Outlier Detection using Nonparametric Depth-Based Techniques in Hydrology

Insia Hussain[1]
College of Computer Science and Information System
Institute of Business Management, Karachi, Pakistan

***Abstract*—Several issues arise when extending the methods of outlier detection from a single dimension to a higher dimension. These issues include limited methods for visualization, marginal methods inadequacy, lacking a natural order and limitation in parametric modeling. The intension to overcome and address such limitations the nonparametric outlier identifier, based on depth functions, is introduced. These identifiers comprise of four threshold type outlyingness functions for outlier detection that are Mahalanobis distance, Tukey depth, spatial Mahalanobis depth, and projection depth. The object of the present research is the application of the proposed nonparametric technique in hydrology. The study is intended to be executed in two different frameworks that are multivariate hydrological data analysis and functional hydrological data analysis. The event of a flood is graphically represented by hydrograph whose components are used for computing flood characteristics that are peak(p) and volume(v). These characteristics are frequently employed for the various types of analysis in the multivariate study. Whereas, hydrograph is exhaustively employed in the analysis of functional data so that all the important information regarding flood event are not missed while analysis. The proposed technique in a multivariate framework is applied to the bivariate flood characteristics $(\boldsymbol{p}, \boldsymbol{v})$ while in functional framework proposed approach is applied to the initial two scores of principal components denoted as $(\boldsymbol{z_1}, \boldsymbol{z_2})$, since initial two principal components capture major variation of data employed for analysis.**

Keywords*—Outlyingness functions; nonparametric techniques; flood characteristics; principal component scores; multivariate analysis; functional analysis

I. Introduction

The “outlier” observations in any data set is crucial to be detected and identified for nonparametric or parametric inferences. “Outliers” are the observations that are inconsistent or far from the majority of data points or within the chunk of data points with unusual behaviour. The presence of unusual observations in the data set acts as an outlier that can impact adversely the outcomes of estimation, inference, and testing procedures. Therefore, outliers are required to be identified and treated so that inferences are not violated due to unusual observations [1,2].

Outliers identified marginally suffer inadequacy of checking, in each coordinate, an outlier can find to be nonoutlying. Approaches that are algorithmic and take into account underlying geometry are required. A suitable function of outlyingness may be formulated with a threshold specified. A suitable choice can be Mahalanobis distance which is a highly tractable function of outlyingness but constrained for having elliptical contours of symmetric outlyingness, even though whether the model under consideration is symmetric elliptically.

The author in [3] introduced a nonparametric technique which is based on functions of depth and orders the multidimensional data in center-outward. Higher depth represents higher centrality whereas lower depth greater outlyingness. One can associate with any depth function an equivalent function of outlyingness. For a suitable selection of depth function, actual geometrical structure and data shape are formed by equal outlyingness contours. In general, four different affine invariant functions of outlyingness were derived which are based on Mahalanobis distance outlyingness (MO), projection depth outlyingness (PO), halfspace or Tukey depth outlyingness (TO), and Spatial Mahalanobis outlyingness (SO). Related to these outlyingness functions the corresponding points are “outliers” having values of outlyingness exceed the constrained threshold of a particular function.

The nonparametric approaches introduced by [3] have been practiced by [4] and [5] in hydrology while [4] executed multivariate hydrological data analysis using two frequently employed flood characteristics; peak(p) & volume(v), for the identification of unusual observations i.e. outliers.

The author in [5] came up with groundbreaking research and extended the work of [4] by conducting functional hydrological data analysis. The nonparametric outlier identification technique was practiced in hydrology by [5] in such a way that the initial two scores of principal components were employed for the detection of outliers in a functional context. In multivariate analysis, employed flood characteristics are dependent and mutually correlated whereas scores of principal components employed in functional analysis are uncorrelated.

The execution of research in the functional framework follows the claim made by [5] that the characteristic of flood use in conducting the multivariate hydrological study are computed by subjective approach and do not encounter the complete series of employed data set, therefore, inferences of multivariate study suffer lack of authenticity. Hence it is crucial to conduct research in a functional framework so that authentic estimation regarding the associated risk of flood is obtained by incorporating complete phenomena produced through employed data series.

The objective carried by present research is the implementation of nonparametric techniques based on depth functions in both the context of a study that is a multivariate and functional framework using hydrological data of Kotri Barrage on Indus River in Pakistan.

II. Literature Review

The methods going to be presented are based mainly on the statistical notion of depth functions. These functions provide convenient ranking tools for ordering data variables. Depth functions were initiatively practiced in hydrology by [6]. Several techniques of univariate analysis were extended to execute multivariate analysis developed through analogy. The variables that are dependent mutually affect the performance badly when analysing data component-wise, whereas moment-based techniques required the moment's existence.

