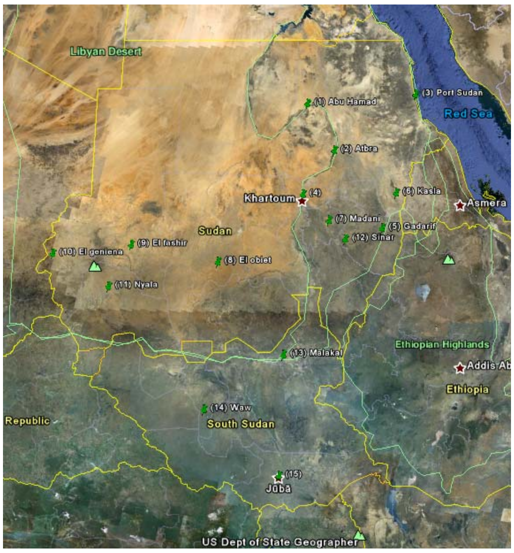

Figure 1. Satalite map of Sudan shows the study area

Published by Atlantis Press, Paris, France.

II. THE DATA

The data used have been collected from 15 raingauging stations, the records of which are published by the Sudan National Meteorology Department. Altitude and latitude have been assumed as initial statistics of hydrologic homogeneity, and station selection criteria were based on these characteristics. All sites used in this procedure are located between 31°50_ and 33°25_N latitudes, and 1530 and 2300 m altitudes (MSL). Records used for the analysis have ended the same year, and there are no gaps in the records (Table I and Fig. 1).

TABLE I. THE HYDROLOGICAL AND GEOGRAPHICAL CHARACTERISTIC OF THE RAINFALL SITES SELECTED IN SUDAN

Site name	Code	Data length	Mean ARF	L-cv	L-skew.	L-kurt.	lat.	Long.	Area.	Elev.
Abu Hamed	02ABH640	58	10.6034	0.7029	0.5437	0.3206	19.533	33.333	122.1	299.6
Atbra	02ATB680	58	59.3793	0.4231	0.1992	0.1345	17.667	33.967	30.4	316.3
Port Sudan	03PSD641	15	63.8933	0.42	0.0645	0.0767	19.583	37.217	218.8	525.3
Khartuom	04KHA721	102	150.8023	0.2835	0.1752	0.168	15.6	32.55	22.1	39.5
Gadarif	05GDF752	41	603.5854	0.1053	-0.0233	0.205	14.033	35.4	75.2	363.5
Kasala	05KSL730	28	240.7354	0.198	-0.023	0.1379	15.467	36.4	36.7	522.6
Madani	06WMD751	58	305.9828	0.1762	0.0403	0.096	14.383	33.483	27.5	416.7
Elobiet	08OBT771	90	359.3592	0.171	0.086	0.1399	13.183	30.217	185.3	492.8
Elfashir	09FSH760	84	262.332	0.2223	0.193	0.1969	13.617	25.333	296.4	405.7
Elgeniena	09GEN770	30	447.322	0.1901	0.0509	0.1614	13.483	22.45	79.4	526.2
Nyala	10NYL790	43	393.3309	0.1376	0.0674	0.1113	12.05	24.883	127.3	775.8
Sennar	12SNR762	60	458.7983	0.1173	-0.0342	0.1466	13.55	33.617	37.8	692
Malakal	13MLK840	58	742.5498	0.1068	0.0463	0.1436	9.55	31.65	77.7	392.4
Wau	14AWE852	28	1078.941	0.1003	0.042	0.0979	8.767	27.4	93.9	487.8
Juba	15JUB941	43	961.4919	0.1105	0.0177	0.1987	4.867	31.6	22.9	468.2

III. APPLICATION OF REGION OF INFLUENCE APPROACH

This section introduces the Region of Influence approach as a candidate method. Firstly, its concept is discussed, following the presentation of its procedural components. The final sub-section gives a tentative evaluation of the method, highlighting advantages and promising features as well as limitations and shortcomings.

A. The ROI Procedure

Further issues concerning the ROI approach are presented in this section in the context of a flowchart type representation, pointing out the sequential steps, including some implications for the practical application. Fig. 2 shows the flowchart with the ROI procedure's main components.

- Ungauged site of interest: starting point is the site of interest on which everything is focussed and for which a rainfall quantile estimate is desired. The derivation of its drainage basin is the first step.
- Derive basin descriptors: The subsequent step involves the derivation of basin descriptors to characterise the site of interest. These attributes are used as reference values in the following region identification process for the definition of similarity among basins.
- Region identification: This comprises the identification of those gauging stations that ultimately build the site of interest's region. This step itself can be broken down into several sub-components as illustrated in Fig. 3. Firstly, there is the attribute selection.

The outcome of this step is the choice of those basin attributes that are considered in the distance calculation. The successive step is the weight assignment, associating weights with the selected attributes.

Given the characterisation of the site of interest, the selected attributes and the assigned weights, the distance is calculated according to (1). The necessary information, the characterisations of all available gauging stations with respect to the selected attributes, is retrieved from an attribute database. The distance computation yields the similarity measures according to which the gauging stations can be sorted.