Review in detail regarding techniques use for conducting classical multivariate analysis, it is referred to follow [7,8]. Techniques that are developed on the basis of depth, avoid the earlier drawbacks science depth functions are ordered using multivariate inward and outward ranking [9]. Indeed, techniques based on depth aren't component-wise, also, they are affine invariant and moment-free. Numerous techniques of outlier detection are enabled by ranking based on depth. The number of depth function formulas have been derived for executing the multivariate study. Depth region location inference considered by [3] is evaluated on sample space. Description of connection and general treatment related to multivariate quantile and centre ranked functions can be studied through [10,11]. For other inferential applications of depth see [12,13]. Numerous studies conducted in hydrology using various nonparametric approaches. The functions based on depth have been recently employed for the detection of outliers by [14,15]. According to [16], nonparametric models are suitable for capturing subtle aspects related to the frequency estimation of a flood. Flood inundation and flood damage were analysed using hydrologically distributed models through nonparametric techniques [17]. Similar other studies recently conducted in hydrology for outlier detection and risk estimation using nonparametric approaches are [18,19]. Characteristics of drought evaluation were assessed in a multivariate context implementing a nonparametric approach by [20-22]. Further research of [23] discussed data cleaning of water consumption and estimation of uncertainty regarding hydrologic modeling. Depth notion in regression was practiced and the performance of runoff model was evaluated, see work of [24-26]. Author in [27] used parametric and nonparametric multivariate approaches for designing rainfall framework whereas [28] applied rank-based nonparametric techniques to study trends of rainfall.

Multidimensional data is reduced by of analysis of functional principal component (AFPC) techniques to attain an easy approach for analyzing hydrological data. Notable work includes profile classification of streamflow, minimum indicators selection and functional data analysis application on streamflow are the studies executed on the basis of AFPC. Simulation of drought interval and drought changes were analysed by [29,30]. [31-33] studied rainfall variability modeling, pattern identification, and outlier detection. Other relevant studies include work of [34-38], are also preferred for acquiring information about the useful application of AFPC in hydrology.

This paper is organized in such a way that the discussion regarding proposed methodologies is presented in Section 3. Section 4 provide description related to hydrological data employed for executing present research. Section 5 provides an application of the discussed methodology on employed hydrological data and obtained results are provided in Section 6 whereas Section 7 contain the conclusion drawn from the research.

III. Methodology

This section contains methods for computing bivariate series of flood characteristic (p, v) and also bivariate series of principal component scores (z_1, z_2). Both the computed series (p, v) and (z_1, z_2) are required for obtaining outliers in multivariate and functional context, respectively, using proposed threshold type nonparametric techniques which will also be discussed later in this section.

A. Flood Characteristics

The flood peak (p) and volume (v) are the fundamental and most studied flood characteristics [39-41] and their computation based on the work of [41].

The bivariate series (p, v) are generated through hydrograph components using following formulas.

The flow peak series p_j is calculated as.

$$p_j = y_{hj}(t_k) \tag{1}$$

where $y_{hj}(t_k)$ is the highest recorded observation of flow on a *k*th day in a *j*th year.

The flow volume series v_j is calculated as.

$$v_j = \sum_{l=SD_j}^{ED_j} y_j(t_k) - \frac{1}{2}\,(y_{ij}(t_k) + y_{fj}(t_k)) \tag{2}$$

where $y_j(t_k)$ are the recorded observations of flow on a *k*th day in a *j*th year, $y_{sj}(t_k)$ and $y_{ej}(t_k)$ are the recorded observation of flow on starting (SD_j) and ending day (ED_j) respectively, in the *k*th year of flood time span.

B. Analysis of Functional Principal Component

Analysis of principal component (APC) practices in a multivariate study for reducing the dimensionality through the computation of new variables which are the linear combination for original values so that the maximum of data variation could be captured. After the conversion of data as functions, analysis of functional principal component (AFPC) permits us to compute new functions so that special kind of variation for curve data could be revealed [5]. The AFPC method maximizes sample variance scores as orthonormal constraints. It divides the functional centred observations in orthogonal basis form and defined as follows.

Let functional observations be $y_j(t), j = 1, \dots, n$ obtained after smoothing the discrete observations $\left(y_j(t_1), \dots, y_j(t_T)\right), j = 1, \dots, n$. By definition, the curve of mean is a same variation for most of the curves which can be fixed by centering. Let $(y_j^*(t) = y_j(t) - \bar{y}(t))_{j=1,\ldots,n}$ be functional centered observations where $\bar{y}(t)$ represents the function of mean for $(y_1(t), \ldots, y_n(t))$. Now AFPC is applied to $(y_j^*(t))_{j=1,\ldots,n}$ for creating a set of small functions, known as harmonics which reveals the type of variation important for analysis. The first principal component $(y_j^*(t))_{j=1,\ldots,n}$ denoted as $w_1(t)$ be a function so that variance regarding corresponding scores $z_{j,1}$ of real value is as follows.

$$z_{j,1} = \int_C w_1(s)\, y_j^*(s) ds, j = 1, \ldots, n \tag{3}$$

is maximized under $\int_C w_1(s)^2 ds = 1$ constraint. The next $w_l(t)$; a principal component computed by maximization of variance related to corresponding scores $z_{j,l}$:

$$z_{j,l} = \int_C w_l(s) y_j^*(s) ds, j = 1, \ldots, n \tag{4}$$

under $\int_C w_l(s) w_k(s) ds = 0, l \geq 2, l \neq k$ constraints.

C. Detection of Outliers

The approaches for detection of outliers employed by [4] in the multivariate context was adapted by [5] in functional context; applying functions of outlyingness on the scores of initial two principal components. The purpose of this adaption is to create a comparison between multivariate and functional results.

Functions of outlyingness in a multivariate context were described and employed for detecting outliers. These functions have values ranging [0,1] interval. The outlyingness of a particular point is measured related to the whole sample. A value of outlyingness close to 1 shows high outlyingness, and a value close to 0 shows centrality. An observation is determined to be an outlier by defining a threshold i.e. the outlyingness value corresponds to an outlier must exceed their respective threshold values. Reference [3] introduced outlyingness functions which are based on the functions of depth, are going to be presented in the following section.

1) Outlyingness functions: A depth function is transformed to depth outlyingness for a F given distribution and $x \in R^d$. Reference [3] studied as follows.

a) Half space

$$O_{HO}(x, F) = 1 - 2HO(x, F) \tag{5}$$

b) Mahalanobis

$$O_{MO}(x, F) = d^2{}_{A(F)}(x, \mu(F)) / [1 + d^2{}_{A(F)}(x, \mu(F))] \tag{6}$$

c) Projection

$$O_{PO}(x, F) = PO(x, F) / [1 + PO(x, F)\,] \tag{7}$$

where $HO(., F)$, $d^2{}_{A(F)}(., \mu(F))$ and $PO(., F)$ are given by [4], a location measure is $\mu(F)$ and $A(F)$ is non-singular measure of scale matrix.

Spatial

$$O_{so}(x, F) = \|E(Sign(x - X))\| \tag{8}$$

d) Spatial Mahalanobis

$$O_{Ms}(x, F) = \left\| E[Sign(\mathbf{C}^{-\frac{1}{2}}(x - X)] \right\| \tag{9}$$

where the Euclidean norm is $\|.\|$, F-distribution is X and the sign multidimensional function is $Sign(.)$ given by $Sign(x) = x/\|x\|\ if\ x \neq 0\ and\ Sign\ (0) = 0$ also, $\mathbf{C}$ is any positive definite affine invariant $d \times d$ symmetric matrix.

2) Threshold: An essential step in the detection of an outlier is the appropriate selection of the threshold. It relates to true positive and false positive rates. α_n denoted for a false positive arbitrary rate which is defined as the proportion of misidentified nonoutliers as outliers. This constant relates closely to the ε_n true positive rate by which the theoretical proportion for real outliers are represented (also known as contaminants). Ideally, α_n suppose to be smaller than ε_n. Reference [3] fixed the false outliers' ratio $\delta = \alpha_n/\varepsilon_n$ and also used another coefficient $\beta = \varepsilon_n\sqrt{n}$, in order to define a threshold for the values of outlyingness as $(1 - \alpha_n)$ quantile.

$$\rho_n = F^{-1}_{O(X,F)}(1 - \alpha_n) = F^{-1}_{O(X,F)}\left(1 - \delta\beta_n/\sqrt{n}\right) \tag{10}$$

where false positive rate α_n is represented as $\alpha_n = \delta\beta_n/\sqrt{n}$ and true positive rate ε_n represented as $\varepsilon_n = n\varepsilon_n/n$; a number of true outliers are $n\varepsilon_n$ and a number of observations are n, in such a way that $\alpha_n < \varepsilon_n$. For further calculations and applications, readers are referred to follow [4].

IV. Data Description

The major source of hydrological data is daily streamflow. The daily flow data series of the Kotri barrage are available from Sindh Irrigation department, Sindh Secretariat, Karachi, Pakistan.

A daily flow observations (m^3s^{-1}) of Kotri barrage which is located between Jamshoro and Hyderabad in Sindh province on the Indus River, Pakistan. It has a discharge capacity of 875,000 cusecs (i.e. approximately 24800 m^3s^{-1}). Fig. 1 indicates the geographical location of the Kotri Barrage.

Some studies contain data of complete year while some consider section of a year having high flow observations. Hydrological data observations of the present study contain a duration of 6 months ($i.e. T = 183$ days) per year spanning 1977 to 2017 (i.e. n=41 years) since high flow period is observed during the months April to September, in Pakistan. The series of observations are $Y_j = \left(y_j(t_1), \ldots, y_j(t_T)\right)'$, $j = 1, \ldots, n$, $k = 1, \ldots, T$, where n=41 years, $T = 183$ days and $y_j(t_k)$ is the recorded flow observation on t_k day in the jth year. Before any computation is performed the streamflow observations which are recorded on measurement scale in cusec (a volume flow rate) are required to be converted into cubic meter per second (m^3s^{-1}).

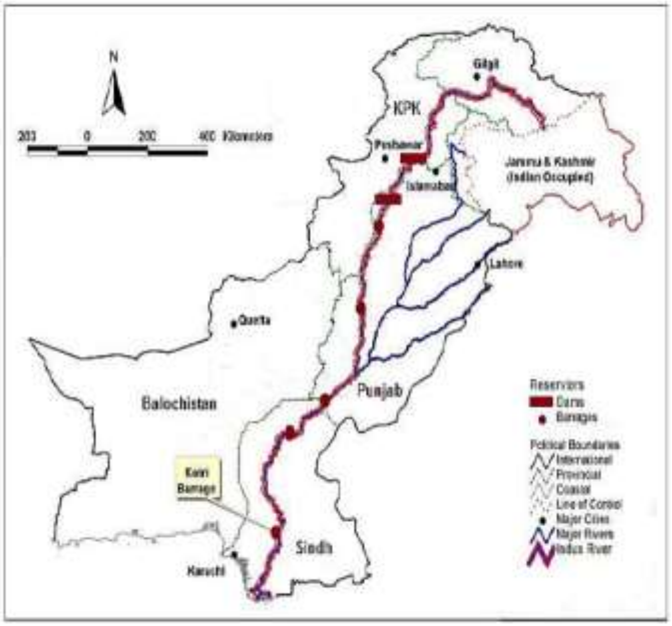

Fig. 1. Geographical Location of Kotri Barrage.