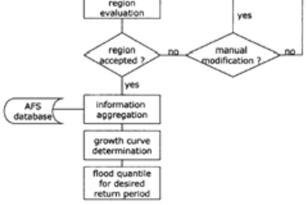

Figure 2. Flowchart of the ROI procedure's main components

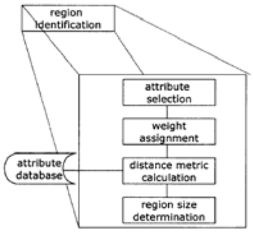

Figure 3. Components of the region identification step

This in turn represents the basis for the region size determination, which comprises the selection of the ROI sites.

- Region evaluation: The identified region has to undergo an evaluation with respect to its homogeneity and the fulfilment of the method's prerequisites. Fig. 4 shows a more detailed representation of this step with a number of tests and examinations that a region should pass in order to be accepted. This region diagnosis encompasses statistical homogeneity tests, an analysis concerning similarity in the seasonal behaviour of the ROI stations, and the examination of the goodness-of-fit of the Index Flood relationship and the investigation of the region's simple scaling behaviour.

Since some of these tests require a visual interpretation of the results, their graphical display is deemed to be beneficial for the evaluation. The synoptically evaluation of all test results leads to the subsequent step, the decision about acceptance or rejection of the region.

- Region accepted? This decision depends whether the pooling group can be rated acceptably homogeneous according to the outcome of the former region evaluation. If the region is rejected, a modification of the pooling group composition is necessary. On the one hand, this may be affected by a manual region revision, which could also be referred to as "heuristic search process" and includes some subjectiveness since the user's regional expert knowledge (knowledge of the regional hydrology conditions) is brought in. It comprises the removal of single stations from the region ensemble. Such stations may have been identified as not belonging to the ROI, because they showed discordant behaviour in one or more of the region evaluation tests. But also additional information that could not have been parameterised, like certain stations' peculiarities about which the user has knowledge may influence this revision.

On the other hand, without manual modification, a new region may be identified by running again through the whole region identification process with a different set of considered attributes and assigned weights. The modified region or the newly identified region has to undergo again the region evaluation.

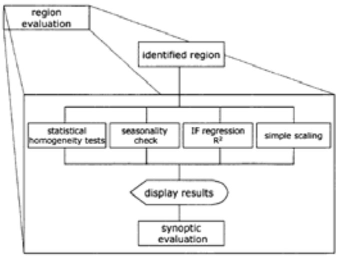

Figure 4. Components of the region evaluation

This loop has to be repeated until it passes the region acceptance step [11]. It is thus an iterative search process in a partly interactive manner. In the utmost case no acceptable region can be found so that no meaningful information transfer for the respective site of interest is possible.

- Information aggregation and growth curve determination: Once a region has been accepted, the rainfall series of the ROI stations have to be pooled into regional statistics. For this purpose, the stations' rainfall records are retrieved from a database containing the annual rainfall series of all available gauging stations. The data aggregation and, subsequently, the derivation of the growth curve are carried out, respectively.
- Rainfall quantile for desired return period: From the derived growth curve, the dimensionless rainfall quantile for the desired return period is computed. Obtaining the absolute value of the site of interest's rainfall quantile estimate still requires the multiplication with the Index Flood.

B. *Flexibility*

The ROI method is a flexible approach not only from its original idea in identifying unique regions for every site of interest, but also regarding its procedure in that for some of its components various options exist, which themselves contain a certain level of flexibility. From all this, the present ROI approach is understood as a methodological framework with an underlying heuristic idea. This framework, in essence the structure of the procedure illustrated in the flowchart of Fig. 2, is kept modular in that several options may serve for the single steps. It has also the advantage of easily integrating advancements of some components. This is unlike some applications where the approach has been carried out with strict rules and default options. For instance, the UK Flood Estimation Handbook, FEH (1999), suggests to consistently using three attributes, weighted equally, to calculate the distance measure.

C. *Degreen of Freedom and Subjectiveness*

This flexibility, which was desired to make regionalisation applicable for the heterogeneous environments found in Sudan, entails unfortunately subjectiveness. Main sources for subjectiveness are:

- The choice for one of the optional approaches to pass the individual steps of the procedure (although preference is mostly associated with one of the options).

- The need for expert judgement for attribute selection and weight assignment (importance of attributes is presumed to vary in space).
- The region evaluation and revision, and the decision whether to accept a region or not.

From the applied point of view, this is most likely the main reason for objection, since a considerable number of degrees of freedom is left to the user. If taken as a tool for the application oriented engineering hydrology community, then it is probably not what the engineering community would like to have in that too much is asked of the single user, overburdened with personal expertise and judgement requirements.

D. Information Updating

A further important feature of the ROI approach is its flexibility also with respect to the extension of the data pool with data becoming available in time. The data pool that the application exploits consists of the rainfall records and the basin descriptors of all gauging stations (the attribute- and the AFS databases in Fig. 3). A continuous updating with newly available information ensures that the estimation of rainfall quantiles at the site of interest is based on the best available data sets at that respective point in time. Such new information can originate from

- new gauging stations included in the database
- additional rainfall records of already considered gauging stations
- new basin descriptors

The possible emergence of new attributes is in favour of only providing a general framework with some optional procedures for the attribute selection rather than prescribing strictly specific attributes, since such a recommendation may become obsolete with the availability of new data. This is unlike the case of predefined regions, where the identification of the regions, the establishment of their growth curves etc. has been carried out with the data sets given at a certain point in time. Information that became available since the region determination is not considered.