V. Application

The two most studied and examined characteristics of the flood that is peak (p) and volume (v) are focused here. The series of bivariate (p,v) are computed by using (1) and (2) and results are displayed in Table I.

According to [4], an approach developed by [3] are based on the function of depth outlyingness and the threshold corresponded. The four functions of depth outlyingness are evaluated for the (p,v) series of bivariate observation i.e., Mahalanobis (MO), Projection (PO), Spatial (SO) and Tukey (TO). The values of depth outlyingness correspond to each (p,v) observation for years 1977-2017 are reported in the last four columns of Table I. The thresholds correspond to each outlyingness functions are computed by selecting 15% false outlier ratio and the number of true outliers as 5, this selection is similar to the choices made by [4] in such a way that the outlyingness value corresponds to an outlier must exceed their respective threshold values.

Hence, 98% quantile is a corresponding threshold for the values of outlyingness. The computed values of the threshold for MO, PO, SO & TO are 0.9412, 0.9040, 0.9719, and 0.9444, respectively. The values of threshold approximately remain constant if the number of true outliers is considered greater than 5 with changed false outlier ratio i.e. 5%, 10% and 20%. The detected outliers correspond to MO, PO, SO & TO with respect to their respective threshold values are graphically displayed by Fig. 2.

Reference [5] employed the procedure for detecting outliers which are based on the function of depth outlyingness and the threshold corresponded. As discussed earlier and also practiced in preceding section, four functions of depth outlyingness are evaluated for the series of the bivariate score (z_1, z_2) i.e., Mahalanobis (MO), Projection (PO), Spatial (SO) and Tukey (TO).

TABLE I. Multivariate Results For Flood Peak and Volume

Year	Peak	Volume	MO	PO	SO	TO
1977	7490	248765	0.0979	0.5424	0.4134	0.4634
1978	15747	249063	0.8782	0.8631	0.4183	0.9512
1979	7342	305373	0.4843	0.7099	0.6352	0.7561
1980	5776	170479	0.0852	0.2978	0.0255	0.2195
1981	7149	246426	0.1473	0.5673	0.3586	0.5610
1982	5560	129340	0.1783	0.4059	0.2671	0.3171
1983	9367	260061	0.1161	0.5753	0.4844	0.5610
1984	7913	290839	0.2922	0.6491	0.5849	0.7073
1985	3662	126804	0.3419	0.5121	0.3348	0.5610
1986	10160	185277	0.6149	0.7608	0.1526	0.9024
1987	2771	128432	0.4982	0.6217	0.2893	0.9024
1988	14527	467773	0.6348	0.7848	0.7808	0.8049
1989	6276	112997	0.3567	0.6141	0.3900	0.6585
1990	6355	243994	0.3066	0.6250	0.3110	0.6585
1991	5309	276870	0.6496	0.7430	0.5363	0.9512
1992	15241	618581	0.8350	0.8484	0.8783	0.9024
1993	9617	217016	0.3713	0.6765	0.1981	0.7073
1994	19109	921882	0.9482	0.9043	0.9756	0.9512
1995	17998	483519	0.7882	0.8274	0.8288	0.8537
1996	8520	417460	0.7610	0.8073	0.7321	0.9024
1997	6898	145428	0.2501	0.5854	0.1765	0.4634
1998	6263	181396	0.0444	0.2874	0.1065	0.2195
1999	4133	59546	0.4171	0.5835	0.5856	0.8049
2000	1372	27595	0.5406	0.6543	0.8807	0.9512
2001	1969	39701	0.4927	0.6301	0.6815	0.8537
2002	2581	32254	0.4782	0.6272	0.7895	0.8537
2003	4171	146269	0.3006	0.4783	0.1627	0.5122
2004	898	30884	0.5784	0.6626	0.8236	0.9512
2005	6800	236405	0.1577	0.5614	0.2491	0.5122
2006	7922	154970	0.3857	0.6698	0.0733	0.7073
2007	3323	147582	0.4653	0.5966	0.1364	0.8049
2008	2882	87966	0.4016	0.5513	0.4880	0.6098
2009	2111	36592	0.4870	0.6291	0.7316	0.8049
2010	28244	694249	0.9404	0.9044	0.9267	0.9512
2011	4459	45005	0.5054	0.6305	0.6391	0.9512
2012	2115	22688	0.5078	0.6420	0.9725	0.9512
2013	8475	174738	0.3751	0.6731	0.0634	0.6585
2014	3005	24519	0.5024	0.6425	0.9248	0.9512
2015	14155	325111	0.6981	0.7957	0.6819	0.8537
2016	3257	86015	0.3600	0.5389	0.5355	0.5610
2017	5730	97637	0.3715	0.5963	0.4407	0.7561

Legend

Highest
2nd Highest
3rd Highest

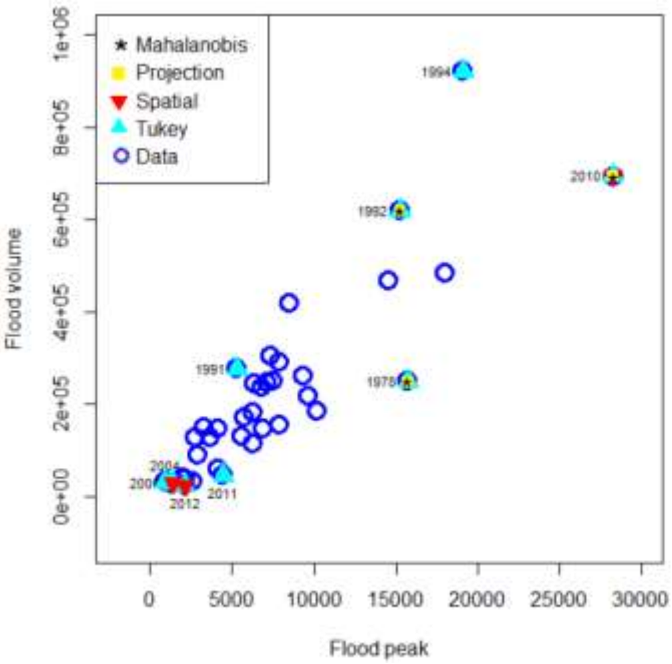

Fig. 2. Detected Outliers using Flood Peak and Volume.

The thresholds correspond to each outlyingness functions are computed by selecting 15% false outlier ratio and the number of true outliers as 5, this selection is similar to the choices made by [4] in such a way that the outlyingness value corresponds to an outlier must exceed their respective threshold values. Hence, 98% quantile is a corresponding threshold for the values of outlyingness. The computed values of the threshold for MO, PO, SO & TO are 0.9106, 0.8905, 0.9264, and 0.9444, respectively. The values of threshold approximately remain constant if the number of true outliers is considered greater than 5 with changed false outlier ratio i.e. 5%, 10% and 20%. The computed outlyingness values of MO, PO, SO & TO for years 1977-2017 are tabulated in Table II whereas Fig. 3 displays the detected outliers correspond to MO, PO, SO & TO with respect to their respective threshold values.

VI. Results

A. Multivariate Result

The year 1994 contain outlyingness values greater than their respective threshold values by MO, PO & SO functions. Several years including years 1978, 1994, 2010 and 2012 are detected by TO function as outliers. The year 2010 is detected by MO and PO, and year 2012 is detected by SO functions as the closest value of outlyingness with respect to their threshold values. In addition, the year 1978 corresponds to the third highest MO and PO values whereas the year 2010 correspond the third highest SO value compare to their respective threshold values. Hence, it can objectively be inferred from Table I that the years 1994 and 2010 are identified as outliers by all the four functions of outlyingness. Whereas, the year 1978 is detected by the three and the year 2012 is detected by the two functions of outlyingness. For illustrative purpose a scatter plot constructed between bivariate (p,v) series (i.e. flood peak and flood volume) is displayed through Fig. 2 so that the above interpretation can explicitly comprehensible. The years 1978, 1990, 1994, 2000, 2004, 2010, 2011, 2012 and 2014 computed as outliers by the outlyingness functions, among them the years 1978 and 1992 are present outside compare to the rest of the years whereas the years 1994 and 2010 are appear as outliers.

TABLE II. Functional Results For Principal Component (z_1, z_2)

Year	z_1	z_2	MO	PO	SO	TO
1977	-2.29	-2.339	0.1671	0.5215	0.3226	0.5122
1978	13.09	-10.776	0.8342	0.8565	0.8512	0.9024
1979	6.218	7.085	0.6323	0.8256	0.6295	0.7561
1980	-3.056	0.173	0.1077	0.2189	0.0381	0.1707
1981	9.388	8.516	0.7435	0.8548	0.7702	0.9512
1982	-5.564	3.525	0.4119	0.5763	0.4923	0.7073
1983	7.676	-2.302	0.4697	0.7249	0.6046	0.6585
1984	-3.352	-4.623	0.3994	0.6705	0.5732	0.8537
1985	-9.07	-1.06	0.5201	0.5805	0.7348	0.9512
1986	-3.226	-2.818	0.2465	0.5537	0.3994	0.6585
1987	1.832	5.992	0.4786	0.7791	0.5310	0.6585
1988	7.617	-3.672	0.5177	0.7498	0.6238	0.7073
1989	-5.09	0.182	0.2501	0.3459	0.2150	0.3171
1990	6.42	-1.683	0.3743	0.6943	0.5049	0.5610
1991	20.652	9.365	0.8839	0.8928	0.9015	0.9512
1992	28.743	3.228	0.9157	0.8888	0.9422	0.9512
1993	7.174	7.721	0.6788	0.8378	0.6965	0.8049
1994	7.345	-21.331	0.9217	0.8938	0.9077	0.9512
1995	9.233	-6.894	0.6926	0.8068	0.7436	0.8049
1996	7.365	-4.316	0.5350	0.7577	0.6280	0.7561
1997	-4.721	-0.359	0.2244	0.2994	0.2017	0.3659
1998	9.949	8.385	0.7490	0.8559	0.7998	0.9024
1999	-6.452	1.793	0.3800	0.4861	0.4139	0.5610
2000	-10.234	2.782	0.6053	0.6467	0.8525	0.9512
2001	-5.921	6.952	0.6195	0.7290	0.7311	0.9512
2002	-9.377	1.845	0.5479	0.5970	0.7410	0.8537
2003	-3.783	-0.194	0.1559	0.2193	0.0807	0.2195
2004	-10.197	2.608	0.6002	0.6415	0.8266	0.9512
2005	0.865	2.131	0.1073	0.6408	0.3050	0.4634
2006	-6.466	-2.704	0.4169	0.6282	0.6128	0.8537
2007	0.884	6.854	0.5359	0.7897	0.5904	0.9024
2008	-8.884	0.856	0.5077	0.5714	0.6634	0.7561
2009	-8.824	1.963	0.5224	0.5829	0.6608	0.8049
2010	-0.407	-19.296	0.9007	0.8898	0.8743	0.9512
2011	-6.425	-1.158	0.3601	0.4990	0.4612	0.6098
2012	-9.277	-0.121	0.5251	0.5829	0.7346	0.9024
2013	-6.19	-2.373	0.3862	0.5996	0.5283	0.7561
2014	-7.242	1.574	0.4233	0.5073	0.4995	0.6098
2015	1.404	-1.802	0.0946	0.5918	0.3483	0.4634
2016	-3.317	5.357	0.4567	0.7064	0.5341	0.9024
2017	-6.487	0.934	0.3596	0.4502	0.3736	0.5610