E. The Concept of ROI

The method was introduced by Acreman [2]-[11]-[12]-[13]-[14] and has been adapted and developed by Bum [3]-[8]. An application of this method can be found, for instance, in Zrinji & Bum [15]. The same authors added a hierarchical feature to the ROI approach [16]-[15]. Tasker applied the concept so that regression analysis has been performed upon the identified ROIs in order to estimate rainfall quantiles. The UK Rood Estimation Handbook FEH (1999) adopted the ROI concept as well.

The fundamental idea of the Region of Influence approach is the determination of a unique region for every site of interest. It does not claim distinct boundaries between different regions (both in geographical and attribute space) and leaves the more traditional approach of breaking down a given study area into a limited number of a priori defined groups. Instead, a tailor-made region considers the information only from those gauging stations that are sufficiently similar to the respective site of interest. The core of the ROI method is, therefore, the definition of a similarity measure in order to quantify the closeness of the gauging stations with respect to the site of interest. This implies the choice and appropriate aggregation of basin attributes into a distance measure. Based on such a similarity index, a subset of all available gauging stations is selected which represents the most similar sites and creates the region of influence of the site of interest. This station selection requires a criterion for deciding how many stations to include. After the region composition has been determined, the rainfall records of the selected stations still need to be aggregated for, eventually, establishing the growth curve.

F. The Similarty Measure

As a similarity index a distance measure is suggested that calculates the closeness of all available stations to the site of interest in a multidimensional attribute space. A commonly used measure is the weighted Euclidean distance. In its most generic form, this distance measure can be written as

$$D_{ij} = \Big[\textstyle\sum_{m=1}^{M} w_m (X_m^i \qquad (1)$$

Dij being the weighted distance between stations j; and the site of interest i, M the number of attributes used to define station similarity, wm the weight applied to attribute m, the standardised value of attribute m for station j and p an exponent, respectively. This type of similarity measure is also very often used in cluster analysis. One may, therefore, refer to the ROI as a sort of dynamic clustering in that for every site of interest the corresponding cluster is determined anew.

The distance measure is based on M selected attributes. It is expected that similarity in certain attributes is equivalent to similarity in the rainfall generation processes with similar extreme flow responses. If several attributes are available, a selection has to be carried out with regard to which of them are most significant and should thus be included in the similarity measure. The distance measure is the core of the region identification, since stations are selected or discarded according to its value. It is therefore crucial which attributes to consider. This is why this section is dedicated to the issue of attribute selection. There, a number of possible approaches are discussed as to how to perform this task and, since intimately related, also deals with the next item, namely the weight assignment to the selected attributes.

The weight wm reflects the relative importance of attribute m. Since an attribute is an indicator of a related rainfall generation process, the weight wm ultimately represents the importance of this process. A potential refinement is that weights are a function of the return period T if it is presumed that the relative importance of the considered attributes changes with the extremity of the rainfall.

Since the attributes chosen for the distance measure will have different units and in most cases also different magnitudes, it is necessary to perform a standardisation of the attribute data prior to calculating the distance measure. As for standardisation, several methods are conceivable [6]-[1].

Such methods are, for example, normalisation based on a range or standard deviation and/or subtraction of the mean.

The two most common normalisation methods are briefly presented below.

Method A: standardisation of the attribute value Xj, by the standard deviation with subtraction of the mean. In formal writing,

$$X_j' = \frac{X_i - M_x}{S_x}$$

Method B: standardisation by the total attributes value range with subtraction of the minimum attribute value. In formal writing,

$$X_j' = \frac{X_j - X_{min}}{X_{max} - X_{min}}$$

Xj; being the actual value of the X attribute for site j; the standardised value, Mx the average value, Sx the standard deviation, Xmin the minimum value and Xmax the maximum value of attribute X.

In the case of method B, all standardised attribute values fall in the range [-1, 1], whereas in case A, the standardised values can also have values exceeding 1, or may fall below -1. The relative spread is more sensitive to the presence of single extremes in case B, because the reference range is determined only by the extremes, whereas in case A, this reference, Sx, depends on the whole sample. Therefore, standardisation method A is the preferred approach.

The exponent p in Equation 1, sometimes referred to as the "norm" of the distance measure, reflects the tolerance of deviations. The higher the value of p the more emphasis is put on high departures, with the extreme case of p → ∞, where only the largest distance of a single attribute becomes relevant for the distance measure. Thus, p is not associated with an attribute but rather to the whole system. A common value is p equal 2. Should p be an odd number, the absolute values of the differences in brackets in Equation 1 have to be taken.

Sometimes it is desired that the stations identified with such a distance measure be not overly scattered in space but are characterised by geographical proximity. In order to reach this goal, it is often necessary to include the geographical distance as an attribute, which can be incorporated in Equation 1 like any other attribute, too. To represent this incorporation in an explicit way, the distance formula can be rewritten in the form[8]-[3]

$$D_{ij}^{d} = \frac{D_{ij} + \frac{d_{ij}}{d_{max}} w}{1 + w} \qquad (2)$$

being the similarity index between station i and j; including spatial distance (combined similarity index), Dij the similarity index based on one or more basin characteristics, d_{ij} the geographic distance between catchments i and j, dmax the maximum geographic distance between any catchment pair, and w the weighting factor reflecting the relative importance of spatial proximity, respectively. Large values of w (as compared to a weight of one assigned to D_{ij}) will promote regions that are geographically contiguous, but not necessarily hydrologically similar in terms of the considered attributes.

Representing the central point on which the whole approach hinges, the distance measure deserves particular attention. Does it consider all relevant factors with their right proportion so that it truly reflects a measure of similarity in rainfall response? This implies that those factors can be identified and the associated indicators, or rather basin attributes are derivable from available information. In practical terms, one can only expect an approximation of this ideal situation. Evaluating the significance of the similarity measure is subject to interpretation. To what extent the dominant rainfall generation processes are captured by the index? As an example of what remains typically unaccounted for are local hydraulic effects. If no suitable parameter is at hand for such strongly site-specific features as a similarity attribute, one has to assess the extent to which this reduces the significance of the similarity measure. If recognised as an essential factor, this may promote research efforts for obtaining a related parameter.

G. The ROI Size

In cluster analysis, one is confronted with the problem of finding a criterion for identifying the limit between within-group similarity and between-group dissimilarity. In other words, a "stopping rule" for incorporating new stations to the cluster is required. As for the ROI approach, a similar problem occurs when it is about the size (number of sites) of the region. Up to which distance further stations should be included in the region? Once the distance has been calculated, the gauging stations can be ranked accordingly and the most similar stations will then be selected to build the region (see Table II). This is sketched schematically in Fig. 5, illustrating an assumed distribution of the distance measure. The more uniform this distribution, the less clear the cutting point for defining very similar and/or dissimilar basins.

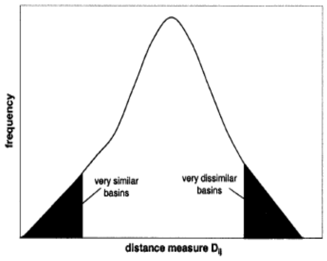

Figure 5. Sketch of an assumed distribution of the distance measure

There are several approaches as to how to carry out the selection of those stations that eventually constitute the region on the basis of the above similarity measure.

Subjective determination, considering the similarity measure:

Although the absolute distance values themselves do not have any physical meaning, in a relative context they can be used to discriminate between more and less similar basins. Thus, the procedure would be to stop incorporating sites when the corresponding distance values are significantly higher than those from the most similar sites. In essence, this is what is outlined in Fig. 5.

Based on a distance measure threshold:

Burn [3]-[7]-[9] discussed some options based on such a threshold. The set Ii, of stations selected for station i is based on the distance measure Dij in that all stations j that have a distance Dij smaller than a certain threshold θ, will be selected.

$$I_i = \{j = D_{ij} \leq \theta_i\} \tag{3}$$

A further point is then still the determination of a weighting function that reflects the relative closeness of each station of the pooling group with respect to the site of interest. This weighting function is used for the data aggregation, that is, the higher the distance, the less similar the station and thus the less weight assigned to the respective rainfall records for the regionally averaged moments. One option is to choose a sufficiently large threshold value θ_i so that all stations are included in a ROI in combination with a weighting function reflecting the respective similarity. Another option would result from choosing a very restrictive threshold so that ROI becomes very small. The resulting stations are then expected to be very similar in rainfall response to the site of interest and the weights of these stations are significantly different from zero, i.e., all one. A special case of this option would be to set θ_i so that the region's size becomes zero, that is, Ii= {}. It would mean that there is no other acceptably similar basin that could be used for a meaningful information transfer.

Between the two above extreme options, some "in-between" options may be conceived. For instance, up to a lower threshold sites are weighted equally and then, further to an upper distance threshold, weights decrease according to the corresponding similarity measure.

Coupling region size and return period:

The UK Flood Estimation Handbook (FEH, 1999) suggests a rule of thumb, referred to as 5T rule. It says that the pooled stations should collectively supply five times as many years of record as the target return period, T. Thus, the pooling group has a size to provide at least 5T station-years of rainfall data. The philosophy of this approach is to reduce the extrapolation span and, thus, uncertainty of the required rainfall quantile while coupling sample size and return period. This is why resultant pooling groups are referred to as site- and return period-oriented. Such an approach, however, ignores the homogeneity of a region, since the homogeneity level of a pooling group is expected to decrease while increasing its ROI size. Moreover, fixing the number of selected stations according to the 5T rule considers the most similar stations irrespective of their absolute distance values. As for the situation in Sudan, a target ROI size valid for the whole country is not appropriate already alone from the fact that the gauging stations are anything but evenly distributed across Sudan.

On the basis of a homogeneity test:

The rationale of this variant can be formulated as follows. Starting from the station list, which is sorted according to the distance measure, stations are incrementally added to the pooling group. After each step, the region is assessed by means of a statistical homogeneity test. The inclusion of further stations stops as soon as the applied test identifies the station ensemble as heterogeneous (an example of such an approach is reported in Zrinji & Bum) [2]-[9]-[16].