Legend

Highest
2nd Highest
3rd Highest

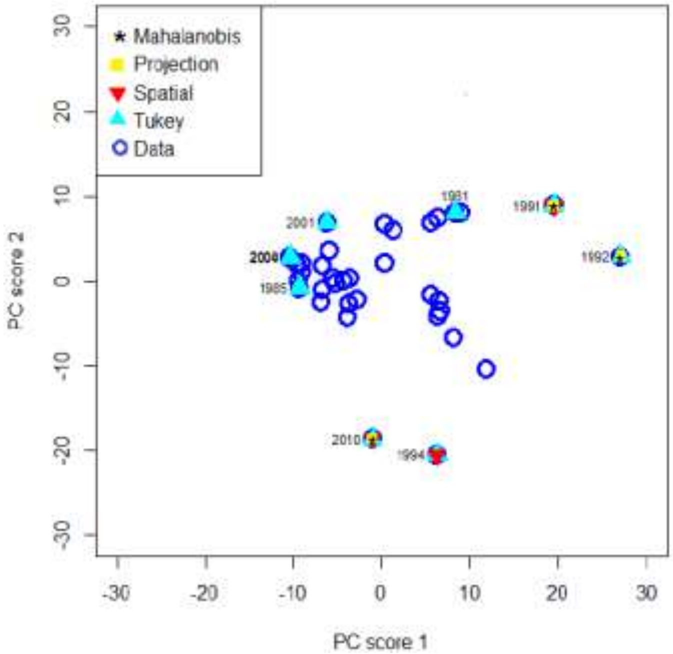

Fig. 3. Detected Outliers using Principal Component Scores.

B. Functional Result

It is observed that the year 1994 contain outlyingness values greater than their respective threshold values by MO and PO functions whereas outlyingness value of the year 1992 is greater than the threshold value by SO function. Several years including 1991, 1992, 1994 and 2010 are detected by TO function as outliers. The year 1991 is detected by the PO, the year 1992 is detected by MO and the year 1994 is detected by the SO functions as a second highest outlyingness values compare to their respective threshold values. In addition, the year 2010 corresponds to the third highest MO and PO values, whereas the year 1991 corresponds to the third highest SO outlyingness value according to their respective threshold values.

Hence, it can distinctly be inferred from the values of Table II, the year 1994 is detected by all the four outlyingness functions as an outlier. Whereas the years 1991, 1992 and 2010 are identified as outliers by the three outlyingness functions. Above interpretation can better be comprehended by the scatter plot constructed between scores of initial two principal components (i.e. PC score 1 & score 2) and represented by Fig. 3 which reveals that the years 1981, 1985, 1991, 1992, 1994, 2000, 2001, 2004 and 2010 computed as outliers by the outlyingness functions, among them the years 1991 and 1992 are present outside compare to the rest of the years whereas the years 1994 and 2010 are appear as outliers.

The functional results are almost consistent with the results of the multivariate framework such that the years 1992, 1994 and 2010 have been detected as the most unusual flows in both the multivariate and functional context.

VII. Conclusion

The nonparametric techniques based on depth function for outlier identifiers have been practiced in two different frameworks of study that are multivariate hydrological data analysis and functional hydrological data analysis. The identification of outlier is essential for the appropriate selection of suitable hydrologic models so that risk associated with flood events can be authentically estimated. The methods employed in the present research are multivariate methods that are superior to previously practiced classical methods that were moment-based, follow normality assumption and component-wise techniques. The implemented techniques are based on depth function notion, free of moment, do not require normality assumption, and also affine invariant.

The proposed approaches have been implemented in two different frameworks of analysis. The intention of executing this study is to gauge the performance of proposed methodologies in both multivariate and functional context. The two most widely practice flood characteristics in hydrological analysis, peak (p) & volume (v) have been included to execute study in multivariate hydrological data analysis. Besides this, two initial scores of principal components (z_1, z_2) used as a series of bivariate variables for executing functional hydrological data analysis since initial two principal components have a capability to capture major variation of data employed for analysis.