Table III shows the applications of this variant rely on Wiltshire's R-test. Although this approach implicitly comprises the statistical homogeneity evaluation of a region, it also entails all limitations of the respective homogeneity test.

TABLE II. THE DISTANCE MEASURED FOR THE SITES IN THE SUDAN

	Site name	dist
15	Juba	137.3621
13	Malakal	336.8076
5	Gadarif	475.8141
12	Sennar	622.7217
10	Geneina	631.8234
11	Nyala	686.4355
8	El Obeid	725.3804
7	Medani	775.8509
6	Kassala	840.2244
9	El Fashir	841.3596
4	Khartoum	930.9524
2	Atbara	1021.595
3	P.Sudan	1022.81
1	Abuhamed	1068.783

TABLE III. THE HOMOGENEITY MEASURED FOR SUDAN STATIONS IN TWO DIFFERENT RETURN PERIODS

Stations	Homogeneity test for 50 years return period			Homogeneity test for 100 years return period		
	H1	H2	H3	H1	H2	H3
Juba	6.095858	-0.80327	-1.40503	7.939557	1.169464	-0.25691
Malakal	5.02124	-0.5171	-1.63533	7.890627	1.127443	-0.40254
Gadarif	6.396919	-0.24879	-1.55198	9.391465	1.379381	-0.2674
Sennar	5.532459	-0.54589	-1.40034	12.56084	1.493463	-0.62482
Geneina	6.293047	-0.08877	-1.45596	8.788771	1.311509	-0.23053
Nyala	2.660354	0.377896	-0.65761	14.21808	1.399999	-0.51418
El Obeid	7.987985	1.117678	-0.15654	12.7013	1.467018	-0.45849
Medani	4.710151	-0.52101	-1.50249	21.57505	3.464991	0.353567
Kassala	4.610747	-0.66535	-1.44541	25.17351	2.906935	-0.81223
El Fashir	10.2593	2.132511	0.734039	23.11148	4.04877	-0.00183
Khartoum	16.31915	4.476205	1.597204	37.15105	9.158606	3.039539
Atbara	19.13413	8.710709	4.105177	33.6009	8.791841	3.144629
P.Sudan	25.90281	8.601055	3.470004	33.30123	9.842978	4.469462
Abuhamed	16.16326	6.672621	2.313062	32.6757	8.013197	3.108208
Wau	5.283505	-0.37234	-1.50577	9.632294	1.440788	-0.18715

An aspect closely related to this section's issue is the question as to how different one may expect the ROI sizes to be across Sudan. How plausible is the assumption of a consistent number of pooling group members across regions? As previously mentioned, the similarity measure has a relative meaning. In absolute terms, however, one can expect that in some parts of Sudan, or rather for some sites of interest, one will find more other resembling basins and thus larger regions than for other sites.

This already descends from the fact that the gauging station density is not uniform and that some areas feature higher variability than other, more homogeneous parts of Sudan. Moreover, there may be sites of interest with particularities and unique features that make it likely to find hardly any closely resembling gauged basin. In all, one cannot presume a consistent ROI size, simply because the predisposition for regionalisation is not the same everywhere. This also explains the impossibility of generally qualifying the ROI approach as suitable or not for Sudan. Rather, it depends on the respective site of interest and its broader surrounding. As empirical evidence thereof, a considerable number of potential sites of interest have been examined with respect to their maximum ROI size that can still be qualified as homogeneous according to an applied statistical homogeneity test.

The resulting differences in the size of homogenous regions among the analysed sites corroborate empirically the above considerations.

The issue of the adequate region size is also addressed here and applied for two different return periods, where the ROI method is systematically applied with varying region sizes. The effect on the prediction performance has been investigated in order to obtain possible indications with respect to the region size magnitude Table IV for 50 years return period shows the unique region for every site of interest (10 selected sites).

TABLES IV. THE UNIQUE REGIONS CHARACTERISTICS FOR EACH SITE (FOR T= 50 YEARS)