The outliers of both the framework are almost consistent but the results of functional analysis can be considered more reliable since it is based on complete information of flood hydrograph whereas flood characteristics (p, v) are not able to generate hydrograph even though more than two characteristics of flood are included in study. Nevertheless, the multivariate results cannot be ignored and must be employed in a parallel complement to functional results so that dynamics of a hydrological event can be analysed to attain comprehensive information related to causes of flood.

References

[1] V. Barnett, and T. Lewis, Outliers in Statistical Data, 3rd ed., John Wiely, Chichester, U.K, 1998.

[2] V. Barnett, Environmental Statistics: Methods and Applications, John Wiely, Chichester, U.K, 2004.

[3] X. Dang, and R. Serfling, "Nonparametric depth-based multivariate outlier identifiers, and masking robustness properties," Journal of Statistical Planning and Inference, vol. 140, no. 1, pp. 198–213, 2010. doi: 10.1016/j.jspi.2009.07.004.

[4] F. Chebana and T. B. Ouarda, "Depth-based multivariate descriptive statistics with hydrological applications," Journal of Geophysical Research, vol. 116, D10120, 2011b. doi:10.1029/2010JD015338.

[5] F. Chebana, S. Dabo-Niang and T. B. Ouarda, "Exploratory functional flood frequency analysis and outlier detection," Water Resources Research, vol. 48, no. 4, W04514, 2012. doi:10.1029/2011WR011040.

[6] F. Chebana, and T. B. Ouarda, "Depth and homogeneity in regional flood frequency analysis," Water Resources Research, vol. 44, W11422, 2008. doi:10.1029/2007WR006771.

[7] T. W. Anderson, An Introduction to Multivariate Statistical Analysis, 2nd ed., John Wiley, Chichester, U. K, 1984.

[8] M. J. Schervish, "A review of multivariate analysis," Statistical Science, vol. 2, no. 4, pp. 413-417, 1987. doi:10.1214/ss/1177013111.

[9] Y. Zuo, and R. Serfling, "General notions of statistical depth function," Annals of Statistics, vol. 28, no. 2, pp. 461–482, 2000b. doi:10.1214/aos/1016218226.

[10] R.Y. Liu, J. M. Parelius, and K. Singh, "Multivariate analysis by data depth: Descriptive statistics, graphics and inference," Annals of Statistics, vol. 27, no. 3, pp. 783–858, 1999.

[11] J. Zhang, "Some extensions of Tukey's depth function," Journal of Multivariate Analysis vol. 82, no. 1, pp. 134–165, 2002.

[12] Y. Zuo, and R. Serfling, "On the performance of some robust nonparametric location measures relative to a general notion of multivariate symmetry," Journal of Statistical Planning and Inference, vol. 84, no. 1–2, pp. 55–79, 2000a. doi:10.1016/S03783758(99)00142-1.

[13] C. H, Müller, "Depth estimators and tests based on the likelihood principle with application to regression," Journal of Multivariate Analysis, vol. 95, no. 1, pp. 153–181, 2005.

[14] I. Hussain, and M. Uddin, "Functional and multivariate hydrological data visualization and outlier detection of Sukkur Barrage," International Journal of Computer Applications, vol. 178, no. 28, pp. 20-29, 2019. doi:10.5120/ijca2019919097.

[15] I. Hussain, "Outlier detection using graphical and nongraphical functional methods in hydrology," International Journal of Advanced Computer Science and Applications, vol. 10, no. 12, pp. 438-445, 2019. doi: 10.14569/IJACSA.2019.0101259.

[16] G. A. Griffiths, S. K. Singh, and A. I. McKerchar, "Flood frequency estimation in New Zealand using a region of influence approach and statistical depth functions," Journal of Hydrology, vol. 589, pp. 125-187, 2020. doi 10.1016/j.jhydrol.2020.125187.

[17] M. Karamouz, F.ASCE, F. Ahmadvand, and Z. Zahmatkesh, "Distributed hydrologic modelling of coastal flood inundation and damage: Nonstationary approach," Journal of Irrigation and Drainage Engineering, vol. 143, no. 8, 2017. doi: 10.1061/(ASCE)IR.1943-4774.0001173.

[18] W. Fan, L. Heng, D. Chao & D. Lieyun,"Knowledge representation using non-parametric Bayesian networks for tunneling risk analysis," Reliability Engineering and System Safety, Elsevier, vol. 191(C), 2019. doi: 10.1016/j.ress.2019.106529.

[19] L. Millán-Roures, I. Epifanio, and V. Martínez, "Detection of Anomalies in Water Networks by Functional Data Analysis," Mathematical Problems in Engineering, 2018. doi: org/10.1155/2018/5129735.

[20] Y. Zhang, S. Huang, Q. Huang, G. Leng, H. Wang, and L. Wang, "Assessment of drought evolution characteristics based on a nonparametric and trivariate integrated drought index," Journal of Hydrology, vol. 579, 2019. doi 10.1016/j.jhydrol.2019.124230.

[21] K. T. Peterson, V. Sagan, and J. J. Sloan, "Deep learning-based water quality estimation and anomaly detection using Landsat-8/Sentinel-2 virtual constellation and cloud computing," GIScience & Remote Sensing, vol. 57, no. 4, pp. 510-525, 2020. doi: 10.1080/15481603.2020.1738061.