	l1	**l2**	**lcv**	**lca**	**lkur**	**ni**	**Si**	**wi**	**di**
Unique Regions									
elobeid	359.3592	61.4524	0.171	0.086	0.1399	90	1	90	0
elfashr	262.332	58.3283	0.2223	0.193	0.1969	84	0.6897	57.931	0.6925
geneina	447.322	85.0489	0.1901	0.0501	0.1614	30	0.4	12	0.7549
wau	1078.941	108.1868	0.1003	0.042	0.0979	28	0.2966	8.3034	1.0594
malakal	742.5498	79.2736	0.1068	0.0463	0.1436	58	0.2	11.6	1.1047
atbara	59.3793	25.1246	0.4231	0.1993	0.1345	58	1	58	0
medani	305.9828	53.9265	0.1762	0.0403	0.096	58	0.8092	46.9342	1.4601
kassala	240.7354	47.658	0.198	-0.023	0.1379	28	0.6184	17.3158	1.6679
khartou	150.8023	42.7477	0.2835	0.1752	0.168	102	0.5263	53.6842	1.8098
abuhamd	10.6035	7.454	0.703	0.5437	0.3206	58	0.1908	11.0658	1.822
nyala	393.3309	54.1253	0.1376	0.0674	0.1112	43	1	43	0
sennar	458.7983	53.8186	0.1173	-0.0342	0.1465	60	0.8286	49.7211	1.1326
geneina	447.322	85.0489	0.1901	0.0509	0.1613	30	0.5896	17.6892	1.518
elobeid	359.3592	61.4524	0.171	0.086	0.1398	90	0.4701	42.3107	1.6889
kassala	240.7354	47.658	0.1979	-0.0229	0.1379	28	0.1115	3.1235	1.8538
medani	305.9828	53.9265	0.1762	0.04032	0.0959	58	1	58	0
kassala	240.7354	47.658	0.1979	-0.0229	0.1379	28	0.7751	21.7054	0.6935
juba	961.4919	106.2412	0.1104	0.0177	0.1987	43	0.6666	28.6666	0.9908
gadarif	603.5854	63.5795	0.1053	-0.0232	0.205	41	0.5	20.5	1.052
geneina	447.322	85.0489	0.1901	0.0509	0.1613	30	0.341	10.2325	1.1355
malakal	742.5498	79.2736	0.1067	0.0462	0.1436	58	0.2248	13.0387	1.1363
malakal	742.5498	79.2736	0.1067	0.0462	0.1436	58	1	58	0
gadarif	603.5854	63.5795	0.1053	-0.0232	0.205	41	0.8	32.8	0.2404
wau	1078.941	108.1868	0.1002	0.0419	0.0979	28	0.6586	18.4413	0.6555
geneina	447.322	85.0489	0.1901	0.0509	0.1613	30	0.562	16.862	0.8858
elobeid	359.3592	61.4524	0.171	0.086	0.1398	90	0.4586	41.2758	1.1046
juba	961.4919	106.2412	0.1104	0.0177	0.1987	43	0.1482	6.3758	1.1314
kassala	240.7354	47.658	0.1979	-0.0229	0.1379	28	1	28	0
medani	305.9828	53.9265	0.1762	0.0403	0.0959	58	0.8923	51.7538	0.6935
geneina	447.322	85.0489	0.1901	0.0509	0.1613	30	0.6692	20.0769	0.8178
sennar	458.7983	53.8186	0.1173	-0.0342	0.1465	60	0.5538	33.2307	1.1228
juba	961.4919	106.2412	0.1104	0.0177	0.1987	43	0.323	13.8923	1.235
gadarif	603.5854	63.5795	0.1053	-0.0232	0.205	41	0.1576	6.4653	1.3364
sennar	458.7983	53.8186	0.1173	-0.0342	0.1465	60	1	60	0
kassala	240.7354	47.658	0.1979	-0.0229	0.1379	28	0.7931	22.2069	1.1228
nyala	393.3309	54.1253	0.1376	0.0674	0.1112	43	0.6965	29.9517	1.1326
geneina	447.322	85.0489	0.1901	0.0509	0.1613	30	0.5482	16.4482	1.1521
juba	961.4919	106.2412	0.1104	0.0177	0.1987	43	0.4448	19.1275	1.502
wau	1078.941	108.1868	0.1002	0.0419	0.0979	28	0.2965	8.3034	1.5783

medani	305.9828	53.9265	0.1762	0.0403	0.0959	58	0.2	11.6	1.6674
elfashr	262.332	58.3282	0.2223	0.1929	0.1969	84	1	84	0
elobeid	359.3592	61.4524	0.171	0.086	0.1398	90	0.6769	60.923	0.6924
gadarif	603.5854	63.5795	0.1053	-0.0232	0.205	41	0.3307	13.5615	1.3527
p.sudan	63.8933	26.8342	0.4199	0.0644	0.0767	15	0.173	2.5961	1.3693
geneina	447.322	85.0489	0.1901	0.0509	0.1613	30	0.1153	3.4615	1.3763
gadarif	603.5854	63.5795	0.1053	-0.0232	0.205	41	1	41	0
malakal	742.5498	79.2736	0.1067	0.0462	0.1436	58	0.8655	50.2032	0.2404
wau	1078.941	108.1868	0.1002	0.0419	0.0979	28	0.6754	18.9114	0.8876
geneina	447.322	85.0489	0.1901	0.0509	0.1613	30	0.5836	17.5082	0.9849
medani	305.9828	53.9265	0.1762	0.0403	0.0959	58	0.4852	28.1442	1.052
elobeid	359.3592	61.4524	0.171	0.086	0.1398	90	0.295	26.5573	1.1479
geneina	447.322	85.0489	0.1901	0.0509	0.1613	30	1	30	0
elobeid	359.3592	61.4524	0.171	0.086	0.1398	90	0.8909	80.1818	0.7549
wau	1078.941	108.1868	0.1002	0.0419	0.0979	28	0.5636	15.7818	0.7636
kassala	240.7354	47.658	0.1979	-0.0229	0.1379	28	0.4618	12.9309	0.8178
malakal	742.5498	79.2736	0.1067	0.0462	0.1436	58	0.36	20.88	0.8858
gadarif	603.5854	63.5795	0.1053	-0.0232	0.205	41	0.149	6.1127	0.9849

TABLE V. THE SITES AND REGIONAL QUANTILES

Quant. Site	0.01	0.5	0.9	0.99	0.999	0.9999
Atbra	60.12669	157.5629	229.1984	293.0033	336.1474	365.9251
Gadarif	221.4859	580.4074	844.2877	1079.323	1238.251	1347.941
Kasala	155.5332	407.5772	592.8809	757.9286	869.532	946.5596
Madani	188.9195	495.0665	720.1469	920.6233	1056.183	1149.745
Elobiet	145.9777	382.5368	556.4559	711.3636	816.1104	888.4057
Elfashir	121.8521	319.3154	464.4911	593.7973	681.2327	741.5798
Elgeniena	183.1483	479.9429	698.1473	892.4995	1023.918	1114.622
Nyala	151.9929	398.2999	579.3856	740.6765	849.7396	925.0139
Sinar	185.7107	486.6578	707.9151	904.9864	1038.244	1130.217
Malakal	238.5225	625.052	909.2298	1162.344	1333.496	1451.624
Reg. quant						
	0.372473	0.97607	1.419838	1.815096	2.082365	2.266831

IV. CHOICE OF THE REGIONAL FREQUENCY DISTRIBUTION

The establishment of the growth curve requires the selection of a theoretical distribution function. Tools for the decision and discrimination between distributions are goodness-of-fit tests (e.g. chi-square, Kolmogorov-Smirnov) or graphical plots, allowing a visual inspection of which distribution best fits the empirical distribution.

Also moment-ratio diagrams provide a means of comparing certain theoretical distributions with the sample characteristics (see Fig. 6).

From the diagram (Fig. 6) the Generalised Extreme Value (GEV) distribution is chosen as regional distribution throughout the method, the parameters of which are based on L-moments.

V. THE DATA AGGREGATION AND DETERMINATION OF THE REGIONAL GROWTH CURVE

For the purpose of the data pooling of the considered ROI stations, the approach of regionally averaged moments is pursued. The ROI approach provides the similarity measure, D_{ij}, for the selected stations of a pooling group.

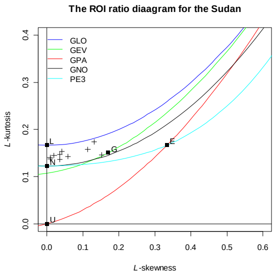

Figure 6. Moment ratio diagrams

Hence, it is self-evident to employ this information for the derivation of weights. Such weights support the aim of allowing a varying influence of the individual ROI stations with respect to the growth curve determination while calculating weighted regionally averaged moments (see Tables IV).

Since it is desired that, apart from the similarity measure, also information on the reliability of the stations' rainfall series influences the regional moments, overall weights [4]-[5] are suggested for the data aggregation.

At the end of the data aggregation procedure are the regional moments. They provide the necessary information to estimate the parameters of the growth curve, for which the GEV distribution is, applied [10]-[13]-[15]. Table 6 shows the sites and regional quantiles calculated by the methods. Table VI and figure 6 show the Regional growth curve and bounds

TABLE VI. THE REGIONAL GROWTH CURVE AND BOUNDS

	f	qhat	RMSE	bound.0.05	bound.0.95
1	0.01	0.372473	0.019721	0.337199	0.418347
2	0.5	0.97607	0.00505	0.966871	0.987138
3	0.9	1.419838	0.011721	1.398044	1.442785
4	0.99	1.815096	0.040477	1.732545	1.893162
5	0.999	2.082365	0.077012	1.937572	2.229784
6	0.9999	2.266831	0.11364	2.044969	2.481002

VI. DISCUSSIONS OF ROI RESULTS

The adequate region size is addressed and applied for two different return periods, where the ROI method is systematically applied with varying region sizes. The effect on the prediction performance has been investigated in order to obtain possible indications with respect to the region size magnitude Table IV for 50 years return period shows the unique region for every site of interest (10 selected sites).

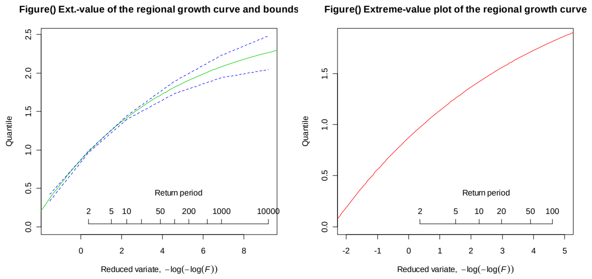

Figure 6. Shows extreme value of the Regional growth curve and bounds

Diagnostics of model results, it compares the estimated quantile for 100 years and corresponding drainage area (A km2), longitude (Xdis) and latitude (Ydis) at different selected sites (10 sites) is carried out, which respectively results in the following derived equations:

$$M_Q = 1951.81A^{-0.96}X_{dist}^{12.09}Y_{dist}^{-108.97} \qquad (4)$$

$$M_Q = 1408.51A^{-0.46}X_{dist}^{7.33}Y_{dist}^{-87.70} \qquad (5)$$

With coefficient of determination R2 of 0.7155 and 0.7969, the root mean squared error (RMSE) of 125.7424 and 82.3325, the mean absolute error (MAE) of 101.9746 and 60.8757, the percentual root mean squared error (RMSEP) of 24.8% and 25.8%, and the percentual mean absolute error (MAEP) of 16% and 19.6% for the defined sites.