[22] J. Rhee, K. Park, S. Lee, S. Jang, and S. Yoon, "Detecting hydrological droughts in ungauged areas from remotely sensed hydro-meteorological variables using rule-based models," Natural Hazards, vol. 57, 2020. doi:10.1007/s11069-020-04114-5.

[23] R. Padulano, G. D. Giudice, "A nonparametric framework for water consumption data cleansing: an application to a smart water network in Naples (Italy)," Journal of Hydroinformatics, vol. 22, no. 4, pp. 666–680, 2020. doi: org/10.2166/hydro.2020.133.

[24] S. Samadi, D. L. Tufford, and G. J. Carbone, "Estimating hydrologic model uncertainty in the presence of complex residual error structures," Stochastic Environmental Research and Risk Assessment, vol. 32, pp. 1259–1281, 2018. doi: org/10.1007/s00477-017-1489-6.

[25] Y. Zuo, "On general notions of depth for regression," arXiv e-prints, 2018.

[26] S. Pool, M. Vis, and J. Seibert, "Evaluating model performance: towards a non-parametric variant of the Kling-Gupta efficiency," Hydrological Sciences Journal, vol. 63, no. 13-14, pp. 1941-1953, 2018. doi: 10.1080/02626667.2018.1552002.

[27] M. A. Sherly, S. Karmakar, T. Chan, and C. Rau, "Design Rainfall Framework Using Multivariate Parametric-Nonparametric Approach," Journal of Hydrologic Engineering, vol. 21, no. 1, 2016. doi: org/10.1061/(ASCE)HE.1943-5584.0001256.

[28] W.W.U.I. Wickramaarachchi, T.U.S. Peiris, and S. Samita, "Rainfall Trends in the North-Western and Eastern Coastal Lines of Sri Lanka Using Non – Parametric Analysis," Tropical Agricultural Research, vol. 31, no. 2, pp. 41-54, 2020. doi: 10.4038/tar.v31i2.8366.

[29] U. Beyaztas, and Z. M. Yaseen, "Drought interval simulation using functional data analysis" Journal of Hydrology, vol. 579, 2019. doi: 10.1016/j.jhydrol.2019.124141.

[30] J. Xia, P. Yang, C. Zhan and Y. Qiao, "Analysis of changes in drought and terrestrial water storage in the Tarim River Basin based on principal component analysis," Hydrology Research, vol. 50 no. 2, pp. 761–777, 2019. doi: 10.2166/nh.2019.033.

[31] M. A. Hael, "Modeling of rainfall variability using functional principal component method: a case study of Taiz region, Yemen," Modeling Earth Systems and Environment, vol. 2, no. 7, 2020, doi: 10.1007/s40808-020-00876-w.

[32] M. A. Hael1, Y. Yongsheng, and B. I. Saleh, "Visualization of rainfall data using functional data analysis", SN Applied Sciences, vol. 2, no. 2, 2020. doi:10.1007/s42452-020-2238-x.

[33] S.M. Shaharudin, N. Ahmad, N.H. Zainuddin, and N.S. Mohamed, "Identification of rainfall patterns on hydrological simulation using robust principal component analysis", Indonesian Journal of Electrical Engineering and Computer Science, vol. 11, no. 3, pp.1162-1167, 2018. doi: 10.11591/ijeecs.v11.i3.pp1162-1167.

[34] J. Suhaila, Application of Functional Data Analysis in Streamflow Hydrograph. In: Kor LK., Ahmad AR., Idrus Z., Mansor K. (eds) Proceedings of the Third International Conference on Computing, Mathematics and Statistics (iCMS2017). (2019) Springer, Singapore.

[35] M. A. T. M. T. Rahman, S. Hoque, and A. H. M. Saadat, "Selection of minimum indicators of hydrologic alteration of the Gorai river, Bangladesh using principal component analysis," Sustainable Water Resources Management, 2017. doi: org/10.1007/s40899-017-0079-6.

[36] S. Xiao, Z. Lu, and L. Xu, "Multivariate sensitivity analysis based on the direction of eigen space through principal component analysis," Reliability Engineering & System Safety, vol. 165, 2017. doi: 10.1016/j.ress.2017.03.011.

[37] S. K. Sharma, S. Tignath, S. Gajbhiye , and R. Patil, "Application of principal component analysis in grouping geomorphic parameters of Uttela watershed for hydrological modeling," International Journal of Remote Sensing & Geoscience, vol. 2, no. 6, 2013.

[38] C. Gyamfi, J. M. Ndambuki, and R. W. Salim, "Simulation of sediment yield in a semi-arid River Basin under changing land use: an integrated approach of hydrologic modelling and principal component analysis," Sustainability, vol. 8, no. 11,2016. doi: 10.3390/su8111133.

[39] S. Yue, T. B. Ouarda, B. Bobée, P. Legendre, and P. Bruneau, "The Gumbel mixed model for flood frequency analysis," Journal of Hydrology, vol. 228, pp.88-100, 1999. doi:10.1016/S0022-1694(99)00168-7.

[40] J. T. Shiau, "Return period of bivariate distributed extreme hydrological events," Stochastic Environmental Research and Risk Assessment, vol. 17, pp.42–57, 2003. doi:10.1007/s00477-003-0125-9.

[41] S. Naz, M. J. Iqbal, S. M. Akhter, and I. Hussain, "The Gumbel mixed model for food frequency analysis of Tarbela," The Nucleus, 53(3), pp. 171-179, 2016.