VII. THE FINAL EVALUATION

Therefore, the ROI approach, as it is conceived in this paper, is not viewed as a rigidly specified procedure, but is rather understood as a methodological framework supporting the user in identifying and qualifying the region for the site of interest.

Integration of regional expert knowledge is seen as an essential component of the method. This is accomplished in an interactive and iterative manner in conjunction with the developed software program.

From the conceptual point of view, the ROI approach may be qualified to possess the necessary flexibility to tackle the unfavorable conditions - due to the high heterogeneity - for information transfer in Sudan. There are, however, a number of relevant issues that do not generally allow advocating or rejecting the approach as far as its suitability for application in Sudan is concerned.

One point is that the method parse may be very flexible, but requires a corresponding data availability covering and reflecting the heterogeneous conditions in space. The suitability, or rather, the success of the ROI method thus depends to a significant extent on the given information density and quality. Information refers to both rainfall records, as actual information to be transferred, and basin attributes, used to identify the stations among which the information is transferred. Concerning rainfall records, information density is related to the spatial density of gauging stations.

Information quality is related to the gauging stations' record lengths as well as the reliability of the rainfall measurements. With regard to the basin attributes, information density and quality refer to the meaningfulness of the available attributes. To what extent do the derived and selected basin characteristics capture all dominant processes? As both aspects, information density of rainfall records and basin attributes, vary in space, the preconditions for the successful ROI application differ spatially.

While for some sites a sufficient number of acceptably similar gauging stations may be identified, no such adequate pooling groups may result for other sites. The evaluation of the suitability of the identified region for information transfer is suggested to be based on a synoptic view of a number of diagnostic tools that have been presented.

From the results presented above, one may conclude that the ROI is an efficient regionalization approach and can be easily extended to the case of ungauged sites with some revisions to the attributes selected in the distance measure. As the concept itself is site oriented, a final evaluation may thus come to the conclusion that also the assessment of the approach's suitability is site-specific. It depends on the respective information available around the site of interest as well as on the adequacy of the basin attributes in reflecting the important processes. Hence, the recommendation is to apply the ROI method with the given constraint of limited available input information and then to evaluate from case to case if the best identifiable region is acceptably homogenous and contains sufficient and reliable rainfall data for performing the desired information transfer.

REFERENCES

[1] Acreman, M. C. & Wiltshire, S. E. (1987). Identification of regions for regional flood frequency analysis. EOS, 68(44), 1262, an abstract.

[2] Burn, D. H. (1989). "Cluster Analysis as Applied to Regional Flood Frequency", Journal of Water Resources Planning and Management, Vol. 115, No. 5, pp. 567-582.

[3] Burn, D. H. (1990). Evaluation of regional flood frequency analysis with a region of influence approach. Water Resources Research, 26(10), 2257-2265.

[4] Coles, S. G., and J. A. Tawn. (1996). A Bayesian analysis of extreme rainfall data. Applied Statistics-Journal of the Royal Statistical Society Series C 45:463-478.

[5] Cunnane, C. (1988). Methods and merits of regional flood frequency analysis. Journal of Hydrology, 100, 269-290.

[6] Hartigan, J.A., 1975: Clustering algorithms. John Wiley and Sons, New York.

[7] Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. Biometrika 57:97-109.

[8] Hosking, J, R; J, R, Wallis & E, F, Wood (1985). Estimation of the generalized extreme -value distribution by the method of probability weighted moments. Technometrics, 27(3), 251-261.

[9] Hosking, J.R (1986). The theory of probability weighted moments. IBM Res. Rep. RC12210, IBM, Yorktown Heights, N.Y.

[10] Hosking, J.R. & Wallis, J.R. (1988). The effect of intersite dependence on regional flood frequency analysis. Water Resources Research, 24(4), 588-600.

[11] Badreldin G. H. Hassan, Isameldin A. Atiem, Li Jianzhu, Feng Ping (2012) “At Site and Regional Frequency Analysis for Sudan Annual Rainfall by Using the L-Moments and Nonlinear Regression Techniques”, International Journal of Engineering Research and Develpment, e-ISSN: 2278-067X, p-ISSN: 2278-800X, www.ijerd.com, Vol. 3, Issue 6, September 2012, PP. 13-19.

[12] Jin, M. & Stedinger, J. R. (1989). Flood frequency analysis with regional and historical information. Water Resources Research, 25(5), 925-936.

[13] Kuczera, G. (1982). “Combining Site-Specific and Regional Information: An Emprical Bayes Approach",Water Resources Research, Vol. 18, No. 2, pp. 306-314.

[14] Lettenmaier, D. P. & Potter, K.W. (1985). Testing flood frequency estimation methods using a regional flood generation model. Water Resources Research, 21(12), 1903-1914.

[15] Lettenmaier, D. P., J.R. Wallis & E. F. Wood (1987). Effect of regional heterogeneity on flood frequency estimation. Water Resources Research, 23(2), 313-323.

[16] Wallis, J. R. and Wood, E. F. (1985). "Relative Accuracy of Log-PearsonII proceedures", Journal of Hydraulic Engineering, Vol. 111, No. 7, pp. 1043-1056.

[17] Zrinji, Z. & Burn, D. H. (1996). Regional flood frequency with hierarchical region of influence. Journal of Water Resources Planning and Management, 122(4), 245-252.